Multimed Tools Appl (2017) 76:15413–15434
DOI 10.1007/s11042-016-3841-0

Adaptive 3D shape context representation for motion trajectory classification

Weihua Liu[1] · Zuhe Li[2] · Geng Zhang[1] · Zhong Zhang[3]

Received: 14 December 2015 / Revised: 8 June 2016 / Accepted: 4 August 2016 /
Published online: 31 August 2016

Abstract The measurement of similarity between two motion trajectories is one of the fundamental task for motion analysis, perception and recognition. Previous research focus on 2D trajectory similarity measurement. With the advent of 3D sensors, it is possible to collect large amounts of 3D trajectory data for more precise motion representation. As trajectories in 3D space may often exhibit a similar motion pattern but may differ in location, orientation, scale, and appearance variations, the trajectory descriptor must be invariant to these degrees of freedom. Shape context is one of the rich local shape descriptors can be used to represent the trajectory in 2D space, however, rarely applied in the 3D motion trajectory recognition field. To handle 3D data, in this paper, we first naturally extend the shape context into the spatiotemporal domain by adopting a spherical neighborhood, and named it 3D Shape Context(3DSC). To achieve better global invariant on trajectories classification, the adaptive outer radius of 3DSC for extracting 3D Shape Context feature is proposed. The advantages of our proposed 3D shape context are: (1) It is invariant to motion trajectories translation and scale in the spatiotemporal domain; (2) It contains the whole trajectory points in the 3DSC ball volume, thus can achieve global information representation and is good for solving sub-trajectories problem; (3) It is insensitive to the appearance variations in the identical meaning trajectories, meanwhile, can greatly discriminate the distinct meaning trajectories. In trajectory recognition phase, we consider a feature-to-feature alignment between motion trajectories based on dynamic time warping and then use the one nearest neighbor (1NN) classifier for final accuracy evaluation. We test the performance of proposed 3D SC-DTW on UCI ASL

✉ Weihua Liu
lwh86117@163.com

Geng Zhang
524764784@qq.com

[1] Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China

[2] School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

[3] The University of Texas at Arlington, Texas, USA

large dataset, Digital hand dataset and the experimental results demonstrate the effectiveness of our method.

Keywords Gesture trajectory · Shape context · Dynamic time warping · Gesture Classification

1 Introduction

Motion trajectory analysis is one of fundamental tasks in many types of research area, such as activity recognition [19, 21, 33], robots action [15], anomaly detection [22], video surveillance [20], etc.

The studies in motion trajectory classification can be divided into three main categories with respect to the classification scheme: model based, feature based and trajectory based. Model based [11] classification aims to build a parameterized model for a collection of motion trajectories, and classifying a new motion trajectory by determining the model that best fits to it. The statistical models such as Gaussian, Markov and Hidden Markov Models [14], Random Forest [18], Neural network are frequently used in this category. In the feature based studies [28], the high dimensionality of motion trajectories are transformed into a set of features with the help of suitable feature extraction methods. For example, Lin and Hsieh [16] develop a Kernel-based representation for motion trajectory retrieval and classification. Bashir, et al. [1]propose a Principle Component Analysis-based approach for trajectory indexing and retrieval. Although such methods state above can optimally give a set of features, an extensive survey claims that they become unfeasible for out-of-core processing which is usually the case for large datasets[6].

Our motion trajectory classification framework falls into the trajectory-based category. In this category, two of the fundamental problems are trajectories representation and the sequences alignment. An effective representation can yield significantly better performance than the raw trajectory data. Moreover, alignment methods can also be used to improve the performance of a classifier by integrating the trajectories representation into the alignment.

A basic way of comparing two trajectories is use the original data which rely on the absolute positions of motion, yet ineffective in computation and invariant to motion trajectory translation and scale [8]. Hence, the most challenge in motion trajectory classification is the specific trajectory adjustments either in alignment or representation steps. To maintain the scale and translation invariant during alignment, the raw trajectories data are often converted in the form of geometric invariant signatures such as curvature, torsion, and their derivatives, etc. To develop an effective system for real-world applications, many researchers have devoted themselves to find a compact yet discriminative representation for trajectories. In the existing work, Yang et al. [31] represented trajectories in segment level rather than point level, and index the trajectories by four segment sequences classes to recognize them. However, this methods can only represent simple shapes and are inefficient in complex and long term trajectories. In Wu et al' s [26] paper, three geometric invariant signature descriptions for motion characterization are developed. Such flexible descriptions give the signature high functional adaptability to meet various application requirements in trajectory representation, perception and recognition, however, computing three differential features and two global features for each discrete point is computationally expensive. Also, reliable finite difference estimation of higher order derivatives is difficult due to their high sensitivity to noise. Yang et al. [29, 32] also present a mixed invariant signature descriptor in the basis of differential feature with global invariants for motion perception and recognition, however, it has the same problems with Wu's method.

Dynamic time warping [4, 13, 30] is the most representative non-linear mapping method which can handle the local time shifting by finding an optimal path with minimal sum of distance from a pairwise distance matrix. In the traditional perspective, the pointwise based alignment usually resorts to use the local information which is adjacent to each point [12]. Currently, Z, Zhang [34] have recently suggested using shape context [2] as a rich local shape descriptor to replace the raw observed position value in conventional DTW in 2D space. To keep the locality of a shape context, they set the outer radius of shape context to one tenth of the length of the trajectories. As an alternative, this approach generates a more feature-to-feature alignment between motion trajectories and thus serves as a robust similarity measure. However, this matching mechanism cannot handle a situation that a trajectory partially similar with its sub-trajectory. In addition, from a shape point of view, ignoring the relationship of each point to its global sequence context, are largely unstable and incline to result in pathological alignment [25]. In our point of view, to build effective trajectory recognition, it will be helpful to refer to the concept of global shape descriptor on each trajectory point, and it will outperform the local descriptor. For this purpose, we propose an adaptive outer radius of 3D shape context mechanism. During trajectory feature representation, the global trajectories information are extracted with the largest point-wise distance in each trajectory equals to the outer most circle radius of shape context descriptor. In this way, the global information of each trajectory point can be included and extracted as pointwise feature for trajectories alignment.

Another challenge of motion trajectory classification is that the appearance variations among the identical meaning trajectories may cause false discrimination. In the most motion trajectory applications, such trajectories perturbations are emerged because these trajectories are drawn repeatedly by different users. As far as we study, most of researchers are not conscious of this problem. In our previous work [17], the 2D shape order context descriptor combined with DTW was proposed. It is greatly insensitive to trajectories perturbations and highly invariant to trajectories scale and translation. However, this previous work did not refer to the 3D trajectories.

As we state above, motion trajectory classification approaches in 2D space are wildly studied and already achieves better performance. It should be noted that emphasize on using projected 2D trajectory in some kind of viewpoint, may lost the authentic meaning in the 3D space. Compared with projected 2D trajectories, more rich information can be drawn from 3D trajectories in spatiotemporal domain and therefore better performance can be achieved in trajectory-based schemes. Also, with more and more complicate as behaviors and activities are performed, 3D trajectory analysis should be further considered. Based on 2D shape context descriptor, the extension of 3D shape context descriptor has been proposed in [9]. It relies on a specific subdivision of the spherical volume around the trajectories point that needs to be described. Matthias et al. [10] propose to use of 3D shape context to recognize the spatial and temporal details inherent in human actions. However, such 3D shape context descriptors is not restrictively in 3D space, but in 2D space plus time domain. S.H Zhao et al. [35] generalized strategy for dynamic 3D depth data matching and apply this strategy in action retrieval task. In Zhao's paper, they use nine planes to segment the ball-like descriptor into 48 homogeneous regions, resulting in the equal probability of the bins for capturing the contour points in each static depth frame. And then employ dynamic time warping (DTW) to measure the temporal similarity between two 3D dynamic depth sequences. Most previous works on 3D shape context are mainly used in the application of static object such as pose estimation[23], 3D object recognition [3] and registration [27], etc, however, rarely apply in motion trajectory classification.

In this paper, we present a novel feature extraction method for motion trajectories. The proposed approach is very straightforward and simplicity for implementation. It takes a motion trajectory of (x, y, z) coordinates, as the input data. Then we perform a two-stage of normalization to yield a compact trajectories. The first stage implicitly normalize the raw trajectories to the same points length. The second stage, called trajectory span distance normalization, aims to normalize points coordinates by dividing the largest distance which is calculated from any of pointwise in each trajectories. After preprocessing, we extend the traditional shape context to spatiotemporal to capture the global representation for each 3D trajectory point. A series of the resulting representation is guaranteed to be most efficient feature for trajectories alignment by using dynamic time warping. Finally, the trajectories are classified by one nearest neighbor (1NN) classifier. The proposed motion trajectory classification method not only remarkably reduces the miss matching problem, thus improves the classification accuracy, but also effectively avoids the trajectories appearance perturbation and can be greatly invariant to motion trajectories scale and translation.

The rest of our paper is organized as follow. Section 2 presents a framework of our proposed method. In Section 3, we first normalize the trajectories to the same length, then briefly review the shape context descriptor and explicitly expatiated the proposed 3D shape context. In Section 4, the trajectory alignment method based on DTW is introduced and the combination of 3D-SC and DTW for trajectory recognition is elaborated. This paper is concluded in Section 6.

2 Overview of the framework

The main idea behind our motion trajectory classification approach is to obtain a series of global 3D trajectory shape descriptors in spatiotemporal domain, to replace the raw observed values in finding the alignment between two motion trajectories. Figure 1 shows the block diagram of the proposed method. Our representation method for 3D motion trajectory is composed of three functional units, namely pre-processing, 3D shape context representing and dynamic time warping. In the flow diagram, during the pre-processing step, trajectories are firstly normalized to the same length and then each of the trajectory points coordinate is divided by the maximum pointwise Euclidean distance. The pre-processing results are then represented by 3D shape context descriptor in spatiotemporal domain. In the resulting feature space, we integrate the 3D SC descriptor into a dynamic time warping framework for feature-wise aligning and finally use the standard nearest-neighbor algorithm for trajectory classification.

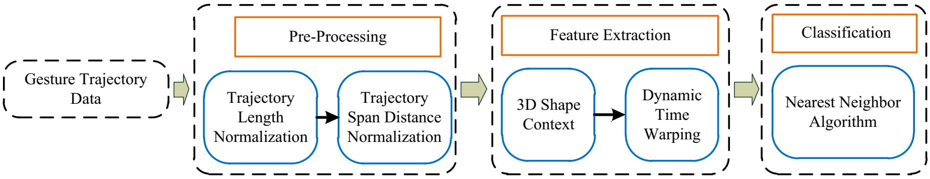


Fig. 1 Block diagram of the 3D shape context based gesture trajectory classification algorithm

3 Trajectory Pre-processing and representation

3.1 Pre-processing

Motion trajectory length normalization is important because of the fact that the majority of trajectories recognition algorithms are need to work with the same length of trajectories points. Since the trajectories are discrete sequences, the number of sampling points may differ from one trajectory to another. Also, the performance of 3D shape context feature will be affected by the distance inequality of two adjacent trajectory points. To avoid directly calculating the 3D-SC feature on a row data, we need to normalize the length of trajectories ahead. This trajectory length normalization based on sampling several neighboring points will reduce the sensitivity of the 3D-SC feature to the trajectory points' intervals. Also, using a fixed amount of points intervals can guarantee that all the identical trajectories in different scales can be represented by the unified histograms.

Given L sampling points along a 3D trajectory, we can represent the set of points as

$$\{x_k, y_k, z_k\}_{k=1}^{L} \tag{1}$$

where $\{x_k, y_k, z_k\}$ denote the $\{x, y, z\}$ coordinates of the $k-th$ point on the trajectory respectively. Some examples of 3D trajectories in the compact Australian Sign Language dataset are shown in Fig. 2.

The starting and ending points of gesture trajectory are manually drawn and the stationary points caused by the signer's holding behavior, are eliminated by considering the relationship between adjacent points' positions.In our experiments, these stationary points can be detected by examining the condition of

$$|x_k - x_{k-\Delta k}| + |y_k - y_{k-\Delta k}| + |z_k - z_{k-\Delta k}| \le \varepsilon \tag{2}$$

where $\Delta k = 2$ and $\varepsilon = 0.01$.

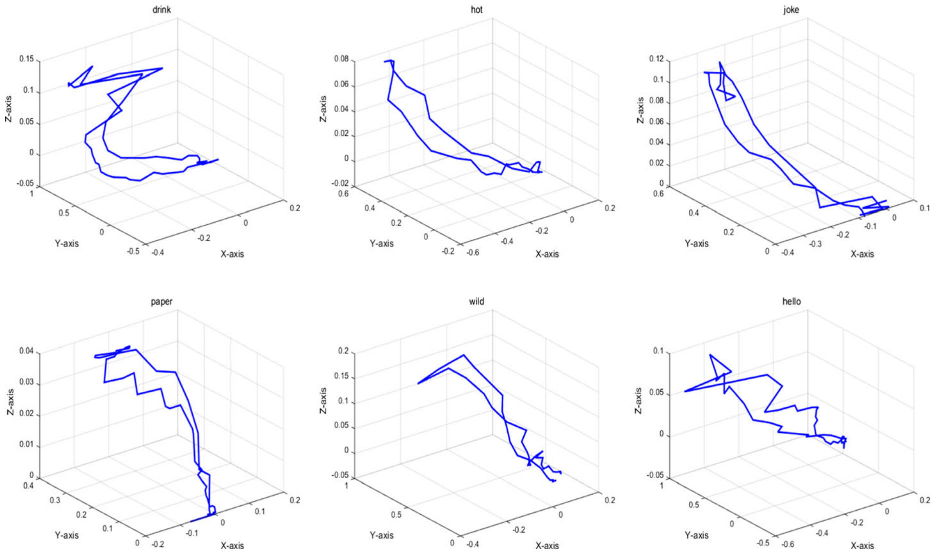


Fig. 2 3D gesture trajectories from ASL dataset

A parametric spline approximation [5] is implemented to smooth the trajectory curve. Then, we perform resampling so that the resampled trajectories are of equal length as shown in Fig. 3. In order to finding the optimized trajectory length for trajectory representation, in the experiment section, we employ many different normalized trajectories length for final classification. As a result, the optimization trajectory length can be approximately set at 70 sample points after experiment evaluation.

For the second step of pre-processing, we proposed a trajectory span distance normalization method, which transform each 3D trajectory to a common domain.

Given a set of resample points $\{x_k, y_k, z_k\}_{k=1}^{L'}$ from the previous step, normalized coordinates are given by

$$\left\{x_k', y_k', z_k'\right\}_{k=1}^{L'} = \left\{\frac{x_k}{d_{\max}}, \frac{y_k}{d_{\max}}, \frac{z_k}{d_{\max}}\right\}_{k=1}^{L'} \tag{3}$$

where $d_{\max} = \underset{\substack{i,j \in L' \\ i \neq j}}{\arg\max} \left\{\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2}\right\}$

In such a case, the distance of any two points is involved in the range between 0 and 1. As a result, the scale variations contained in the raw trajectory data can be effectively removed.

3.2 The review of shape context

In this section, we first review the shape context descriptor [2] which can be utilized in representing a object point by measuring the distribution of relative positions of neighboring points. Obviously, the full set of vectors used as shape descriptors contains global details since it configures the entire shape relative to the reference points. This set of vectors is identified as a highly discriminative descriptor which can represent the shape distribution over relative positions. The shape context of p_i is defined as a coarse histogram h_i of the relative coordinates of the remaining $n-1$ points:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\} \tag{4}$$

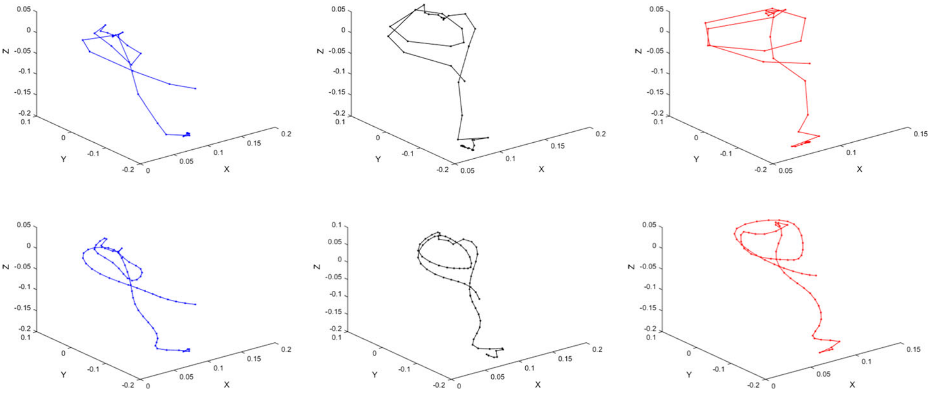


Fig. 3 Example of gesture trajectory normalization; the upper line shows the original trajectory with different points number; the bottom line shows the normalized trajectory with the same points number

The bins are uniform in log-polar space, making the descriptor more sensitive to the positions of nearby sample points than to those of points farther away. The histogram similarity of pair wise points (p_i, q_j) between two trajectories can be denoted as follow:

$$C_{ij} = \frac{1}{2}\sum_{k=1}^{K}\frac{\left[h_i(k)-h_j(k)\right]^2}{h_i(k)+h_j(k)} \tag{5}$$

where $h_i(k)$ and $h_j(k)$ denote the K-bin normalized histogram at p_i and q_i, respectively.

3.3 The adaptive 3D shape context descriptor

Based on shape context, the 3D shape context descriptor is very straightforward. In 2D shape context, a point histogram is built based on 2D log-polar coordinate system as shown in Fig. 4 (left).In 3D shape context, the pervious descriptor is extended to 3D space by building a point histogram based on 3D spherical coordinate system as shown in Fig. 4 (right). Denoting an origin on one of a trajectory point, the 3D shape context captures the 3D spatial and 1D temporal distribution of all other trajectory points around it. Along radial direction, bins are arranged uniformly in log-polar space which makes it more sensitive to positions of nearby points than to those of remaining points farther away. If there are i bins for the radius,j bins for azimuth and k bins for elevation, the 3D shape context is partitioned to $i \times j \times k$ bins in total.

The traditional coordinate representation denotes each position of trajectory points as x, y, z in Cartesian coordinate. It can be represented in the spherical coordinate as follow:

$$\begin{cases} x = r\sin\theta\cos\theta \\ y = r\sin\theta\sin\varphi \\ z = r\cos\theta \end{cases} \tag{6}$$

Generally, a 3D shape context descriptor can be described by five parameters: the number of $_\theta$ bins along the azimuth dimension, the number of log(r) bins along the radial dimensions, the number of φ bins along the elevation dimension, outer radius, and inner radius. Outer radius is the radius of the outer most circle in 3D SC ball, and the inner radius is the radius of the inner most circle in 3DSC ball. The (x, y, z) position of each point in the Cartesian coordinate can be converted to (r, θ, φ) in the spherical coordinate as follow:

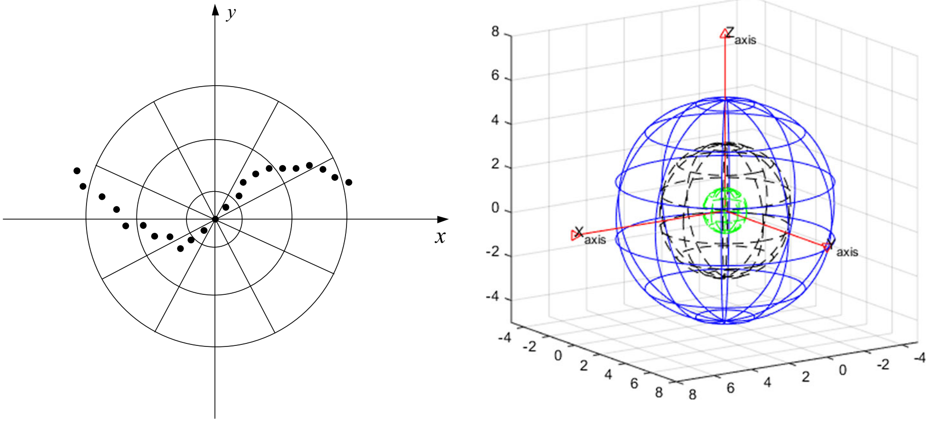


Fig. 4 The illustration of log-polar coordinates in 2D (left) and 3D (right) space

$$\begin{cases} r = \sqrt{x^2 + y^2 z^2} \\ \theta = arccos\left(\frac{z}{r}\right) \\ \varphi = actan\left(\frac{y}{x}\right) \end{cases} \tag{7}$$

Obviously, the outer radius controls the 3DSC ball volume. It restricts the number of trajectory points that involve in each 3D shape context. It also indicates the width of the time window from the perspective of time series. In previous works, there are two main kinds of strategies to determine the outer radius in 2D shape context: one is to set the outer radius to one tenth of the motion trajectories length, which means that at most 10 % of the points will be covered by each shape context. The other is compute the average distance of any two points in the motion trajectories and take it as the maximum size of the outer radius. In our work, instead of using 3DSC to extract the local information from each trajectory, the global information of the whole motion trajectory is concerned, as shown in Fig. 5.

In Eq. (7), we treat a origin in the spherical coordinate the same as a origin in the Cartesian coordinate. However, for the 3DSC feature extraction, each trajectory point should be treated as a reference point with origin translate from (0,0,0) to the current point position (x_n, y_n, z_n). Hence, for 3DSC descriptor representation, the parameter of each reference point can be expressed as:

$$\begin{cases} r_n = \sqrt{(x_r - x_n)^2 + (y_r - y_n)^2 + (z_r - z_n)^2} \\ \theta = \arccos\left(\frac{z_r - z_n}{r_n}\right) \\ \varphi = \arctan\left(\frac{y_r - y_n}{x_r - x_n}\right) \end{cases} \tag{8}$$

where (x_r, y_r, z_r) denote the all other trajectory points aside from the current point position (x_n, y_n, z_n). Consequently, for the *n-th* trajectory point (x_n, y_n, z_n) as the 3DSC descriptor's origin, the corresponding outer radius can be expressed as follow:

$$r_n^{\max} = \underset{\substack{(x_r, y_r, z_r) \in L' \cup \\ (x_r, y_r, z_r) \neq (x_n, y_n, z_n)}}{argmax} \left(\sqrt{(x_r - x_n)^2 + (y_r - y_n)^2 + (z_r - z_n)^2}\right) \tag{9}$$

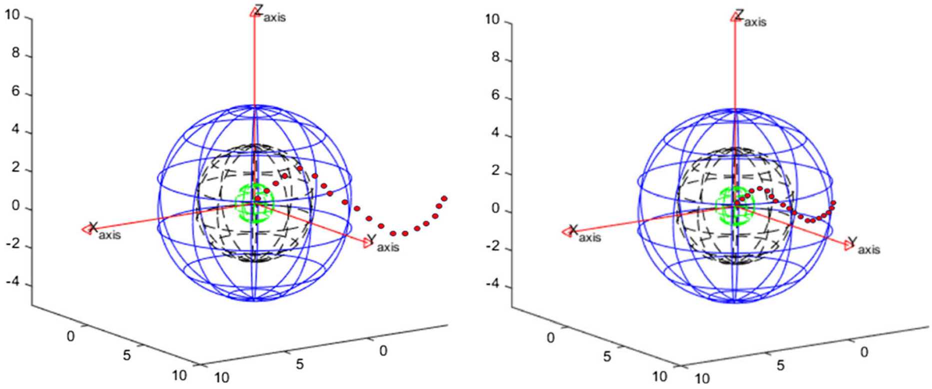


Fig. 5 Local 3D shape context (left) vs. Global 3D shape context (right)on gesture trajectory representation

As seen from above equation, the adaptive outer radius sets $\left\{r_n^{\max}\right\}_{n=1}^{L'}$ in a trajectory can be established by assembling each outer radius which generate from each trajectory point. We would like to point out that although different length of outer radius $\left\{r_n^{\max}\right\}_{n=1}^{L'}$ are yielded by setting different 3DSC ball origin, all the trajectory points can also be involved in each 3D shape context compactly.

In our trajectory representation strategy, each outer radius equals to the local maximum trajectory span distance which calculated according to the current 3DSC origin. In addition, a spherical grid is defined by means of subdivisions along the azimuth, elevation and radial dimensions. To account for generality, the number of subdivisions can be different along each dimension. In our experiments, the azimuth and elevation dimensions are equally divided into 12 and 8 spaces respectively. Typically, the outer radius of $\left\{r_n^{\max}\right\}_{n=1}^{L'}$ in each trajectory is 2^k times larger than the inner radius. Hence, the radial dimension is logarithmically divided into 5 spaces, which means $k = 5$. For the 3DSC histogram computation, each bin accumulates a weighted sum of the trajectory points number falling thereby.

The benefit of using adaptive outer radius mechanism is that it makes it possible for generating the global information which can increase the diversity of pairwise points during distance calculation. Furthermore, the global 3D shape context can give a better discrimination for matching a motion trajectory with its sub trajectory. In this case, the pairwise points with global information give relatively higher matching score for the same meaning trajectories, and reversely give relatively lower matching score for different meaning trajectories but which partially has the same shape appearance from one to another.

From a shape point of view, trajectories of the same class could be seen as similar shapes but with small non-rigid shape deformations, as shown in Fig. 6. 3D Shape contexts are extremely rich descriptors in that they can give appropriate tolerances for these trivial deformations, meanwhile, are only sensitive to those discriminative deformations. In contrast, trajectories of different classes always contain deformations large enough to be grasped by 3D shape contexts.

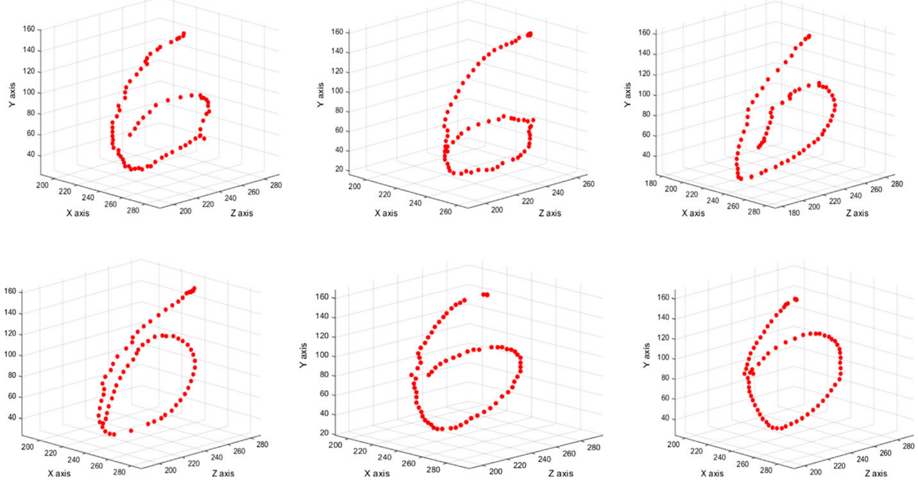


Fig. 6 Gesture digital "6" under shape deformations perform by different signers

4 Gesture trajectory alignment

4.1 Dynamic time warping

In time series analysis, dynamic time warping (DTW) is a well-established algorithm for comparing temporal sequences which may vary in time or speed. DTW addresses the main problem of aligning two sequences in order to get the most suitable distance measure of their overall difference. Compared with Euclidean distance, DTW can overcome the time distortion problem by finding a time-flexible alignment between two given time series, where the total cumulative distance is minimized. Each point of the time series is aligned to at least one point of another time series.

More specifically, suppose $\mathbf{X} = \{x_1, x_2, \cdots x_m\} \in \mathbf{R}^m$ and $\mathbf{Y} = \{y_1, y_2, \cdots y_n\} \in \mathbf{R}^n$ denote two time series with length $_m$ and n, respectively. To align two sequences using DTW, an m by n matrix is construct. The value of the (i^{th},j^{th}) cell of the matrix is the base distance between the two feature vectors x_i and y_i, namely $\delta(x_i, y_i)$.

A warping path W defines an alignment between X and Y can be formally written as $W = w_1, \cdots, w_T$, where $\max(m, n) \leq T \leq m + n - 1$. Each $w_t = (i,j)$ specifies that feature vector x_i of the X sequence is matched with Y feature vector y_i. The warping path is typically subject to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$;

Temporal continuity: Given $w_t = (a, b)$, and $w_{t-1} = (a', b')$, then $a - a' \leq 1$ and $b - b' \leq 1$;

Temporal monotonicity: Given $w_t = (a, b)$ and $w_{t-1} = (a', b')$, where $a - a' \geq 0$ and $b - b' \geq 0$.

From the point of view of above restrictions, an exponential number of warping paths can be found; however, DTW computes the optimal path that will minimize the following warping cost:

$$DTW(\mathbf{X}, \mathbf{Y}) = \min\left\{\sum_{t=1}^{T} \delta(w_k)\right\} \tag{10}$$

To find the optimized path, DTW can be recursively calculated using dynamic programming which computes the cumulative distance $DTW(i,j)$ with the distance $\delta(x_i, y_i)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements as the follow:

$$DTW(i,j) = \delta(x_i, y_i) + \min\left\{DTW(i-1, j-1), DTW(i, j-1), DTW\left(i-1, j\right\}\right. \tag{11}$$

In this way, we can find the best warping path W^* and the global matching score D^* by back tracing the cumulative distance matrix.

4.2 Using 3D shape context in DTW

The standard dynamic time warping typically using successive sequence locations as the trajectory feature for the cost matrix computation. In our proposed method, each element of the cost matrix is acquired by computing the histogram similarity between two 3D shape context. The new base distance between each pair of points can be defined as:

$$\delta_{3D\text{-}SC}(p, q) \equiv C_{pq} \tag{12}$$

where C_{pq} is defined in Eq. (5); The δ_{3D-SC} can be substitute $\delta(\cdot)$ for computing DTW.

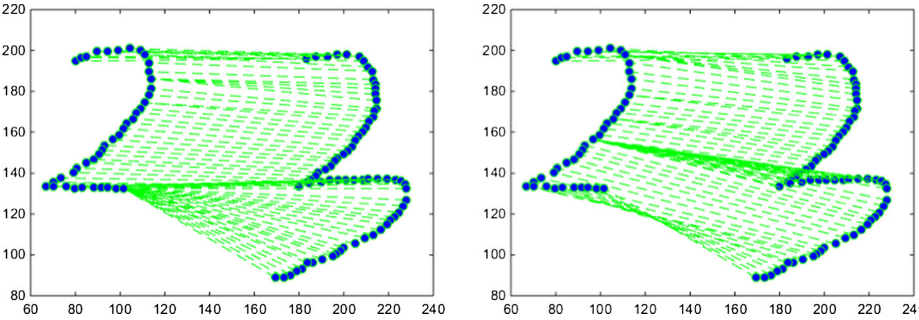
One thing must be clarified is that different from SC-DTW [34] which uses shape context to generate the alignment, our 3DSC-DTW consider to use the matrix cost of global 3D shape context feature rather than original Euclidean distance as the cumulative value. The merit is that it is greatly invariant to the trajectory translation and scale.

One of the significant reason in combination of the global 3DSC feature and DTW is that it can deal with the sub-trajectories problems. Without lose of generality, we take the digital gesture in the 2D space as an example. As we can see in Fig. 7, the digital gesture “2” can be treated as the sub-gesture of the digital gesture “3”. By using the local 3DSC feature, the alignment of digital“2” and the partial of digital“3”,as shown in Fig. 7a, coincide with the alignments of digital “2”. Only the end point of digital “2” is left to match the rest points of digital “3”. Toward this end, the final decision score may be relatively lower and can readily cause the miss classification. On the contrary, the alignments of digital “2”and“3” in Fig 7b, which make use of the global 3DSC feature, can achieve relatively higher matching score, hence has strong discrimination.

Another reason for embedding the 3DSC descriptor into DTW is that it can greatly resist to trajectory appearance perturbation. Unlike pose models, trajectory data encompasses a notion of time flow. Even trajectories have the same appearances, they may represent different meanings due to the different directions of time flow. Typically, all the gesture trajectories no matter for training or testing should be captured under a fixed canonical coordinate frame. However, the rotation of trajectory should be considered in two situations:(1) all the gestures are captured under the fixed canonical coordinate frame. In this case, as we mentioned before, even the gesture trajectories may have slightly appearance difference or axis inclination, the 3D shape context descriptor is insensitive to such deformations, and can greatly eliminate the presence of noise. (2) Gestures are captured under different coordinate frame. In this case, we should transform the trajectory into a canonical coordinate frame according to the translation and scale parameters. Otherwise, it is hard to determine whether two gestures have the same meanings or not even they have similar shape.

4.3 Time complexity

Suppose P denotes the number of bins in a 3D shape context histogram. The time complexity of computing a gesture point histogram is $P = r * a * b$, where r is the number of radial bins, a



(a) Alignments with the local 3D-SC descriptor (b) Alignments with the global 3D-SC descriptor

Fig. 7 Alignments of two digital gestures with global (left) and local (right) 3D shape context descriptor. **a** Alignments with the local 3D-SC descriptor **b** Alignments with the global 3D-SC descriptor

is the number of θ angular bins, b is the number of φ angular bins. Recall that m and n represent the lengths of two time series after gesture normalization. DTW has to consider all cells in the warping matrix; thus, it has a time complexity of $O(m \cdot n)$. In 3DSC-DTW, the base distance is the difference between two histograms instead of two real numbers. Hence, SC-DTW has a time complexity of $O(P \cdot m \cdot n)$.

5 Experiments

In this section, we conduct a series of experiments to evaluate the proposed method. Experiments are conducted on three types of datasets: two types of Australian Sign Language dataset(compact and large) from UCI KDD archive [24] and the 3D hand digital dataset [7]. The compact ASL trajectory dataset consists of 95 sign classes (words), and 27 samples were captured for each sign. The large ASL dataset also contain 95 signs examples. Each sign has 70 examples and with 6650 sign samples in total.

For ASL datasets(compact and large) evaluation, we first utilize the compact ASL dataset to investigate the optimization of trajectory normalization length. Secondly, the benefits of using adaptive outer radius and scale invariance in trajectory classification are implemented, and then we compare our results to the state-of-the-art methods. The trajectory recognition performance under varying training size is also tested based on the large ASL database. Finally, we made a evaluation on the 3D hand digital dataset to test the discrimination capacity between sub-trajectory and full-trajectory as well as the robustness of proposed method.

5.1 The benefit of using trajectory length normalization

The propose of this experiment is to evaluate the impact of various of trajectories normalization in the performance of 3DSC based trajectory classification technique. As shown in Table 1, the classification accuracy with normalized trajectories length are overall higher than original trajectories length. Thus, the effectiveness of using normalized trajectory can be verified. The best results were obtained when the normalization length approximately equals to 70 sample points. This result may suggests the best normalization length should be fixed neither too larger nor too small. Consequently, we choose each trajectories equals to 70 sample points as the optimal normalization length.

Table 1 Classification accuracy with varying the sample length of motion trajectories

Classes	Original Length	Normalization Length										
		10	20	30	40	50	60	70	80	90	100	110
2	93.83	96.30	97.53	98.15	97.53	98.46	97.53	**98.62**	98.46	97.22	97.53	97.53
4	82.41	91.05	94.29	95.83	95.99	96.14	**96.91**	95.68	95.83	95.6 8	96.30	95.52
8	69.52	81.64	88.66	89.27	89.35	90.12	90.51	**91.43**	90.12	88.66	89.74	88.97
10	68.70	80.06	87.04	88.07	87.16	89.04	89.59	**90.14**	89.04	87.28	87.59	87.04

5.2 The benefit of using adaptive outer radius

In this section, the compact ASL dataset are utilized and a 9-fold-cross validation was conducted for trajectory classification by varying the scale of adaptive outer radius. One of ninth trajectories from each category serve as testing samples and the others serve as training samples. We repeated this test 7 times. After 7 round evaluations, the average classification rate is computed for the final comparison.

The propose of this experiment is to demonstrate the advantage of using adaptive outer radius. Figure 4 shows the classification accuracy under varying the scale of adaptive outer radius. The scaled adaptive outer radius is defined as follow:

$$\left\{r_n{'}\right\}_{n=1}^{L'} = \kappa\cdot\left\{r_n^{\max}\right\}_{n=1}^{L'} \tag{13}$$

where k is the scale factor which control the size of the 3D shape context;

The x axis in the Fig 8 represents the scale factor from 0.1 to 1, with step 0.1. For example, 0.1 means we set outer radius of each 3D shape context equals to 10 % of the corresponding local maximum trajectories span distance. We consider that the scale factor from 0.1 to 0.9 generate the local feature, otherwise it generate the global feature. With the increasing of the outer radius, the classification accuracy of 4, 8, and 10 classes are gradually getting higher. All these three types of classes reach the maximum classification rate 95.68 %, 91.43 %, 90.14 % respectively under the scale factor equals to 1, which means all the gesture points are involved in each 3D shape context ball volume and the global 3DSC descriptor are generated for histogram distribution computing.

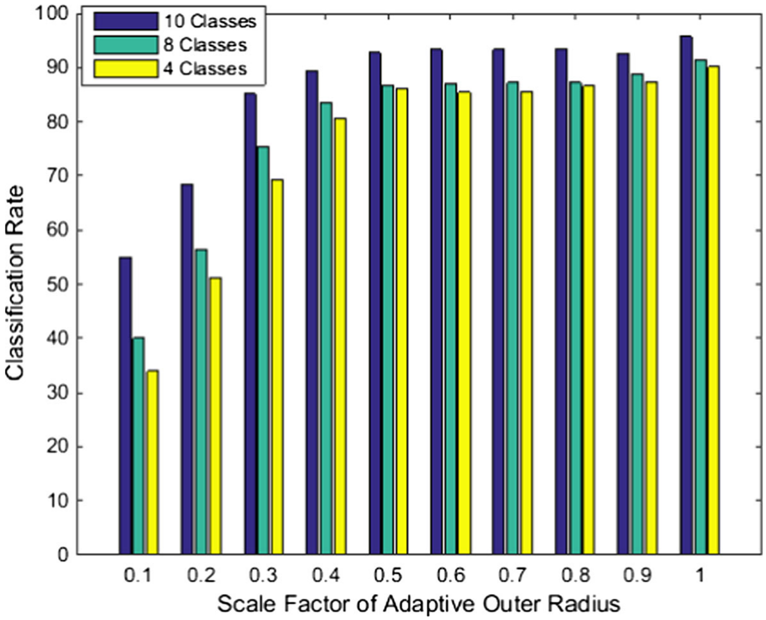
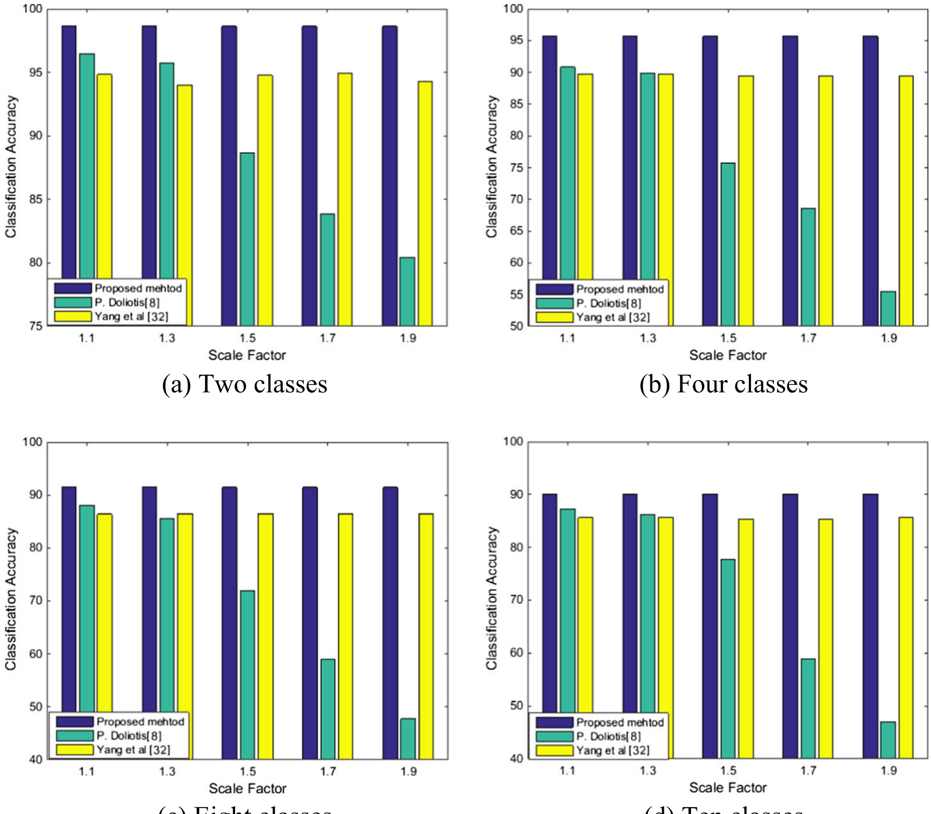


Fig. 8 Classification rate under varying adaptive outer radius

5.3 Scale invariance performance

In this part, we evaluate the performance of classification accuracy by changing trajectory scale and translation in 2, 4, 8, and 10 classes respectively. As shown in Figs. 9 and 10, a considerable improvement of our proposed method on scale invariance can be obviously seen. To apply a certain amount of scaling to the input gestures, we multiply the *x* and *y* coordinates of each trajectory point by a set of small increments ([1.1, 1.3, 1.5, 1.7, 1.9]). To apply a certain amount of translation to the input trajectories, we add a set of small increments ([0.01, 0.03, 0.05, 0.07, 0.09]) in meters to the *x* and *y* coordinates of the position of each gesture point. With gradually increasing scale and translation factors, the classification rate of Euclidean Distance based DTW method [8] rapidly dropped, however, the proposed method and Mix signature method [32] still remain a stable accuracy. Moreover, the performances of our method outperform Mix signature method. This advantage owes to the 3D shape context descriptor with adaptive outer radius can automatically represent the global information for each pairwise points as a robust similar measure.



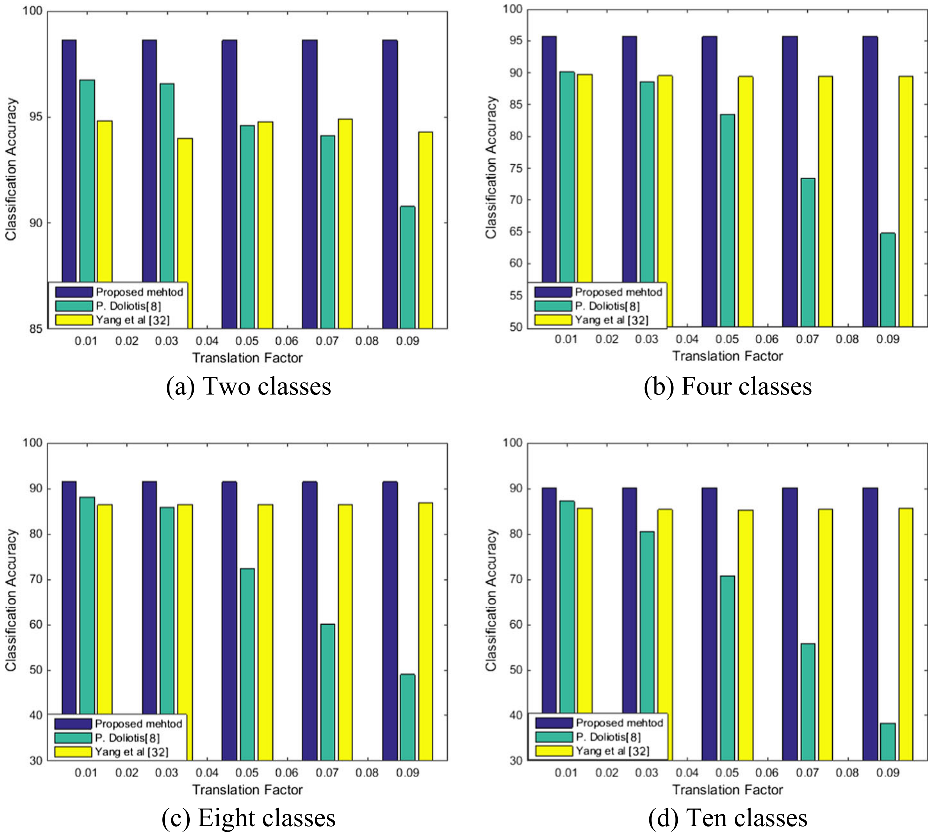
(a) Two classes

(b) Four classes

(c) Eight classes

(d) Ten classes

Fig. 9 Classification accuracy vs. gesture scale: **a** Two classes, **b** Four classes, **c** Eight classes, **d** Ten classes



(a) Two classes

(b) Four classes

(c) Eight classes

(d) Ten classes

Fig. 10 Classification accuracy vs. gesture translation: **a** Two classes, **b** Four classes, **c** Eight classes, **d** Ten classes

5.4 Comparison with other methods

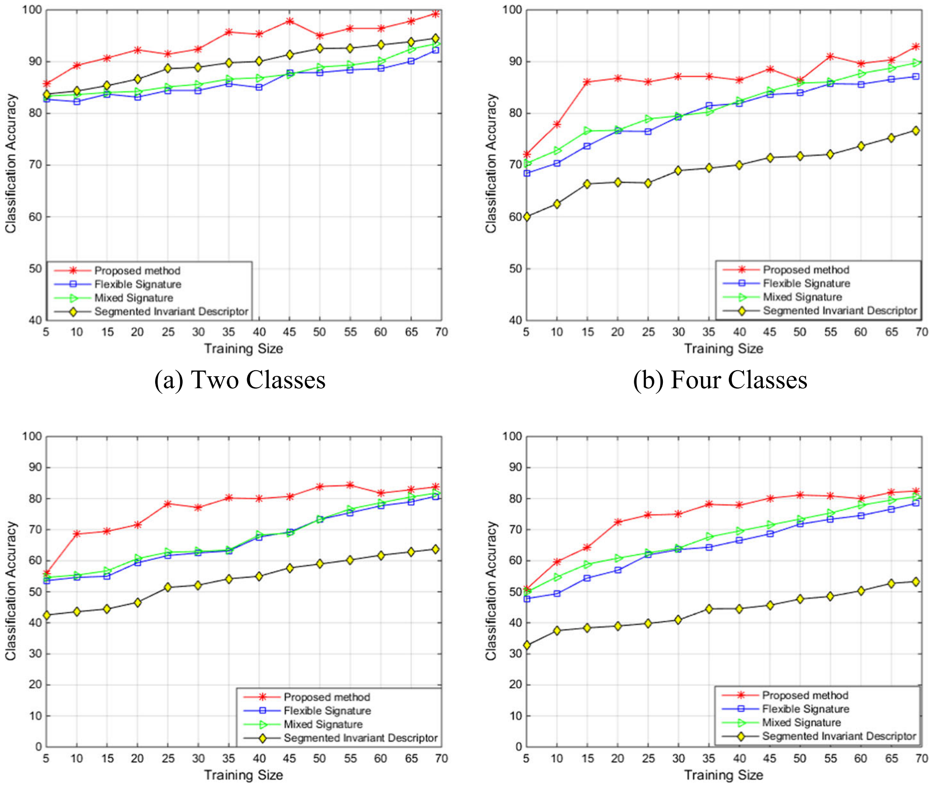
In the third experiment, the compact ASL dataset is used to evaluate the trajectory recognition performance. Since this database and the chosen classes were used in [26, 31, 32] for experiments, we implement experiment on the equivalent situation for comparison. Consider that trajectory recognition also relies on efficient recognition engine, we test the performance of the adaptive 3DSC descriptor by utilizing another two recognition engines as well, which is support vector machine and Lock-step measure. Lock-step measure means a one-to-one correspondence matching between time series as they compare *i*-th point of one time series to *i*-th point of another time series. According to the experimental results represented in Table 2, we can observe that: 1) For all of the approaches, the proposed method achieves the highest recognition rate in matching within 2, 4, 8 and 10 classes respectively. 2)The performance of alignment based method DTW [8, 26, 32], Euclidean distance(Solution 1) outperform the discriminate methods SVM(Solution 2). 3)As the number of classes increase, the recognition ratio of all method gradually dropped, however, the decline rate from class 10 to class 2 of our proposed method is 8.48 %, less than that of all other methods. That is to say, our method is more flexible for multiclass recognition.

Table 2 Comparison with the *state-of-the-art* methods on compact ASL dataset

Method	Description	Classes Accuracy			
		2	4	8	10
Wu et al. [26]	Flexible signature + DTW + 1NN	92.54	87.11	83.52	80.06
Yang et al. [32]	Mixed Signature + DTW + 1NN	94.83	89.74	86.36	85.64
Yang et al. [31]	Segmented Invariant Descriptor	95.04	76.73	63.48	53.96
P. Doliotis [8]	Euclidean Distance + DTW + 1NN	97.46	91.83	89.09	88.28
Solution 1	Adaptive 3DSC + Lock Step	97.75	89.20	85.78	83.63
Solution 2	Adaptive 3DSC + SVM	82.47	75.15	61.81	49.94
Solution 3	Adaptive 3DSC + DTW + 1NN	**98.62**	**95.68**	**91.43**	**90.14**

5.5 Recognition with varying training size

This section utilized the large ASL dataset to evaluate trajectory recognition performance under varying training size. In this experiment, for fair comparison, we follow to extract 2,4,8,10 classes in [26, 31, 32] to evaluate the classification accuracy. Since the dataset has 70



(a) Two Classes

(b) Four Classes

(c) Eight Classes

(d) Ten Classes

Fig. 11 The classification accuracy versus training size on Large ASL dataset

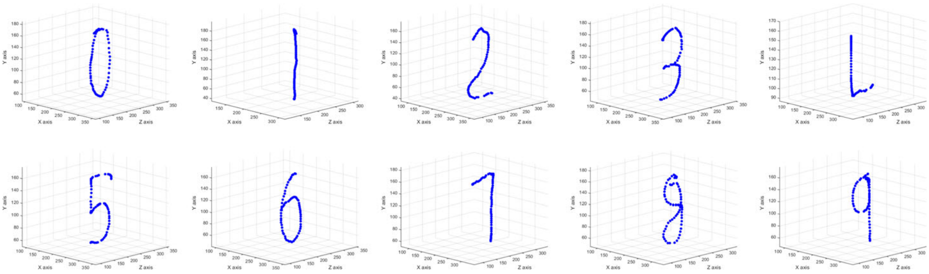


Fig. 12 The visualization of the 3D Hand-Signed Digit

examples for each sign, we decide to take leave one out cross validation strategy. During evaluation, we successively select each gesture example from each category as testing data, and treated the remaining examples of each category as training data. For the propose of examining the relationship between training size and recognition rate. We randomly extracted a certain number of examples from the remain training data. As shown in Fig 11, with gradually increasing the training size, the classification accuracies of 2, 4, 8, 10 classes improved significantly and achieve maximum with using 69 training samples. Also, we can observe that the proposed method outperforms all other state-of-the-art methods, which indicates that our proposed method is also suitable for large datasets. Moreover, we also test other amounts of classes, due to the space limitation we did not show here. Nevertheless, the tendency of classification accuracies, in general, remain the same.

5.6 Results on 3D HSD dataset

In this section, the experiments are conducted on the 3D Hand-Signed Digit which can be visualized as shown in Fig. 12. This hand gesture datasets is a commonly used benchmark for

Table 3 Comparison of Recognition Rate by other state-of-the-art methods

Method	Description	Accuracy(%)
Wu et al. [26]	Flexible signature + DTW + 1NN	86.6
Yang et al. [32]	Mixed Signature + DTW + 1NN	92.5
Yang et al. [31]	Segmented Invariant Descriptor	62.7
P. Doliotis [8]	Euclidean distance + DTW + 1NN	88.6
Solution 1	Adaptive 3DSC + Lock Step	90.2
Solution 2	Adaptive 3DSC + SVM	66.8
Solution 3	Adaptive 3DSC + DTW + 1NN	**98.4**

Table 4 Sub-gesture table for Hand Signed digit from “0” to “9”

Sub-gesture	Super-gestures
“1”	{“7”,“9”}
“2”	{“3”}
“5”	{“8”}
“7”	{“2”,“3”}

Table 5 The missclassification numbers between subgesture "1" and the corresponding supergestures

		“7”	“9”	Others
“1”	Scale = 0.5	2	1	1
	Scale =1.0	1	1	0
	Total	5	1	

gesture recognition with 10 categories performed by 12 different people. In training examples, 300 digit exemplars with 30 per class were stored in the database. In test examples, 440 digit exemplars with 44 per class were captured.

Table 2 illustrates.

Table 3 illustrates the performance comparison of different algorithms by using 300 training data and 440 testing data. As expected, our proposed method(Solution 3) yield a higher recognition rate than other methods. It is worth noting that our proposed adaptive 3DSC descriptor achieve a relatively higher performance when it combine with Lock-step measure (Solution 2).

In Table 4, we manually define the full meaning gestures and the corresponding sub-gesture for the hand signed digits recognition. From Table 4, we can see that gesture “1”,“2”,“5”,“7” can be defined as the sub-gesture of {“7”,“9”},{“3”},{“8”},{“2”,“3”} respectively.

In the following tables, we examine the miss matching numbers between the super-gestures and the corresponding sub-gestures. For simplicity, we choose outer radius scale factor equals to 0.5 to represent the local 3DSC descriptor. As we can see from the following Tables 5, 6, 7 and 8, the total misclassification numbers that caused by sub-gestures are larger than other gestures. That explains why the proposed global 3DSC representation can be effective on decreasing misclassification and restraining the ambiguity among partial similar gestures.

Table 6 The missclassification numbers between subgesture "2" and the corresponding super-gesture

		“3”	others
“2”	Scale =0.5	3	1
	Scale =1.0	0	1
	Total	3	2

Table 7 The missclassification numbers between subgesture "5" and the corresponding super-gesture

		“8”	Others
“5”	Scale =0.5	3	1
	Scale =1.0	1	0
	Total	4	1

Table 8 The missclassification numbers between subgesture "7" and the corresponding super-gestures

		“2”	“3”	Others
“7”	Scale =0.5	2	1	1
	Scale =1.0	1	0	1
	Total	4	2	

6 Conclusions and future work

In this paper, we present a novel motion trajectory classification method in the spatiotemporal domain. An invariant descriptor - 3D shape context with adaptive outer radius is presented. This descriptor able to flexible extract rich global shape context information for motion trajectory representation. An effective alignment algorithm based on Dynamic Time Warping which replaces the raw distance feature by 3D shape context descriptor is proposed for calculating the matching similarity. We compare the classifying performance with our proposed descriptor to the previous descriptors in the three benchmark datasets. The experiments results show that the proposed method achieves the state-of-the-art performance in both accuracy and efficiency for motion trajectory classification in the spatiotemporal domain.

There are still several future tasks to improve our current work. It is in urgent need to establish a real-time motion trajectory recognition or classification system to automatically segment the motion trajectory by detecting the start and end frame. In addition, how to recognize the motion trajectory from different viewpoint and how to apply the proposed 3D motion trajectory strategy to various application, such as activity recognition, anomaly detection, video surveillance is still worth study.

Acknowledgments This work is supported by National Natural Science Foundation of China (Grant no. 61501456), "Light of West China" Program of Chinese Academy of Sciences.

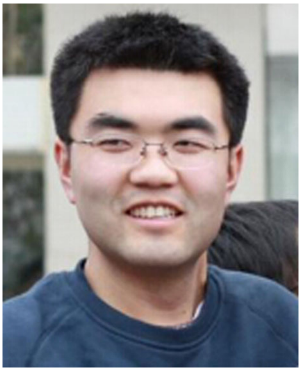
References

1. Bashir FI, Khokhar AA, Schonfeld D (2007) Real-time motion trajectory-based indexing and retrieval of video sequences[J]. IEEE Trans Multimed 9(1):58–65
2. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
3. Chen L, Yao H, Sun X (2012) Action retrieval based on generalized dynamic depth data matching[C]// Visual Communications and Image Processing. 1–4
4. Cheng H, Dai Z, Liu Z (2013) Image-to-Class Dynamic Time Warping for 3D hand gesture recognition[C]// Multimedia and Expo (ICME), 2013 I.E. International Conference on. IEEE, 1–6
5. DEBOOR C, Fix GJ (1973) Spline approximation by quasiinterpolants[J]. Journal of Approximation Theory 8(1):19–45
6. Ding H, Trajcevski G, Scheuermann P et al (2008) Querying and mining of time series data: experimental comparison of representations and distance measures[J]. Proc Vldb Endowment 1(2):1542–1552
7. Doliotis P (2013) Viewpoint invariant gesture recognition and 3d hand pose estimation using RGB-D[J]
8. Doliotis P, Stefan A, McMurrough C. et al (2011) "Comparing gesture recognition accuracy using color and depth information," in Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments, pp. 20–22,ACM
9. Frome A, Huber D, Kolluri R et al (2004) Recognizing objects in range data using regional point descriptors.[J]. Lect Notes Comput Sci 3023:224–237
10. Grundmann M, Meier F, Essa I (2008) 3D Shape Context and Distance Transform for action recognition[C]// Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 1–4
11. Kim I C, Chien S I. Analysis of 3D Hand Trajectory Gestures Using Stroke-Based Composite Hidden Markov Models.[J]. Applied Intelligence, 2001, 15(2):131-143
12. Jonathan A, Athitsos V, Yuan Q, Sclaroff S (2009) A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Trans Pattern Anal Mach Intell 31(9):1685–1699
13. Kaya H, Gündüz-Öğüdücü Ş (2015) A distance based time series classification framework[J]. Inf Syst 51: 27–42

14. Li Z, Horain P, Pez A, et al (2009) Statistical Gesture Models for 3D Motion Capture from a Library of Gestures with Variants[C]// International Conference on Gesture in Embodied Communication and Human-Computer Interaction. Springer-Verlag, 219–230
15. Liang B, Zheng L (2015) A Survey on Human Action Recognition Using Depth Sensors[C]// International Conference on Digital Image Computing: Techniques and Applications. IEEE, 1019–1029
16. Lin WY, Hsieh CY (2013) Kernel-based representation for 2D/3D motion trajectory retrieval and classification[J]. Pattern Recogn 46(3):662–670
17. Liu W, Fan Y, Li Z et al (2015) RGBD video based human hand trajectory tracking and gesture recognition system [J]. Math Probl Eng 2015:1–15
18. Liu W, Fan Y, Lei T, Zhang Z (2014) Human gesture recognition using orientation segmentation feature on random Forest. In Proceedings of IEEE China Summit & International Conference on Signal and Information Processing (SIP'14), pp. 480–484, Xi'an, China, July
19. Lu G, Zhou Y, Li X et al (2016) Efficient action recognition via local position offset of 3D skeletal body joints[J]. Multimed Tools & Appl 75(6):3479–3494
20. Niu W, Long J, Han D et al (2004) Human activity detection and recognition for video surveillance[C]. IEEE Int Conf Multimed Expo 1:719–722
21. Psarrou A, Gong S, Walter M (2002) Recognition of human gestures and behavior based on motion trajectories. Image Vis Comput 20(5–6):349–358
22. Suzuki N, Hirasawa K, Tanaka K, et al (2010) Learning motion patterns and anomaly detection by Human trajectory analysis[C]// IEEE International Conference on Systems, Man and Cybernetics. IEEE, 498–503
23. Tombari F, Salti S, Stefano LD (2010) Unique shape context for 3d data description[C]// Proceedings of the ACM workshop on 3D object retrieval. ACM, 57–62
24. UCI KDD ASL Archive, http://kdd.ics.uci.edu/databases/auslan2/auslan.html
25. Vlachos M, Gunopulos D, Das G (2004) Rotation invariant distance measures for trajectories[J]. Proceedings of Sigkdd, 707-712
26. Wu SD, Li YF (2009) Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition. Pattern Recogn 42(1):194–214
27. Xiao D, Zahra D, Bourgeat P et al (2010) An improved 3D shape context based non-rigid registration method and its application to small animal skeletons registration [J]. Comput Med Imaging & Graphics Off J Comput Med Imaging Soc 34(4):321–32
28. Xing Z, Pei J, Keogh E (2010) A brief survey on sequence classification[J]. Acm Sigkdd Explor Newslett 12(1):40–48
29. Yang JY, Li YF and Wang KY (2010) "Mixed signature descriptor with global invariants for 3D motion trajectory perception and recognition," in Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, pp. 1952–1956
30. Yang J, Li Y F, Wang K (2010) A new descriptor for 3D trajectory recognition via modified CDTW[C]// Automation and Logistics (ICAL), 2010 I.E. International Conference on. IEEE, 37–42.
31. Yang J, Li Y F, Wang K (2011) Invariant trajectory indexing for real time 3D motion recognition[C]// IEEE/RSJ International Conference on Intelligent Robots & Systems IEEE/RSJ International Conference on Intelligent Robots & Systems, 3440–3445
32. Yang J et al (2012) Mixed signature: an invariant descriptor for 3D motion trajectory perception and recognition [J]. Math Probl Eng 2012(1):488–488
33. Yuting Su1, Haiyi Wang1, Peiguang Jing1, Chuanzhong Xu. A spatial-temporal iterative tensor decomposition technique for action and gesture recognition[J]
34. Zhang Z, Tang P, Duan R (2015) Dynamic time warping under pointwise shape context[J]. Inf Sci 315:88–101
35. Zhao S, Chen L, Yao H et al (2015) Strategy for dynamic 3D depth data matching towards robust action retrieval[J]. Neurocomputing 151:533–543

Weihua Liu received his PhD degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, in 2015. He is currently a Researcher in Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China. His current research interests include image processing, computer vision and pattern recognition.

Zuhe Li received the B.S. degree in electronic information science and technology from Zhengzhou University of Light Industry, Zhengzhou, China, in 2004, and the M.S. degree in communication and information system from Huazhong University of Science and Technology, Wuhan, China, in 2008.He is currently pursuing the Ph.D. degree at the Northwestern Polytechnical University, Xi'an, China. His major research interests include Computer Vision and Machine Learning.



Geng Zhang is now a Researcher in Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China. He received his PhD, master and bachelor degrees in Electronic Informationin The Xi'an JiaotongUniversity in 2013, 2010 and 2007 respectively. His research interest is image processing and analysis.

Zhong Zhang received the B.S. degree in Chongqing University, China, in 2007, and the M.S. degree in Wuhan University, China, in 2009 and received his PhD in the university of Texas at Arlington in 2015. His research interest is computer vision, machine learning, gesture recognition, fall detection.