Research Article

Chinese Sign Language Recognition Based on DTW-Distance-Mapping Features

Juan Cheng ¹, Fulin Wei ², Yu Liu ¹, Chang Li ¹, Qiang Chen ¹ and Xun Chen ³

¹Department of Biomedical Engineering, Hefei University of Technology, Hefei, Anhui 230009, China

²Department of Artificial Intelligence, Beijing Normal University, Beijing 100875, China

³Department of Electronic Engineering & Information Science, University of Science and Technology of China, Hefei 230026, China

Correspondence should be addressed to Yu Liu; yuliu@hfut.edu.cn

Received 17 December 2019; Revised 1 July 2020; Accepted 9 July 2020; Published 6 August 2020

Academic Editor: Wanquan Liu

Copyright © 2020 Juan Cheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sign language is an important communication tool between the deaf and the external world. As the number of the Chinese deaf accounts for 15% of the world, it is highly urgent to develop a Chinese sign language recognition (CSLR) system. Recently, a novel phonology- and radical-coded CSL, taking advantages of a limited and constant number of coded gestures, has been preliminarily verified to be feasible for practical CSLR systems. The keynote of this version of CSL is that the same coded gesture performed in different orientations has different meanings. In this paper, we mainly propose a novel two-stage feature representation method to effectively characterize the CSL gestures. First, an orientation-sensitive feature is extracted regarding the distances between the palm center and the key points of the hand contour. Second, the extracted features are transformed by a dynamic time warping- (DTW-) based feature mapping approach for better representation. Experimental results demonstrate the effectiveness of the proposed feature extraction and mapping approaches. The averaged classification accuracy of all the 39 types of CSL gestures acquired from 11 subjects exceeds 93% for all the adopted classifiers, achieving significant improvement compared to the scheme without DTW-distance-mapping.

1. Introduction

Sign language (SL), a structured form of hand gestures, is the combination of various signs, hand movements, or body and facial expressions to deliver information. The motivation of SL recognition (SLR) is to provide a translation system to bridge the communication between the deaf and the healthy hearing society. From the literatures, cameras, data gloves, accelerometers (ACCs), and surface electromyography (sEMG) sensors are some common acquisition devices to capture gesture signals [1–3]. Currently, many SLs have been studied worldwide for the purpose of recognition, typically as American SLs, Australian SLs, Arabic SLs, Greek SLs, Korean SLs, and Chinese SLs (CSLs) [4–7]. According to the global estimates on prevalence of hearing loss provided by WHO, about 360 million people of such kind are in the world, among whom approximately one-fifth are Chinese [8]. Thereby, it is urgent to develop a practical and effective CSL recognition (CSLR) framework.

In the past decades, researchers have proposed various solutions to improve the performance of CSLR [9, 10]. For instance, Fang et al. employed data-gloves as input devices and designed a CSLR system based on a hierarchical decision tree, which could recognize 5113-isolated-sign vocabulary with accuracy rates of 91.6% and 83.7% for the registered and unregistered sets, respectively [2]. Li et al. [3] proposed an automatic component-level CSLR framework using the combined information from both ACC and sEMG signals. The overall classification results were 96.5% and 86.7% for a vocabulary of 120 signs and for 200 sentences, respectively. Later, Ma et al. employed a probabilistic model, hidden conditional random field (HCRF), to recognize CSLs based on sEMG and ACC signals. The average recognition rate was 91.51% for 120 high-frequency CSL words [11]. In 2015, Zhang et al. proposed a new system for both isolated and continuous CSLR based on video data. For isolated CSLR, the histogram of oriented displacement (HOD) was used to

describe the trajectories, while the multi-SVM was adopted for classification. As for continuous CSLR, a dynamic programming method with warping templates obtained by DTW was adopted. The averaged accuracies of 450 phrases and 180 sentences were 88.0% and 85.2%, better than those obtained by HMMs, as well as CRF and its advanced techniques [12]. By considering the typical posture and motion simultaneously, Wang et al. proposed a sparse observation representation approach for CSLR, yielding better performance than traditional DTW and HMMs [13]. Besides the CSLR systems with traditional machine learning frameworks, there are a lot of studies employing advanced deep learning techniques [9, 14–16]. For instance, Yang and Zhu proposed a CSLR system based on convolutional neural networks using video data [15]. In [9], Huang et al. delivered a novel sequence-to-sequence learning method based on keyframe centered clips (KCCs) for isolated CSLR. The CSL sequence learning was realized by establishing an encoder-decoder using a long short-term memory (LSTM) network, achieving an average accuracy of 91.18% for 310 CSL words. Xiao et al. adopted a dual LSTM combined with a couple conditional HMM to fuse hand and skeleton sequence information to recognize continuous CSLs using RGB-D data [16].

Currently, the sign gestures employed by CSLR systems generally refer to the handbook edited by China Association of the Deaf [17], which has about 5000 normalized sign gestures. However, since the number of frequently used Chinese characters is about 3500, the number of frequently used words is far more than tens of thousands. It is obvious that these gestures are not enough for practical applications. Besides, most of these sign gestures are symbolic and consist of complex components, which are sometimes hard to be standardized due to regional differences and personal habits. To overcome these limitations, a completely new version of phonology- and radical-coded CSL execution has been introduced by the same association. In this version, each Chinese character can be expressed by executing the coded gestures twice using both hands. The first execution of phonology-coded gesture is near the mouth or the chest, representing Mandarin Pinyin. The second execution of radical-coded gesture is near the waist, representing the fore and the end radicals. In this way, almost every Chinese character can be composed of the only 13 basic gesture shapes [17, 18], with each having 3 orientations (totally 39 gestures), shown in Figure 1. The defined 13 gesture shapes are shown in Figure 1(a) and named after four capital letters. The 3 orientations are shown in Figure 1(b) and abbreviated as U, H, and D, corresponding to vertical upwards, horizontal, and vertical downwards, respectively.

The coded rules to express a Chinese character can be seen from Table 1 in [18]. Taking the Chinese character “huang” (yellow) for instance, the pronunciation is formed by the fast sounding of basic initial consonant code “h” (right hand, “ASLB” with a horizontal orientation) and the basic final consonant code “uang” (left hand, “ASLW” with a horizontal orientation). It is expressed by the first execution of phonology-coded gestures using both hands. However, common characters “huang” (empire), “huang” (yellow),

and “huang” (lie) share the same pronunciation “huang” but can be distinguished from the fore and the end radicals, as seen in [18]. These radicals are expressed by the second execution of radical-coded gestures using both hands. Thereby, two executions of coded gestures can determine one Chinese character. By representing Chinese characters one-by-one based on the defined coded gestures, a word or a complete sentence can be formed to express a specific meaning. It can be seen from Figure 1 that all the coded gestures are somewhat static and will not be influenced when the same character is expressed in different words or different sentences. Meanwhile, the number of the coded gestures is limited and constant, which are relatively simply designed so that the standardization of them becomes much easier. Consequently, the most important thing is to recognize the 13 hand shapes with 3 different orientations for this version of CSL including relatively static gestures without the careful consideration of movement epenthesis (ME) problem.

This paper aims to develop an orientation-sensitive and robust feature to represent the above CSL gestures that are executed with different number of stretching-out fingers and different orientations. The main contributions of this paper can be summarized into the following two aspects:

- (1) We propose an orientation-sensitive feature extraction approach based on the distances between the palm center and the key points of the hand contour for CSLR. The key points are extracted by considering the information of stretching-out fingers. The proposed features are capable of representing different orientations of the same gesture shape, as well as characterizing the spatial distribution of stretching-out fingers with the same number but different hand shapes.
- (2) We introduce a DTW-based feature mapping method to further improve the recognition accuracy. The feature mapping method calculates the DTW distances between the current sample and a set of reference samples covering all categories of gestures, aiming to improve the effectiveness of feature representation. To the best of our knowledge, it is the first time that the DTW algorithm is employed for feature mapping in visual recognition problems.

The rest of this paper is structured as follows. Section 2 introduces some related works. In Section 3, we present the proposed CSLR framework, mainly including the extraction of orientation-sensitive features and the feature mapping based on DTW distances. The detailed experimental results and discussions are provided in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Orientation-Sensitive Feature Extraction. Since cameras are simple, noncontact, convenient, and natural devices to capture gesture videos, vision-based approaches have drawn much attention in the field of CSLR [4, 6, 9]. In order to well recognize gestures, feature extraction is an important procedure, which sometimes should consider the rotation or

FIGURE 1: The coded gestures of CSL. (a) The 13 basic gestures indexed from 1 to 13. (b) The 3 orientations abbreviated as U, H, and D, respectively.

TABLE 1: p values derived from the ANOVA analysis.

Classifier	LDA	SVM	KNN
p values	$5.52e-16$	$2.48e-13$	$2.62e-16$

orientation information of gestures. To address this problem, researchers have put forward some solutions. The first type is to develop orientation-sensitive descriptors. For instance, the histogram of gradient (HOG) is a typical rotation-dependent indicator to describe the appearance and the shape of a local object within an image using the distribution of intensity gradients [19]. Wang et al. proposed a novel sparse observation representation approach to convert the high-dimensional HOG features into low-dimensional sparse vectors while obtaining better classification performance [13]. The other type is to evaluate the angles between the actual orientation and the designated orientation of the gesture. For instance, Huang et al. used Gabor-filter-based method to estimate the angles between the gesture orientations and the designated orientation, which later helped to correct the hand pose into the upright orientation [20]. Priyal and Bora utilized the geometry information of gestures to correct gesture orientation by calculating the histogram of palm length to palm width ratios [21].

2.2. Dynamic Time Warping. DTW, proposed by Itakura, is an algorithm to measure the similarity of two distorted trajectories, which overcomes the scale invariance of two models by aligning them in the similarity matrix [22]. If the lengths of two time series that need to compare similarities are not equal, the time axis length of one sequence or both sequences needs to be “warped” for obtaining a better alignment result [23]. The goal of DTW is to find an optimal match path between these two unequal sequences by using a dynamic programming approach, allowing shifting of the time axis. This matching process requires compensation for length differences and the nonlinear properties of length differences in two sequences should be taken into account [24]. Thanks to its small training data requirement and high

accuracy, DTW has been widely used in both sequential and nonsequential models. Readers can refer to the survey on DTW for detailed information [25]. Currently, more and more advanced DTW techniques, such as VQ-DTW [26], weighted-DTW [27], and Soft-DTW [28], have been developed. They have been widely used in biometric identification [29], static and dynamic gesture recognition [30, 31], medical examination [32], etc. Most of the DTW-based studies calculated the similarity of two sequences for classification directly according to the minimum cost or using cluster algorithms. In this paper, the DTW-based distance is employed for feature mapping to improve the effectiveness of feature representation.

3. The Proposed CSLR Method

The flow chart of our Chinese sign language recognition framework based on DTW-distance-mapping features is illustrated in Figure 2. First, gesture videos are acquired and framed to images. Second, each frame of image is transformed to hue, saturation, and value (HSV) color space for hand segmentation [33] and the segmented image is resized. Third, a palm center is localized based on hand boundary contour points and a constant number of key points are determined. The orientation-sensitive features are then extracted based on the distances between the palm center and all the key points. Afterwards, the orientation-sensitive feature is mapped into DTW distances via DTW algorithm between a certain sample and reference samples. Finally, the DTW-distance-mapping features are sent to classifiers for training and testing.

3.1. Hand Segmentation. Different hand segmentation approaches have been available in literature to separate the hand region from the background [33, 34]. Among them, the color-based method is common, which is a distinctive cue of hands and is invariant to scale and rotation [35]. Before performing this method, the framed images from acquired gesture videos in RGB color space are first converted to the HSV color space. Both thresholds (H_{Th} and S_{Th}) of hue and

FIGURE 2: The flow chart of the proposed CSLR framework.

FIGURE 3: Results of hand segmentation. (a) The original image. (b) Skin-color-based hand segmentation. (c) The cropped image after guided filtering. (d) The resized cropped image.

saturation are defined to segment the hand region. In this paper, if the hue pixels from a region are smaller than H_{Th} and the saturation pixels from the same region are larger than S_{Th} , these pixels will be determined within the hand region. H_{Th} and S_{Th} are experimentally set to 0.1 and 0.2.

Then, Otsu's method is utilized to better separate the skin hand region from the nonskin background region, with the criterion that the intraclass variance between the two regions is minimal [36]. Morphological dilate-erode operations are then adopted to eliminate the noise interference. Subsequently, a guided filter, namely, edge-aware image filtering, is performed to get a smooth hand edge [37]. Scale normalization is utilized to resize the cropped hand regions to a constant size of 256×256 . The results of hand segmentation are shown in Figure 3. Figure 3(a) shows an image of gesture "EXPM," and Figure 3(b) shows the result of skin-color-based hand segmentation. Figures 3(c) and 3(d) show the hand-cropped region after guided filtering and the resized hand-cropped region, respectively.

3.2. Feature Extraction

3.2.1. Palm Center Localization. After the hand segmentation, the hand boundary contour is tracked by the 8-

neighborhood contour tracing algorithm, resulting in a sequence of hand contour points for each image [38]. For the purpose of determining the number of stretching-out fingers, the palm circle with a proper radius value is adopted to eliminate the palm region. To achieve this goal, the palm center is first localized according to the hand boundary contour points. The detailed procedures are as follows:

- (1) First, the centroid coordinate of the segmented hand region in the resized image is calculated. The contour point at the bottom of the centroid is assigned as the initial point. Then, the adjacent point in the clockwise direction will be assigned as the second point. In this way, the sequences of the hand contour points are generated.
- (2) Second, the Euclidean distance between each contour point in the sequence and the centroid is calculated, and its minimums, representing the finger roots, are detected using peak detection algorithm [39].
- (3) Third, by applying the least-square method to the detected finger roots, a fitting circle is obtained. The palm center is defined as the center of the fitting circle. Let R_{min} denote the radius of the circle. In order to cover all the stretching-out fingers without

submerging them, an enlarged circle with the radius $R = 1.1R_{\min}$ is adopted. After eliminating the palm by the enlarged circle (setting all the pixels within the circle to zeros), the number of the connected regions corresponding to stretching-out fingers is calculated.

Figure 4 shows the detected finger roots and the corresponding fitting circle. It can be seen from Figures 4(a) and 4(b) that the distance minimums indicate the locations of finger roots, marked as blue stars. The center of the fitting circle (denoted as the red rectangle) and the isolated fingers after the palm elimination are shown in Figure 4(c).

3.2.2. Key Point Determination. After palm elimination, the hand boundary contour points are redetected to form a new sequence and the distance between each new contour point and the palm center is recalculated. The fingertips are determined as the locations corresponding to distance maximums, while finger roots are determined by the threshold (R). Figure 5 shows the detected fingertips and finger roots of gesture “EXPM,” where Figure 5(a) is in the form of distance sequence, while Figure 5(b) is in the form of hand. It can be clearly seen that the maximums (red stars) indicate the fingertips, while the minimums (blue stars) can represent the finger roots. Compared to the minimums in Figure 4(a) where there exist inaccurate or missing finger roots, the detection of finger roots based on our proposed palm elimination technique is more accurate, shown in Figure 4(b) accordingly. Similar conclusions can be found in [30, 40].

In this paper, five key points are selected to represent each stretching-out finger. One point is the fingertip, two other points are the finger roots, and the remaining two points are the midpoints that relate to knuckles. The midpoints are determined as the middle of the horizontal distances between the fingertip and ambilateral adjacent finger roots. The detected midpoints can be seen as the pink stars in Figure 5(b).

Since at most five fingers can stretch out simultaneously, the maximum number of the key points is 25. With the aim of making all the gestures be classified at once while reducing the computation complexity, the number of all the key points is set to be a constant M , where M is not smaller than 25. Suppose that there are N_i , $1 \leq N_i \leq 5$, $N_i \in \mathbb{N}$ stretching-out fingers detected from the resized hand-segmented binary image I_i ; a total of $5N_i$ points can be formed orderly to represent all the stretching-out fingers. The M key points consist of two parts: one part includes the $5N_i$ points representing the stretching-out fingers and the other part includes all the hand boundary contour points inside the interval from the location of the last finger root clockwise to that of the first finger root, which is equally normalized to $M - 5N_i$ boundary contour points.

Figure 5 gives an example of 40 key points of gesture “EXPM” with 5 stretching-out fingers. So far, 25 key points have been determined. In order to form the total 40 hand boundary contour points, the remaining 15 points are uniformly sampled from the interval clockwise from the rightmost finger root to the leftmost finger root, shown as the green stars in Figure 5(b).

There is a particular gesture “HDGP” where none of the finger are stretching out. In this case, Harris corner detection algorithm [41] is adopted to derive the corresponding hand boundary contour points, and all the points are normalized to M points.

3.2.3. Distance Calculation. All the Euclidean distances between the M key points and the palm center are calculated as the orientation-sensitive features. Specifically, the Euclidean distance $\text{Dist}(k)$ between the k -th key point (x_k, y_k) and the palm center (x_0, y_0) is calculated as

$$\text{Dist}(k) = \sqrt{(x_k - x_0)^2 + (y_k - y_0)^2}. \quad (1)$$

3.3. DTW-Distance-Based Feature Mapping. To further improve the effectiveness of feature representation, the original orientation-sensitive feature of each gesture sample is then mapped into the corresponding dynamic time warping (DTW) distance between it and the *reference samples*. The procedures are as follows: N samples of each type of gesture are selected as the reference samples. For 39 types of coded gestures, the r -th feature vector can be denoted as $F_{\text{Ref},r}$, ($1 \leq r \leq 39N$). The original i -th training feature $F_{\text{Tr},i}$ can be mapped to the new feature $F'_{\text{Tr},i}$ according to

$$F'_{\text{Tr},i} = \{D(F_{\text{Tr},i}, F_{\text{Ref},1}), D(F_{\text{Tr},i}, F_{\text{Ref},2}), \dots, D(F_{\text{Tr},i}, F_{\text{Ref},39N})\}, \quad (2)$$

where $1 \leq i \leq N_{\text{Train}}$. N_{Train} is the total number of training gesture samples. Similarly, the original testing feature $F_{\text{Test},j}$ can be mapped to the new feature $F'_{\text{Test},j}$ according to Equation (2), too. The number of the new feature dimensions only depends on the total number of reference samples. Due to the inherent property of DTW, the original orientation-sensitive features are transformed to the new feature space with higher similarity [42]. Besides, since the new features consist of the DTW distances between the sample and all types of gestures, the relationship between each two types of gestures is considered to strengthen the separability of features.

After feature mapping, the new training and testing features will be sent to the classifiers. The effective classifiers can be linear discriminant analysis (LDA) [18], K-Nearest Neighbors (KNN) [43], support vector machine (SVM) [44], and hidden Markov models [10]. Before sending the features into classifiers, a feature dimension reduction technique, uncorrelated linear discriminant analysis (ULDA), is performed to reduce the feature dimension into $C - 1$, where C is the number of gesture types [45]. On the one hand, Fukunaga proposed the optimal dimension of feature space as $C - 1$ for C -class problem [46]. On the other hand, the optimizing criterion in LDA is the ratio of between-class scatter to the within-class scatter, and, for any C -class problem, $C - 1$ nonzero eigenvalues will be obtained [45]. As

FIGURE 4: The detected finger roots and the corresponding fitting circle. (a) Detected minimums. (b) Detected finger roots. (c) Palm elimination by the fitting circle.

FIGURE 5: The locations of fingertips and finger roots in the distance sequence and the hand image. (a) Distance sequence. (b) The hand image.

it is known that ULDA is an extension of LDA by adding some constraints into the optimization objective of LDA, the feature vectors extracted by ULDA will contain minimum redundancy [47]. Thereby, the dimension of the reduced features after ULDA is $C - 1$ for C -class problem.

4. Experiment and Results

4.1. Data Acquisition. A built-in camera from Vivo X7 (Vivo Inc., Guangzhou, China) was employed to collect gesture videos. It was attached on a laptop, with a distance of 0.2 m away from the subject. All the videos were recorded with a frame rate of 30 frames per second (fps) and a resolution of 1280×720 . With the approval of the Ethics Review Committee of Hefei University of Technology, 11 informed subjects (3 females and 8 males, with the average age of 23.64 ± 1.03 years) were recruited in the experiment. Every subject performed all the 39 types of gestures sequentially with each lasting for about 5 seconds, resulting in a total number of 39 videos (all the videos are publicly available and can be obtained from *Baidu Cloud Engine* 1). Since this paper aims to propose a robust and orientation-dependent feature to classify CSL gestures, the environmental settings of data acquisition were relatively simple, regardless of complex background, intensive illumination variations, different colors of skin, and so forth. Some effective solutions

addressing these problems can be found in [21, 44]. Each acquired video was framed to images on MATLAB platform and downsampled to 10 fps. Finally, for each gesture, 50 images were obtained in order to meet the sample requirement of classifiers and provide samples with different illumination variations and tilted angles, resulting in 21450 samples in total ($50 \text{ frames/gesture} \times 39 \text{ gestures/subject} \times 11 \text{ subjects} = 21450 \text{ frames}$).

4.2. Classification Results. Three commonly used classifiers, linear discriminant analysis (LDA), support vector machine (SVM), and K-Nearest Neighbors (KNN) [43, 48], were employed to verify the effectiveness of the DTW-distance-mapping features.

4.2.1. With DTW-Distance-Mapping versus without DTW-Distance-Mapping. Two classification schemes were designed to demonstrate the feasibility of our proposed mapping features. They were classifying all 39 types of gestures with and without DTW-distance-mapping, termed as *Scheme I* and *Scheme II*, respectively. For *Scheme I*, once the reference samples were determined, the samples from the remaining 10 subjects were further divided into training samples and testing samples by applying the leave-one-subject-out strategy again; namely, each one of the 10 subjects took turns

to be the testing subject, while the samples of the other 9 subjects were used for training. As for *Scheme II*, the above two-level leave-one-subject-out strategy for training and testing was exactly the same as *Scheme I* for fair comparison. That is to say, a reference subject was also selected in each round (at the first level) but was just removed from the experiment instead of being used for DTW feature mapping.

The averaged classification rates of our proposed CSLR framework among all the 11 subjects are illustrated in Figure 6. It can be seen clearly from Figure 6 that the results of DTW-mapping-distance features are all higher than 93%. Specifically, the results are $93.20 \pm 0.80\%$, $99.03 \pm 0.19\%$, and $94.92 \pm 0.87\%$ for LDA, SVM, and KNN, respectively, while those of without DTW scheme are $84.33 \pm 0.97\%$, $97.16 \pm 0.19\%$, and $87.78 \pm 0.44\%$, respectively (in the form of Mean $\pm$ Std). The averaged classification rates of *Scheme I* are improved by 8.87%, 1.87%, and 7.14% compared to *Scheme II*, respectively.

A one-way ANOVA test was conducted on the results in Figure 6 to explain the difference between the two schemes from the statistical view [48, 49]. The p values derived from the ANOVA analysis are listed in Table 1. The maximum p value among all the classifiers is $2.48e-13$ ($p < 0.05$), indicating that there is a significant difference between these two schemes.

Furthermore, we take the results based on LDA as an example to show the specific classification results of each subject in Figure 7. The blue ones represent the results of *Scheme I*, while the yellow ones represent those of *Scheme II*. It can be seen that DTW-distance-mapping features can almost improve corresponding classification accuracy for each subject, with the improvement of at least 6.68%. The best classification accuracy improvement can reach 10.35%, achieved from subject S10.

Taking the results of subject S6 as an example, a specific classification confusion matrix of 39 types of gestures is illustrated in Figure 8. Figure 8(a) represents the results of *Scheme I* with KNN, and the corresponding results of *Scheme II* are shown in Figure 8(b). The horizontal and vertical coordinates represent predicted class and actual class, and the values of the axes represent the gesture types. The first 13 values represent all types of the downward gestures and then the horizontal and upward ones. It can be seen from Figure 8(b) that most types of gestures are well classified, and some of them can be 100% classified, such as “DASLW” and “UPINT.” It can be clearly seen that the misclassified gesture in *Scheme I* is significantly reduced, especially for line 6 (gesture “DEXIF”), line 23 (gesture “HEXTF”), and line 33 (gesture “UEXTF”). These results verify that the separability ability can be improved by DTW-distance-mapping features compared to the original features.

4.2.2. DTW Matching versus DTW-Distance-Mapping. In order to show the better performance of DTW-distance-mapping compared to that of DTW matching, where the latter one means adopting DTW as a template matching for classification, we first give the overall classification accuracy

FIGURE 6: The averaged classification rates of 39 gestures among 11 subjects for two schemes. The blue ones and the yellow ones represent the results of *Scheme I* and *Scheme II*, namely, with and without DTW-distance-mapping.

of 11 subjects using DTW matching and DTW mapping followed by classifiers, listed in Table 2. The accuracy of using DTW matching for all 39 types of gestures is only 29.30%, while that of using DTW mapping followed by classifier is at least 93.20%, with an improvement of at least of 63.90%. Meanwhile, all 39 types of gestures can be divided into subsets according to the number of stretching-out fingers. The numbers of gesture types of each subset are 3, 9, 12, 9, 3, and 3. The overall classification accuracy of each subset is also shown in Table 2, with * – F corresponding to * stretching-out fingers. All the results demonstrate that the strategy of DTW mapping followed by classifiers is superior to that of traditional DTW matching alone. It can be seen that the number of gesture types of 1 – F subset is 9, 3 fewer than that of 2 – F subset. However, the overall classification accuracy of 1 – F subset is lower than that of 2 – F subset during every case. It can be indicated that when the number of the stretching-out finger is fewer, the DTW distance between each two feature sequences will be much similar. By means of DTW mapping, all the features are transformed into minimum cumulative distance of warp path while considering the differences between each two samples [30]. Thereby, the DTW mapping can enhance the separation ability of features to improve the classification accuracy. However, as for 4 – F and 5 – F subsets, although there are only 3 types of gestures in each subset, the traditional DTW matching method is unable to recognize them well, with the classification accuracy of about 74%. It indicates that the traditional DTW matching method cannot well distinguish the same gesture but with different orientations.

To further demonstrate the feasibility of DTW mapping strategy, the classification confusion matrices for all the 9 types of 1 – F gestures (subject 1 was designated as the reference subject) derived from DTW mapping followed by KNN and DTW matching are shown in Figure 9, with Figure 9(a) representing the confusion matrix using DTW matching and Figure 9(b) representing results using the

FIGURE 7: The overall classification accuracy of 39 gestures for all subjects. The blue ones represent the results of *Scheme I*, and the yellow ones represent those of *Scheme II*, namely, with and without DTW-distance-mapping.

FIGURE 8: The DTW-distance-mapping features of all types of gestures with 1 stretching-out finger. (a) *Scheme I* with DTW-distance-mapping. (b) *Scheme II* without DTW-distance-mapping.

TABLE 2: The overall classification accuracy of 11 subjects using DTW matching and DTW mapping followed by classifiers.

	0- <i>F</i> (%)	1- <i>F</i> (%)	2- <i>F</i> (%)	3- <i>F</i> (%)	4- <i>F</i> (%)	5- <i>F</i> (%)	All (%)
DTW matching	59.43	32.85	33.66	38.03	74.69	74.18	29.30
DTW mapping + LDA	84.96	89.66	92.15	94.40	97.69	94.48	93.20
DTW mapping + SVM	85.82	95.61	97.21	98.71	98.41	95.43	99.03
DTW mapping + KNN	96.31	91.61	93.26	95.76	97.61	94.61	94.92

* – *F* stands for the number (*) of the stretching-out fingers. Here, * ranges from 0 to 5. “All” means all 39 types of gestures.

DTW mapping. The parameters of the KNN were set to 100 neighbors, Euclidean distance, and equal distance weight, respectively. According to Figure 9(a), except for the accuracies of 66% for gesture “DEXIF” and 59% for gesture “HEXIF,” the recognition rates of the other 7 gestures are all less than 50% when using DTW as template matching. However, when adopting the strategy of DTW mapping, the average recognition accuracy is significantly improved, with the lowest

recognition rate of 81% for the gesture “DEXTF.” The overall recognition accuracy is 89.50%, with an improvement rate of 62.06% compared to that using DTW matching.

The reasons why the performance of DTW mapping followed by classifiers is superior to that of DTW matching can also be revealed from the feature representation perspective. Taking the 9 types of 1 – *F* gestures for instance, Figure 10 shows the original orientation-sensitive features,

FIGURE 9: The classification confusion matrix for all the 9 1-stretching-out-finger gestures. (a) Confusion matrix of using DTW as template matching. (b) Confusion matrix of using DTW-distance-mapping followed by KNN.

FIGURE 10: The orientation-sensitive features of all 9 types of gestures with 1 stretching-out finger: (a) DEXIF; (b) DEXTF; (c) DEXLF; (d) HEXIF; (e) HEXTF; (f) HEXLF; (g) UEXIF; (h) UEXTF; (i) UEXLF.

FIGURE 11: The DTW-distance-mapping features of all types of gestures with 1 stretching-out finger: (a) DEXIF; (b) DEXTF; (c) DEXLF; (d) HEXIF; (e) HEXTF; (f) HEXLF; (g) UEXIF; (h) UEXTF; (i) UEXLF.

while Figure 11 shows the corresponding DTW-distance-mapping features. It can be seen from Figure 10 that, even with the same number of stretching-out fingers, the corresponding features are different with the same orientation (comparison of each 3 horizontal subfigures). For the same gesture with different orientations, the corresponding features are also different (comparison of each 3 vertical subfigures). However, when only one finger is stretching out, the original feature waveforms are similar but with a shifting of the time axis, which cannot be well distinguished using the DTW matching. The reason is that the DTW distance between these two feature sequences with similar waveform but time shifting is small.

It can be seen from Figure 11 that the DTW-distance-mapping features of all types of gestures in 1 – F subset are separable. Specifically, the DTW-distance-mapping feature of “UEXLF” is quite different from that of “HEXIF,” which indicates that, by considering the relationship between each two types of gesture via DTW distance, the DTW-distance-mapping features are more separable.

4.3. Discussion. The above results clearly demonstrate that (1) the proposed feature extraction approach based on DTW-distance-mapping is effective for CSLR and (2) the DTW-based feature mapping strategy can further improve the recognition accuracy. By using DTW mapping, the features are transformed into minimal cumulative distance of warp path to enhance the feature representation ability, which can overcome the differences between subjects, such as gesture execution habits or different sizes of the palm. To further analyze the results, we compared them with other related studies. In [50], Liu et al. designed a hierarchical strategy based on the number of fingers to solve the “large-category” problem of SLs, and HOG features were used. The results of 25 popular static gestures were 78.9% for training gestures in a fixed orientation and 99.9% for training gestures in multiple orientations. Nevertheless, the results of 39 types of gestures in our study can also reach 99.03% using DTW-distance-mapping combined with SVM. Besides, the maximum dimension of the orientation-sensitive

features is only 40, which is far smaller than that of the HOG features. Subsequently, in 2016, Chen et al. proposed locating palm center, fingertips, and finger roots, as well as root-center-angle features using finger length weighted Mahalanobis distance. The average accuracy of the 15 self-collected gestures with 9 rotations was 94.67%, which benefitted from the usage of the supernumerary depth data from Kinect sensor and the robustness to illumination variations [51]. Ren et al. proposed a novel Finger-Earth Mover's Distance (FEMD) method to measure the dissimilarity between hand shapes, which helped to achieve an average accuracy of 93.2% for 10 gestures. The FEMD was robust to hand articulations, distortions, and orientation or scale changes and could even handle challenging situations including cluttered backgrounds and lighting conditions [40, 52]. However, in our study, the same gesture executed with different orientations is treated as a different gesture, where the FEMD method is suboptimal. Meanwhile, the classification performance of using DTW-distance-mapping features has been proved to be better than that using DTW as template matching. In addition, the time-consumption complexity of our proposed DTW is $O(N^2)$, while that of EMD algorithm is cubic, specifically $O(N^3 \log n)$ [53]. The mean running times of a hand recognition procedure were reported as 4.0012 seconds and 0.0750 seconds for near-convex decomposition followed by FEMD and the thresholding decomposition followed by FEMD, respectively [40]. As for our DTW-distance-mapping method, the mean running time of a hand recognition procedure was 0.07188 seconds (MATLAB 2015b, i5-7500 CPU, 8G RAM).

It is noted that a drawback of our proposed CSLR system is the skin-color-based hand segmentation approach, which has restrictions such as illumination invariance and relative simple background, although this step is not the focus of this work. In the future, we will develop a more robust hand segmentation method to improve the practicability of the proposed CSLR system. The other drawback of our proposed CSLR framework lies in determining the radius of the palm fitting circle, which is important to determine the number of stretching-out fingers. Due to physiological anatomy and gesture execution habit differences between subjects, the radius of the palm fitting circle currently cannot be constant among all the subjects. In the future, we will adopt some adaptive methods to solve this problem. The other problem is that there is a restriction of finger occlusion. If the stretching-out fingers are occluded but the maximum distance between the distance contour point and the palm center is larger than the radius R , the corresponding fingertips can still be detected. However, if the maximum distance is not larger than the radius R due to severe occlusion, the corresponding fingertips cannot be detected by our proposed method. In this case, gestures based on RGB-D data and FEMD method might provide valuable information that can handle severe occlusion problem [40]. Besides, with the development of deep learning technique, deep-learning-based CSLR system will also be further explored.

5. Conclusions

In this paper, we have proposed a novel two-stage feature representation method for Chinese sign language recognition (CSLR). The target is to classify 39 types of static CSL gestures, consisting of 13 types of basic hand shapes with each having 3 orientations. An orientation-sensitive feature extraction approach is presented based on distances between the palm center and the key points of the hand contour. Besides, a dynamic time warping- (DTW-) based feature mapping approach is proposed to further improve the recognition accuracy. Experimental results demonstrate the effectiveness of the proposed feature representation approach. This study will facilitate the practical progress of CSLR systems.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank all the volunteers for their contributions to the experiments. This work was supported by the National Key R&D Program of China (Grant no. 2017YFB1002802), the National Natural Science Foundation of China (Grant nos. 61922075, 61701160, and 41901350), the Provincial Natural Science Foundation of Anhui (Grant no. 1808085QF186), and the Fundamental Research Funds for the Central Universities (Grant nos. JZ2020HGPA0111 and JZ2019HGBZ0151).

References

- [1] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131–153, 2019.
- [2] G. Fang, W. Gao, and D. Zhao, "Large vocabulary sign language recognition based on fuzzy decision trees," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 34, no. 3, pp. 305–314, 2003.
- [3] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for Chinese sign language recognition using accelerometer and semg data," *IEEE Transactions on Bio-Medical Engineering*, vol. 59, no. 10, pp. 2695–2704, 2012.
- [4] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to Arabic sign language recognition," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 4, pp. 551–557, 2014.
- [5] E.-J. Holden, G. Lee, and R. Owens, "Australian sign language recognition," *Machine Vision and Applications*, vol. 16, no. 5, pp. 312–320, 2005.
- [6] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer

- based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.
- [7] J. Feng, G. Wen, H. Yao, D. Zhao, and X. Chen, "Synthetic data generation technique in signer-independent sign language recognition," *Pattern Recognition Letters*, vol. 30, no. 5, pp. 513–524, 2009.
 - [8] http://www.who.int/pbd/deafness/WHO_GE_HL.pdf.
 - [9] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel Chinese sign language recognition method based on keyframe-centered clips," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 442–446, 2018.
 - [10] W. Yang, J. Tao, and Z. Ye, "Continuous sign language recognition using level building based on fast hidden markov model," *Pattern Recognition Letters*, vol. 78, pp. 28–35, 2016.
 - [11] D. Ma, X. Chen, Y. Li, J. Cheng, and Y. Ma, "Surface electromyography and acceleration based sign language recognition using hidden conditional random fields," in *Proceedings of the 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pp. 535–540, Langkawi, Malaysia, December 2012.
 - [12] J. Zhang, W. Zhou, and H. Li, "A new system for Chinese sign language recognition," in *Proceedings of the 2015 IEEE China summit and international conference on signal and information processing (ChinaSIP)*, pp. 534–538, Chengdu, China, July 2015.
 - [13] H. Wang, X. Chai, and X. Chen, "Sparse observation (SO) alignment for sign language recognition," *Neurocomputing*, vol. 175, pp. 674–685, 2016.
 - [14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4207–4215, Las Vegas, NV, USA, June 2016.
 - [15] S. Yang and Q. Zhu, "Video-based chinese sign language recognition using convolutional neural network," in *Proceedings of the 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, pp. 929–934, Guangzhou, China, May 2017.
 - [16] Q. Xiao, M. Qin, P. Guo, and Y. Zhao, "Multimodal fusion based on lstm and a couple conditional hidden markov model for Chinese sign language recognition," *IEEE Access*, vol. 7, pp. 112 258–112 268, 2019.
 - [17] China Association of the Deaf, *Chinese Sign Language (Revised Edition)*, Hua Xia Press, Beijing, China, 2003, in Chinese.
 - [18] J. Cheng, X. Chen, A. Liu, and H. Peng, "A novel phonology- and radical-coded Chinese sign language recognition framework using accelerometer and surface electromyography sensors," *Sensors*, vol. 15, no. 9, pp. 23 303–324, 2015.
 - [19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893, San Diego, CA, USA, June 2005.
 - [20] D.-Y. Huang, W.-C. Hu, and S.-H. Chang, "Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6031–6042, 2011.
 - [21] S. P. Priyal and P. K. Bora, "A robust static hand gesture recognition system using geometry based normalizations and krawtchouk moments," *Pattern Recognition*, vol. 46, no. 8, pp. 2202–2219, 2013.
 - [22] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.
 - [23] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 2001 SIAM International Conference on Data Mining*, pp. 1–11, Chicago, IL, USA, April 2001.
 - [24] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," pp. 359–370, In KDD Workshop, Seattle, WA, USA, 1994.
 - [25] C.-Y. Kao and C.-S. Fahn, "A human-machine interaction technique: hand gesture recognition based on hidden markov models with trajectory of hand motion," *Procedia Engineering*, vol. 15, pp. 3739–3743, 2011.
 - [26] M. Faundez Zanuy, "On-line signature recognition based on VQ-DTW," *Pattern Recognition*, vol. 40, no. 3, pp. 981–992, 2007.
 - [27] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," *Pattern Recognition*, vol. 44, no. 9, pp. 2231–2240, 2011.
 - [28] M. Cuturi and M. Blondel, "Soft-DTW: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 894–903, Sydney, Australia, August, 2017.
 - [29] L. Yang, J. Shen, S. Bao, and S. Wei, "Biometric identification method for ECG based on the piecewise linear representation (PLR) and dynamic time warping (DTW)," *Journal of Biomedical Engineering*, vol. 30, no. 5, pp. 976–981, 2013.
 - [30] C. Hong, Z. Dai, Z. Liu, and Z. Yang, "An image-to-class dynamic time warping approach for both 3d static and trajectory hand gesture recognition," *Pattern Recognition*, vol. 55, pp. 137–147, 2016.
 - [31] H.-R. Choi and T. Kim, "Modified dynamic time warping based on direction similarity for fast gesture recognition," *Mathematical Problems in Engineering*, vol. 2018, Article ID 2404089, 9 pages, 2018.
 - [32] S. Adwan, I. Alsaleh, and R. Majed, "A new approach for image stitching technique using dynamic time warping (DTW) algorithm towards scoliosis x-ray diagnosis," *Measurement*, vol. 84, pp. 32–46, 2016.
 - [33] Y. Tayal, R. Lamba, and S. Padhee, "Automatic face detection using color based segmentation," *International Journal of Scientific and Research Publications*, vol. 2, no. 6, pp. 1–7, 2012.
 - [34] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.
 - [35] H. Lahamy and D. D. Lichti, "Towards real-time and rotation-invariant american sign language alphabet recognition using a range camera," *Sensors*, vol. 12, no. 11, pp. 14 416–514 441, 2012.
 - [36] N. Ohtsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems Man & Cybernetics*, vol. 9, no. 1, pp. 62–66, 2007.
 - [37] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proceedings of the in European Conference on Computer Vision*, pp. 1–14, Heraklion, Crete, Greece, September 2010.
 - [38] F. Chang, C.-J. Chen, and C.-J. Lu, "A linear-time component-labeling algorithm using contour tracing technique," *Computer Vision and Image Understanding*, vol. 93, no. 2, pp. 206–220, 2004.
 - [39] H. S. Shin, C. Lee, and M. Lee, "Adaptive threshold method for the peak detection of photoplethysmographic waveform," *Computers in Biology and Medicine*, vol. 39, no. 12, pp. 1145–1152, 2009.
 - [40] R. Zhou, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

- [41] C. Harris, "Geometry from visual motion," in *Active Vision*, pp. 263–284, MIT Press, Cambridge, MA, USA, 1993.
- [42] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.
- [43] M. Murugappan, N. Ramachandran, and Y. Sazali, "Classification of human emotion from EEG using discrete wavelet transform," *Journal of Biomedical Science and Engineering*, vol. 3, no. 4, pp. 390–396, 2010.
- [44] N. H. Dardas and N. D. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 11, pp. 3592–3607, 2011.
- [45] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis: a brief tutorial," in *Proceedings of the International Joint Conference on Neural Network*, pp. 387–391, IEEE, Anchorage, AK, USA, 1998.
- [46] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Elsevier, Amsterdam, Netherlands, 2013.
- [47] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognition*, vol. 34, no. 7, pp. 1405–1416, 2001.
- [48] J. Cheng, F. Wei, C. Li, Y. Liu, A. Liu, and X. Chen, "Position-independent gesture recognition using semg signals via canonical correlation analysis," *Computers in Biology and Medicine*, vol. 103, pp. 44–54, 2018.
- [49] R. N. Khushaba, "Correlation analysis of electromyogram signals for multiuser myoelectric interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 745–755, 2014.
- [50] S. Liu, Y. Liu, J. Yu, and Z. Wang, "Hierarchical static hand gesture recognition by combining finger detection and hog features," *Journal of Image and Graphics*, vol. 20, no. 6, pp. 781–788, 2015.
- [51] X. Chen, C. Shi, and B. Liu, "Static hand gesture recognition based on finger root-center-angle and length weighted mahalanobis distance," in *Proceedings of the SPIE 9897, Real-Time Image and Video Processing 2016*, SPIE, Brussels, Belgium, April 2016.
- [52] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM International Conference on Multimedia—MM '11*, pp. 1093–1096, ACM, New York, NY, USA, November 2011.
- [53] X. Xu, F. Lin, A. Wang, Y. Hu, M. C. Huang, and W. Xu, "Body-earth mover's distance: a matching-based approach for sleep posture recognition," *IEEE Transactions on Biomedical Circuits & Systems*, vol. 10, no. 5, pp. 1023–1035, 2016.