Research on Dynamic Sign Language Recognition Based on Key Frame Weighted of DTW

ShengWei Zhang, ZhaoSong Zhu(✉), and RongXin Zhu

Nanjing Normal University of Special Education, Nanjing 210038, China
zzs@njts.edu.cn

Abstract. Dynamic sign language can be described by its trajectory and key hand types. Most of the commonly used sign language can be recognized by trajectory curve matching. Therefore, In this paper, a new dynamic sign language recognition method is proposed, which uses trajectory and key hand type to extract features, adopts a key frame weighted DTW (dynamic time warping) algorithm to implement hierarchical matching strategy, and gradually matches sign language gestures from two levels of trajectory and key hand type, so as to effectively improve the accuracy and efficiency of sign language recognition.

Keywords: Sign language recognition · Key frames · Dynamic time warping

1 Introduction

As a special gesture, sign language is a way for deaf mutes to communicate by gesture instead of sound language. According to the latest statistics released by the Ministry of health, there are 20.57 million deaf people in China, accounting for 1.67% of the total population in China [1]. Due to the limitation of physiological factors, it is difficult for the deaf to speak. Therefore, sign language is the mother tongue of most deaf people and is the main tool for deaf people to express their thoughts and feelings, obtain information and participate in social life. However, there are few healthy people who are proficient in sign language. Apart from those engaged in special education, most of them do not understand sign language and have no intention to learn sign language. This has caused great obstacles to the communication between deaf and healthy people.

At present, there are few professional sign language training institutions and sign language translators, which are far from meeting the market demand. In order to further promote the barrier free construction of information exchange and shorten the communication distance between deaf mute and healthy people, it is particularly important to study sign language recognition technology. Sign language recognition is to obtain the sign language data of the deaf through the computer acquisition equipment, use machine learning algorithm, combined with context knowledge, to obtain the meaning of sign language, and then translate it into speech, and convey it to normal people who do not

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 388, pp. 11–20, 2021.
https://doi.org/10.1007/978-3-030-82565-2_2

understand sign language. Sign language recognition technology can not only let us enter the deaf's silent world, but also make the deaf people understand the modern society more comprehensively. So as to further enhance the communication between the deaf and healthy people, and realize barrier free communication.

2 Literature Review

The research on sign language recognition can be traced back to the 1980s. According to the different acquisition methods of gesture data, sign language recognition technology can be divided into two types based on data glove and computer vision.

2.1 Sign Language Recognition Based on Data Glove

Early computer computing power is weak, but computer vision technology needs a lot of complex computing. In contrast, data glove can use sensors to obtain real-time and accurate gesture data, so in the early sign language recognition, the use of data glove has become the mainstream. Some data gloves can not only provide information on the position of the hand, but also record the force applied on each finger [2]. Han proposed a cheap data glove with high recognition accuracy [3]; Hernandez-Rebollar et al. Developed a data glove based method to recognize 26 American Sign Language (ASL), In 2004, a system capable of recognizing 176 asls was developed, with a recognition accuracy of 95% [4]; Kevin and Kim also used data gloves for gesture recognition [5]; In China, Gao Wen et al. Used a number of data gloves, combined with artificial neural network (ANN) and HMM model to train gesture, realized the recognition of isolated words, the recognition rate was more than 90%, and then realized the continuous Chinese sign language recognition system of more than 5000 words [6].

Although the use of data gloves can quickly and accurately get hand features, even subtle movements can be recognized, so it can get a higher recognition accuracy. However, the use of data gloves is complex and expensive, and does not conform to the natural human-computer interaction habits, which is not conducive to further promotion and use, so it can only be used in the laboratory or in specific occasions.

2.2 Sign Language Recognition Based on Computer Vision

Compared with data glove, computer vision technology has many advantages, such as interactive mode more in line with natural habits, cheap equipment and easy to promote. It can be divided into three types: monocular camera, multi camera and somatosensory camera.

Monocular Camera. Monocular means that the input device has only one two-dimensional camera. Camera can collect sign language data more naturally, but tracking and segmenting hand region from complex background is a challenge. The common processing method is to color mark the hand of sign language speaker. Deng et al. Simplified gesture segmentation with color gloves and used parallel HMM for recognition, and the recognition rate of 192 American sign language words reached 93.3% [7]; Manar

et al. Used recurrent neural networks (RNN) to recognize Arabic static sign language, and sign language users wore gloves with highlighted marks, and the recognition rate reached 95.11% [8]. Pattern recognition is also used to recognize gesture regions, but the real-time performance is often poor. Ong et al. Used the boosted cascade classifier in the gray image to detect and track the hand [9]. Zhang Guoliang and others use color gloves to simplify gesture segmentation and use continuous hidden Markov model (CHMM) for recognition, with an average recognition rate of 92.5% [10].

Because the monocular camera can only obtain two-dimensional image information, it can not achieve accurate positioning, so it is difficult to adapt to the complex changes of hands in three-dimensional space, and hand marking is still not in line with natural human-computer interaction. Therefore, researchers try to use multi camera to capture images of different dimensions to make up for this defect.

Multi Camera. Multi camera refers to the use of two or more two-dimensional cameras to obtain two-dimensional image information from different angles, so as to obtain the accurate data of hand in three-dimensional space. Volger et al. Used three orthogonal cameras to locate the arm and estimate the shape and 3D motion parameters of the hand [11, 12]; Utsumi et al. Studied a gesture recognition system using four cameras [13]; Argyro et al. proposed a gesture tracking method based on two video streams with different angles to generate 3D data [14]. They mark the position of the hand in each video stream, and then match the information of the two parts with the angle calibration calculation, so as to obtain the three-dimensional data of the hand position. These methods require high performance of the computer, so it is difficult to process them in real time if their computing power is not strong. Moreover, there are many noises in the depth information estimated by this method. In addition, the use of multi camera, each need to calibrate, the use of inconvenience.

Somatosensory Camera. In recent years, due to the emergence of somatosensory cameras, gesture recognition based on 3D data has made great progress. In the aspect of gesture recognition using a somatosensory camera, Jang et al. Proposed a system based on Kinect to obtain depth information to recognize gesture, and used continuous adaptive mean shift algorithm, CAMSHIFT) to use depth probability and update depth histogram for hand tracking [15]; Chai et al. used Kinect to obtain 3D features of gestures and realized recognition through 3D trajectory matching of hands, with an average recognition rate of 83.51% [16]; Marin used Kinect to locate the hand area, then used leap motion to get the fine information of the hand, and then used support vector machine (SVM), SVM is used as a classifier to recognize gesture, and the recognition rate reaches 91.28% [17]. At present, using Kinect to obtain in-depth information to identify sign language has become the mainstream [18], but there are relatively few domestic related research results, and there are still some problems to be solved in current sign language recognition. For example, how to ensure the stability of hand region segmentation and sign language feature description due to the difference of body and action habits of different sign language speakers? In addition, dynamic sign language is a kind of sign language which is represented by the combination of several changing gestures, which has a large amount of data. For sign language recognition, how to ensure the recognition accuracy and meet the real-time requirements has always been a hot topic for researchers.

3 Key Frame Weighted of DTW Method

The Kinect camera is used as the data acquisition device, and the 3D bone data stream provided by Kinect camera is used for gesture recognition. Kinect provides 30 frames per second depth image and three-dimensional coordinates of 20 human joint points, including the joint point information of left and right hands. The skeleton flow data provided by Kinect was preprocessed. The process of gesture recognition includes four main parts: data acquisition, data preprocessing, data feature extraction and recognition output. After data preprocessing, the original feature data of gesture tracking trajectory is obtained, and the feature extraction of trajectory is carried out, and the hierarchical matching strategy is adopted for sign language recognition.

Kalman filter is used to correct the data returned by Kinect to eliminate singular points and ensure the accuracy and consistency of the trajectory. The hand position data points detected in each frame can be connected to get the gesture trajectory curve. Key frame is a key action in sign language. Through observation and personal experience, it is found that most sign language users will stop for a while to show their emphasis on the key action. In the trajectory curve, that is, near the time when the key frame appears, the data points are particularly dense. Based on this, a key frame extraction algorithm based on track point density is proposed. Firstly, the density set $crow(P)$ of a point P on the trajectory curve ρ was defined as Eq. (1).

$$crow(P) = \{X_i | \forall X_i \in \rho, \delta(P, X_i) \leq \Delta\} \tag{1}$$

$\delta(P, X_i)$ is the Euclidean distance between P and X_i, and Δ is the threshold. However, for the point P on the trajectory curve of sign language, we hope that it is not only close to X_i in space distance, but also adjacent in time. The definition of Eq. (1) ignores the change of time. For example, the positions of the start and end gestures are very close, but the time is quite different. However, according to the definition of Eq. (1), it is obvious that the point of ending gesture will be classified into the density set of the initial gesture, which will cause errors. Therefore, it is necessary to add a time limit when counting the density set of a point on the trajectory curve, that is, data points X_i and X_{i+1} must be adjacent in time. The number of data points in the density set is the point density of P. For a specific sign language, if it is completed by both hands, we need to calculate the point density of the left hand and the right hand trajectory curve respectively, and add them as the overall curve point density of the sign language. For example, for the gesture of sign language word “husband”, the point density curve can be obtained from the trajectory curve (see Fig. 1).

The abscissa represents the time, and the ordinate represents the point density value of the corresponding time. The mean filter with width and length of 5 is used to smooth it (see Fig. 2).

There are three obvious peaks which marked by red circles. Check the key frame images corresponding to these three peaks, and the results are shown in Fig. 3. Except that the last frame is the termination gesture, the remaining two frames are consistent with our expected keyframes. At the same time, the “husband” sign language made by another sign language is processed the same way, and the same keyframes is obtained. Therefore, it is feasible to detect key frames by using point density. However, in fact,

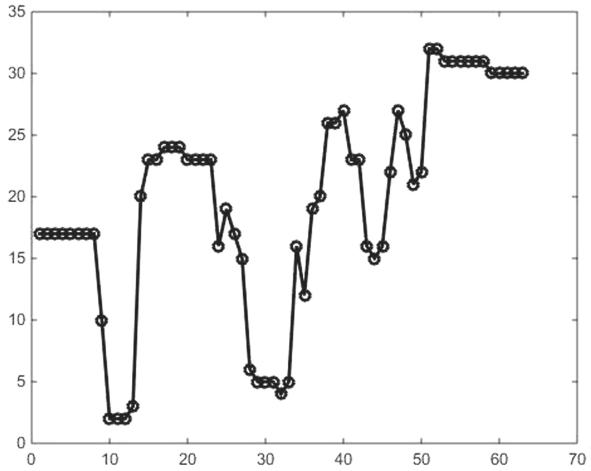

Fig. 1. The point density curve of sign language word “husband”

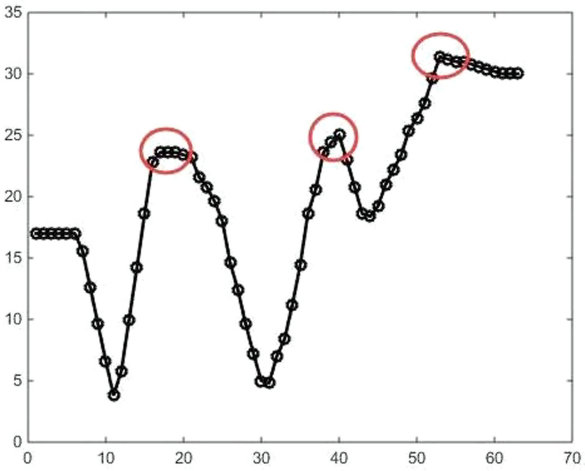

Fig. 2. The point density curve of sign language word ‘husband’ through mean filtering

most sign languages do not have such distinctive features, and can not easily get key frames from point density images.

Then, the point density curve can be segmented into several continuous and equal width windows. In each window, the maximum point density value is found. If the value is greater than the given threshold, the frame to which it belongs is taken as the candidate key frame. It can be seen that the selection of window size and threshold are two key factors affecting the candidate keyframes. If the window width is too small, there will be too many candidate frames, which will lead to too long time to filter the final key frame; if the window width is too large, the candidate frames will be too few, which will lead to missing key frames. Generally speaking, for most sign language, the number of key

Fig. 3. Keyframes of the sign language word "husband"

frames of a gesture will not exceed 6. When the number of candidate frames is two to three times of the number of key frames, it is more appropriate. In the candidate frames, the final key frame can be determined quickly by using the frame subtraction method. The threshold and the threshold Δ in Eq. (1) are defined in an adaptive way.

After the position of the key frame is determined, the trajectory can be matched one level. Trajectory is a typical time series. DTW algorithm is recommended to measure the similarity of time series. DTW distance can find the best alignment and matching relationship between time series by stretching and bending them, so as to measure the similarity of different length time series. Given two time series: $T = \{t_1, t_2, \ldots, t_m\}$ and $R = \{r_1, r_2, \ldots, r_n\}$. The DTW method finds an optimal bending path in the time series T and R, and takes the cumulative distance of this path as the distance between T and R. Even if the length of time series T and R is not equal, the similarity between them can be calculated in this way. The traditional DTW algorithm does not consider the characteristics of sign language itself. In other words, there are two kinds of data points in sign language time series, one is the transition point, the other is the key frame point where the key frame is located. Obviously, their semantic contributions to sign language are different, and they have the same status in the traditional DTW distance, which is obviously unreasonable. Therefore, a key frame weighted DTW algorithm is proposed to improve the accuracy of traditional DTW algorithm for sign language trajectory matching.

Suppose there are two sign language trajectory curves P and Q, P contains m frame of data points, Q contains n frame of data points, then $P = \{p_1, p_2, \ldots, p_m\}$, $Q = \{q_1, q_2, \ldots, q_n\}$.

The distance from point X of curve a to keyframe point was defined as Eq. (2).

$$\delta_{KP}(i) = min\{|i - x|, x \in K_P\} \tag{2}$$

The distance δ_{KQ} from point q_j of curve Q to keyframe point was defined as Eq. (3).

$$\delta_{KQ}(j) = min\{|j - x|, x \in K_Q\} \tag{3}$$

Then the cumulative cost matrix $D(i, j)$ was defined as Eq. (4).

$$D(i,j) = \left(\left|\delta_{KP}(i) - \delta_{KQ}(j)\right| + 1 \times P(i,j)\right) + min\begin{cases} D(i-1.j) \\ D(i-1, j-1) \\ D(i, j-1) \end{cases} \tag{4}$$

Here $P(i, j)$ is the distance from p_i to q_j For point p_i on curve P and point q_j on curve Q, if $\delta_{KP}(i) = \delta_{KQ}(j)$, then there are two cases. One is that point p_i and point q_j approach

key points from the same direction, which indicates that they have the same distance from their respective keyframes, and they should have higher weights, so their corresponding coefficients are smaller; the other case is that point p_i and point q_j approach the key points from left and right respectively, although their values are lower The coefficients are still small, but the coefficients corresponding to other matching points on this matching path will increase, which also ensures the constraint of key frames.

Before measuring the similarity of tracks, we should normalize the trajectory curves, and deal with them both in space and time, so as to eliminate the different effects caused by the differences of sign language users. At the same time, it should be noted that some sign language can be completed by both hands at the same time, and some sign language can be completed with one hand. Therefore, the trajectory curve of gesture is to distinguish left and right hands. For those two hand sign language, the curves of left and right hand should be calculated respectively, and the DTW distance between the left and right curves should be taken as the final matching result. Moreover, there is no difference between the left and right hands in the semantic contribution of sign language, which is equally important.

If the ratio of the DTW distance of the most similar sign language to the sign language to be recognized is less than a certain threshold value, then the recognition can be finished. Otherwise, the first five sign language categories with the smallest distance are returned. In these five categories, the key hand type is used to do the second level matching. The key hand type is also a time series, but the number of elements in the sequence is small, which is less than five. It is easy to use DTW to calculate the distance for such a sequence.

4 Experimental Results and Discussions

This paper compares the time performance of traditional DTW distance, FastDTW distance and our improved DTW distance. As shown in Fig. 4. The abscissa is the length of the two time series, and the ordinate is the time required to match (in MS). The experimental data are randomly generated three-dimensional data, and the length of the two time series is equal. Keyframe points are randomly generated one or two points per 100 points. The experiment was repeated 1000 times and the average value was taken.

As can be seen from Fig. 4, with the increase of the length of time series, the matching time increases in a quadratic way, and the time performance is very poor. Although the matching time of the proposed algorithm is quadratic, the matching time is less than that of the FastDTW distance when the length of the time series is less than 576 because of the global constraint. Although the time of FastDTW distance matching is increasing, it needs to be refined from coarse-grained path, so its advantage is not obvious when the length of time series is short. However, due to the need of backtracking, its performance is not as good as the algorithm in this paper.

The video duration of a sign language word is generally about 2–6 s. The sampling frequency of Kinect camera is 30 frames/s, so a video has 60–180 frames of images, so the length of time series is between 60–180. Figure 4 shows the time performance of traditional DTW distance, FastDTW distance and the improved DTW distance when the time series length is less than 200. As can be seen from the figure, the time required by the algorithm in this paper is the lowest. Most of the matching time is within 1 ms.

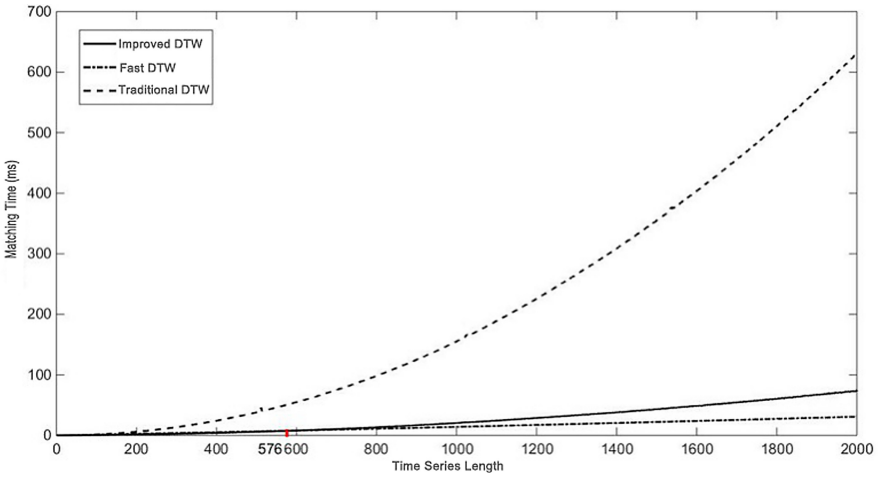

Fig. 4. Time comparison of different DTW algorithms

For 60 sign language templates, all matching needs less than 60 ms. Therefore, when the number of sign language templates is less than a few hundred, real-time recognition can be realized.

Table 1 shows the recognition accuracy of different sign language users when only one level matching and two level matching are added. It can be seen that on the basis of trajectory recognition, adding hand information for matching can effectively improve the accuracy of sign language recognition, which basically reaches more than 90% recognition rate, which verifies the effectiveness of the algorithm in this paper.

Table 1. Sign language recognition accuracy

Sign language tester	1	2	3	4	5
First level matching	0.8343	0.8667	0.8835	0.8967	0.9010
Two level matching	0.9010	0.9265	0.9310	0.9335	0.9668

5 Conclusion

In this paper, a new method of dynamic sign language recognition is proposed, which uses trajectory and key hand to extract features, adopts a key frame weighted DTW (dynamic time warping) algorithm to implement hierarchical matching strategy, due to fully considering the characteristics of dynamic sign language itself, In order to effectively improve the accuracy and efficiency of sign language recognition, sign language gestures are matched step by step from the two levels of trajectory and key hand type.

However, when deaf people play sign language, it is often accompanied by facial expression and lip movement. Recognition of facial expression and lip movement can

better understand the emotion expressed by sign language. In addition, there are many other deep learning network structures at present [19]. For example, the deep network based on attention model is one of the most important core technologies in deep learning technology, and the future research goal is sign language recognition based on attention mechanism.

Acknowledgement. The work described in this paper was fully supported by a grant from the National Philosophy and Social Sciences Foundation of China (No.20BTQ065) and The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 16KJB520026).

References

1. http://www.cdpf.org.cn/sjzx/cjrgk/201206/t20120626_387581.shtml
2. Tarchanidis, K.N., Lygouras, J.N.: Data glove with a force sensor. IEEE Trans. Instrum. Meas. **52**(3), 984–989 (2003)
3. Han, Y.: A low-cost visual motion data glove as an input device to interpret human hand gestures. IEEE Trans. Consumer Electron. **56**(2), 501–509 (2010)
4. Hernandez-Rebollar, J.L., Kyriakopoulos, N., Lindeman, R.W.: A new instrumented approach for translating American Sign Language into sound and text. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 547–552. IEEE (2004)
5. Kim, J.H., Thang, N.D., Kim, T.S.: 3-D hand motion tracking and gesture recognition using a data glove. In: IEEE International Symposium on Industrial Electronics, 2009. ISIE 2009, pp. 1013–1018. IEEE (2009)
6. Gao, W., et al.: A Chinese sign language recognition system based on SOFM/SRN/HMM. Pattern Recogn. **37**(12), 2389–2402 (2004)
7. Deng, J., Tsui, H.T.: A Two-step Approach based on PaHMM for the Recognition of ASL. ACCV (2002)
8. Maraqa, M., Al-Zboun, F., Dhyabat, M., Zitar, R.A.: Recognition of Arabic sign language (ArSL) using recurrent neural networks. J. Intell. Learn. Syst. Appl. **2012**(4), 41–52 (2012)
9. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 889–894. IEEE (2004)
10. Liangguo, Z., et al.: A medium vocabulary Chinese sign language visual recognition system. Comput. Res. Dev. **43**(3), 476–482 (2015)
11. Vogler, C., Metaxas, D.: Toward scalability in ASL recognition: breaking down signs into phonemes. Gesture-based Communication in Human-Computer Interaction, pp. 211–224. Springer Berlin Heidelberg (1999)
12. Vogler, C., Metaxas, D.: Parallel hidden Markov models for American sign language recognition. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, pp. 116–122. IEEE (1999)
13. Utsumi, A., et al.: Hand gesture recognition system using multiple cameras. In: Proceedings of the 13th International Conference on Pattern Recognition, vol. 1, pp. 667–671. IEEE (1996)
14. Argyros, A., Lourakis, M.I.A.: Binocular hand tracking and reconstruction based on 2D shape matching. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 1, pp. 207–210. IEEE (2006)
15. Jang, Y.: Gesture recognition using depth-based hand tracking for contactless controller application. In: 2012 IEEE International Conference on Consumer Electronics (ICCE), pp. 297–298 (2012)

16. Chai, X., et al.: Sign language recognition and translation with Kinect. In: IEEE Conf. on AFGR (2013)
17. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with jointly calibrated leap motion and depth sensor. Multimedia Tools Appl. **75**(22), 14991–15015 (2015). https://doi.org/10.1007/s11042-015-2451-6
18. Raheja, J.L., et al.: Robust gesture recognition using Kinect: a comparison between DTW and HMM. Optik **126**(11), 1098–1104 (2015)
19. Jiang, X., Satapathy, S.C., Yang, L., Wang, S.-H., Zhang, Y.-D.: A survey on artificial intelligence in Chinese sign language recognition. Arab. J. Sci. Eng. **45**(12), 9859–9894 (2020). https://doi.org/10.1007/s13369-020-04758-2