International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 10, No. 7 (2017), pp.21-34
http//dx.doi.org/10.14257/ijsip.2017.10.7.03

Real-Time Dynamic Sign Language Recognition Based on Hierarchical Matching Strategy

Liang Wenle[1], Huang Yuanyuan[2*] and Hu Zuojin[3]

[1]Institute of Computer Sci. &Tech., Nanjing University of Aeronautics & Astronautics, Nanjing, China
[2]Institute of Computer Sci. &Tech., Nanjing University of Aeronautics & Astronautics, Nanjing, China
[3]Institute of math and information science, Nanjing Normal University of Special Education, Nanjing, China
[1]1281907942@qq.com,[2]805861040@qq.com, 3154928@qq.com*

Abstract

Dynamic sign language can be described by its trajectory and the key hand-action. However, a large number of statistical data show that most of the commonly used sign language can be recognized by its trajectory curve. Therefore, a hierarchical matching recognition strategy for dynamic sign language is proposed in this paper. First, the gesture trajectory can be obtained by the somatosensory equipment like Kinect. According to its point density an algorithm of key frame detection is designed and is used to extract the key gestures. Then the dynamic time warping (DTW) algorithm is optimized and used to do the first-level matching, i.e. trajectory matching. If the recognition results can be get currently, then the recognition process can be finished, otherwise the process should go into the second-level, i.e. key frame matching and get the final recognition results. Experiments show that this algorithm not only has good real-time performance, the recognition accuracy is also higher.

Keywords: *Dynamic Sign Language Recognition, Gesture Trajectory, Key Frame, Dynamic Time Warping*

1. Introduction

Hand gesture recognition is one of the key technologies in human computer interaction technology. At the same time, sign language is also the normal communication method between the deaf and the normal people. In China, there are about 2057 million deaf people, but the number of sign language interpreter is far difficult to meet the market demand. Therefore, the study of sign language recognition technology can not only let the deaf and normal people have barrier free communications, but also can improve the computer's ability to perceive and increase the channels and methods of human-computer interaction.

Gesture recognition based on computer vision is a challenging and interdisciplinary research topic. It is a front subject and research hotspot in the field of human computer interaction as well. Compared with the data glove, computer vision has the advantages of interactive mode more in line with natural habits, low cost, easy to promote and so on. Single camera was used in earlier time, but due to the inability to accurately locate the hands position, it can only be used when hands are marked or to identify those relatively simple static gestures [1-2]. In order to overcome the shortcoming that the single camera is difficult to capture the change of hand in three-dimensional space, multiple cameras were used to obtain images in different dimensions so as to make up this defect [3-4]. However, each time when such kind of method is carried out, calibration of all the

ISSN: 2005-4254 IJSIP

cameras is necessary which is very inconvenient. In recent years, since the appearance of the depth camera, the research of gesture recognition based on three-dimensional data has been greatly developed. The Microsoft Corp launched a somatosensory camera Kinect in 2010, which makes the use of depth information to identify sign language has become a trend. Jang and others proposed a system that uses Kinect to acquire depth information to identify the gesture. Based on the algorithm of continuously adaptive mean shift they used depth probability and update depth histogram to track hand position [5]. Chai and *etc.* used Kinect to obtain 3D features of hand gesture and realize the dynamic sign language recognition through matching the 3D gesture trace, and the average recognition rate can reach 83.51% [6]. Marin used Kinect to locate the hand area and furthermore get the fine information through Leap Motion. Then the support vector machine was used as the classifier. The recognition rate reached 91.28% [7]. At present, the use of Kinect to obtain the depth information to identify dynamic sign language has become the mainstream.

2. Existing Problems

1) Description of dynamic sign language. At present, trajectory and hand-shape are mostly used to describe the dynamic sign language. The dynamic sign language can be looked on as a set of different hand gestures. People's hands are flexible objects which have large degree of freedom and varied gestures. Therefore, uncertainty and space-time variability are difficult problems when trajectory is adopted to describe dynamic sign language. In addition to the motion trajectory, hand-shape also contains important semantic information of sign language. Currently, the description of the hand-shape is basically through detecting hand features of each frame in the sign language video. But in fact, not all of the hand-shapes contribute to the semantic meaning of sign language, and only the key hand-shapes which represent the key action are just the key semantic element of sign language [8]. Thus the more accurate description of the dynamic sign language should be the hand gesture trajectory plus the key hand-shapes.
2) Key frame extraction. Key frame is the frame of the key action in dynamic sign language, and the key action is the basic element of sign language semantics [8]. For a long time, hand-shapes in all the frames usually have been used to describe the dynamic sign language, so far the research of how to detect and extract the key frames in the case of natural human-computer interaction is rarely.
3) Similarity measure of dynamic sign language. Currently, the similarity measure of dynamic sign language is done by matching trajectory curve or matching trajectory curve plus all the hand-shapes. However, through a large number of statistical experiments we find that most of the sign language trajectory curve is very different, which means depending no the trajectory curve, we can distinguish a lot of dynamic sign language. But for those sign languages whose trajectory curves are similar, their key hand-shapes will have big difference. Therefore, in order to improve the recognition efficiency and precision, the trajectory and the hand-shape can be considered separately.

From the above we know that through the trajectory and the key hand-shapes, a kind of more accurate and stable description of dynamic gesture can be got. On this basis, reasonable design of similar measure algorithm can help to achieve accurate and efficient recognition.

3. Algorithm Design

A new hierarchical matching strategy for the dynamic sign language is proposed in this paper, *i.e.* first of all, only the trajectory of the gesture is matched which is called first-level matching, and if now can get result, then the recognition process is completed. Otherwise, followed by the matching of the key hand-shapes, *i.e.* the second-level matching, after this, recognition process can be finished. Such matching strategy can be realized depending on the information of both gesture trace and the key hand-shape. According to the point density feature of the track curve, an algorithm of key frame detection is put forward in the paper. Besides, the traditional DTW algorithm is optimized to make it more suitable for the similarity measurement of gesture trace.

3.1. Trace Acquirement

At present, the palm position is generally used to represent the trace of dynamic sign language. For example, in the [9] literature, palm position in each frame is extracted and thus a trajectory curve is formed. The frame rate of Kinect is 30 frames per second, and if the dynamic sign language “graduation certificate” lasts for 2.13 second, then totally 64 frames and 64 corresponding palm points can be obtained. If these points are connected on the time axis, the corresponding trajectory curve then can be got, as shown in Figure 1. As a convenience to display, the 3D data, *i.e.* x, y and z are respectively shown. The horizontal coordinate represents time, and the vertical coordinate is the corresponding space location. Each point on the curve, that is, the characteristic point of the trajectory, has 3D coordinate. Sign language can be divided into single-hand and both-hand. Those single-hand still can be looked on as both-hand if we assume the curve of the other no-move hand is only a point. Therefore, as for the gesture trace curve, each feature point should be seen as a 6-dimension feature vector. This 6-dimension feature vector includes the both two palms’ x, y and z coordinate.

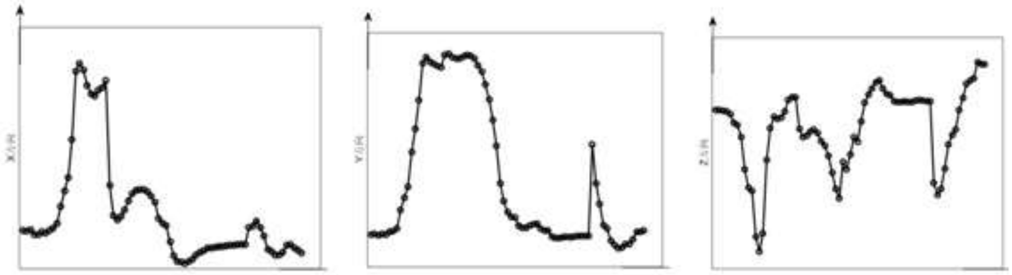


Figure 1. Trajectory Curve of “Graduation Certificate” (Left Hand)

3.2. Key Frame Detection

As for the hand-shapes of the dynamic sign language, it is not necessary to detect hand-shape in each frame of sign language video. Although a video of one dynamic sign language may contain dozens or even hundreds of images, not all of them have semantic information. Only those few key frames that correspond to the key actions carry semantic information, and the rest of the most frames are just transitional actions without any sense. The hand-shape in the key frame is defined as key hand-shape. From the point of stability, when different people play the same dynamic sign language, the key hand-shapes are relatively stable while those transitional hand-shapes are very varied. Consequently, when the key hand-shapes are used to represent dynamic sign language, it can not only reduce the amount of data and improve the recognition efficiency, but also makes the description of sign language more stability and accuracy.

Through a lot of experiments, we find a statistical law, *i.e.* when people play dynamic sign language, in order to explicitly clear the semantics, the player may naturally emphasize the key actions, and the way in the manner is just stay for a longer time in the key action. This stay is only relative to the transition action. It is not a deliberate act, but an instinctive reaction when player doing the sign language expressing. When this instinctive reaction is reflected in the trajectory curve, we can see that near the time of the key action, the feature points are relatively very intensive. According to such a statistical rule, an algorithm of key frame detection based on the point density in gesture trace is proposed.

Suppose there is a gesture trace curve P. If we want to define the point density of point X in P, we should get a points set T_X in which all the points are near to X.

$$T_X = \{Y \mid \forall Y \in P, \delta(X,Y) \le \Delta\} \quad (1)$$

The $\delta(X,Y)$ is the distance between point X and Y in curve P, and the Δ is threshold. As for the T_X, the points belong to it not only in the space position to be similar, but also must be continuous in time. But in the formula (1), the continuity of time is ignored. Thus, during the course of dynamic sign language expressing, the T_X of the point where hands pass many times will be very large. However, this is not in conformity with our original intention for the density definition, so it is necessary to add the limitation of time. Suppose there are n points in T_X, based on the time order, these points are $T_X = \{Y_{t1}, Y_{t2}, ..., Y_{tn}\}$. The subscript ti is the point Y's number in the time series, so the density of point X is defined as the maximal subset of T_X, and this maximal subset should contain most points Y whose subscripts are continuous in time.

$$crow(X) = \max\{|\, Y_{ti} \,|, ti \text{ is continuous, and } |Y_{ti}| \subseteq T_X \quad (2)$$

As for a given dynamic sign language, since it is played by the both two hands, therefore, in order to get the total densities of all the points in trace curve, both the left-hand curve and the right-hand curve should be computed densities for each point firstly, and then add them together. The Figure 2(a) is the density curve of the dynamic sign language “graduation certificate”. The horizontal coordinate represents the total number of frames, and the vertical coordinate is density value of the palm point in corresponding frame. Theoretically, in this curve, the point whose density is larger than a certain threshold value should be corresponding to the key frame. Selection of the threshold can be in several ways, *i.e.* first, the average density value of all the points can be the threshold. Second, the median value of density can be used as threshold. Third, the average value of maximum and the minimum density may be adopted as threshold. From the Figure 2(a) we can see many points of intensity are more than the threshold, but in fact, these points which have larger density are not always corresponding to the key frames. So it can be considered that the intensity curve shown in Figure 2 (a) can be divided into a number of intervals, and in each interval to find the maximum density which is larger than the threshold value. The points corresponding to these selected maximum densities are just the candidate key frame. So, how to divide the intensity curve? Since each key frame contains some semantics, in order to ensure not missing any key frames, the number of intervals should be 1.5 to 2 times of the possible maximum amount of the key frames. Through the study of Chinese sign language, we find that as for the vast majority of sign language words, the number of their key frames is not more than 5 [10]. As shown in Figure 2 (b), the density curve in Figure 2 (a) is evenly divided into 7 intervals, and thus 6 density extreme points are obtained. Based on these points, 6 corresponding candidate key frames can be got, as shown Figure 3.

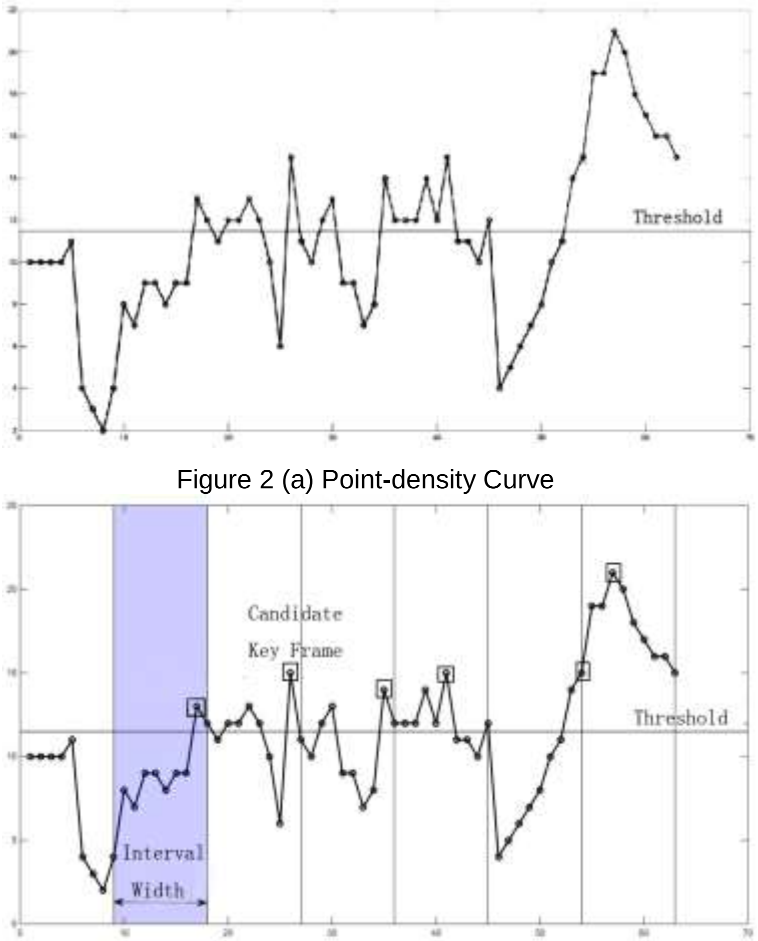
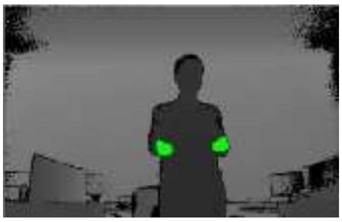


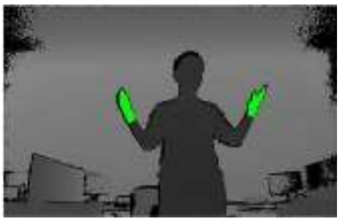
Figure 2 (a) Point-density Curve

Figure 2 (b) Interval Division for the Point-density Curve

Figure 2. Point-density Curve of “Graduation Certificate”

We may get more candidate key frames than the actual number because the interval number of curve division is more than the actual number of key frames. Theoretically, hand-shapes in successive key frames should be certainly not similar. According to this characteristic, we can use frame subtraction method to remove redundant candidate frames. Please note, the last two frames in Figure 3 are the state of end gesture and haven’t any semantics. But due to the shooting time, density of the end point will be relatively large, so it is always selected in the sequence of candidate key frames. However, because the actions of sign language are generally played in the position above the waist, these end gesture can be eliminated by this prior knowledge. After the above steps, we can get the final key frames, and they are the number 1, 2, 4 in Figure 3, which are just in line with the actual situation.





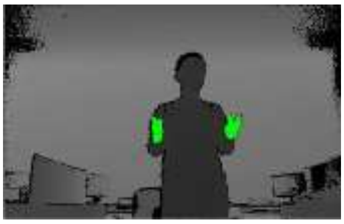


Figure 3(a) Key Frame One Figure 3(b) Key Frame Two Figure 3(c) Key Frame Three

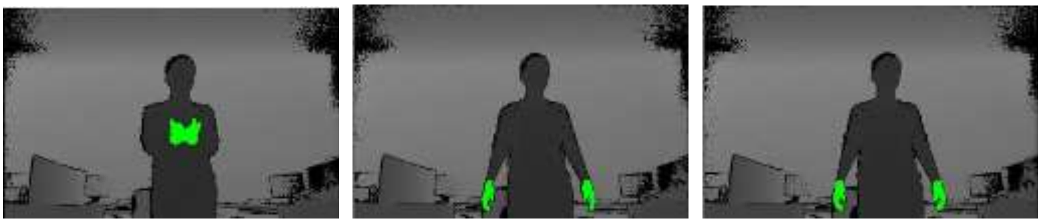


Figure 3(d) Key Frame Four Figure 3(e) Key Frame Five Figure 3(f) Key Frame Six

Figure 3. Candidate Key Frames of the "Graduation Certificate"

3.3. Normalization of Trace Curve

Different signers may have different action habits, so the trace curves of different signers for the same gesture will be varied in duration, speed, range of motion and so on, which makes these curves similar in general trends but significant different in the local details. This will bring many problems to the similarity measure of trajectory. Therefore, it is necessary to normalize the trajectory curve to eliminate these differences. Suppose there are two gesture trace curves P and Q. If we want to compute the similarity between P and Q while the P is seen as template, the Q should be normalized according to P firstly. Suppose the numbers of feature points included in P and Q are respectively m and n, *i.e.* $P=\{p_1, p_2, ..., p_m\}$, and $Q=\{q_1, q_2, ..., q_n\}$. Thus the steps of normalization should be as follows:

1) First, the scaling factor ρ is calculated.

$$\rho = \frac{\| \sum_{i=1}^{m} \frac{p_i}{m} \|}{\| \sum_{j=1}^{n} \frac{q_j}{n} \|} = \frac{n \| \sum_{i=1}^{m} p_i \|}{m \| \sum_{j=1}^{n} q_j \|} \tag{3}$$

2) After normalization, the subscript index of points in Q, *i.e.* $i\,(1 \le i \le n)$ will become $i^{'} = i * \frac{m}{n}$, $(1 \le i^{'} \le m)$. If i' has already existed, then $q_{i'}$ should be put value according to formula (4). Otherwise, $q_{i'}$ ought to be assigned in accordance with formula (5).

$$q_{i'} = (\rho * (q_i - q_1) + p_1 + q_{i'})/2 \tag{4}$$

$$q_{i'} = \rho * (q_i - q_1) + p_1 \tag{5}$$

3) The subscript index i' is rounded to the nearest digit, so it may not be continuous, thus there may remain some points q_t which are not be valued after step(2). As for these points, interpolation operation should be done based on formula (6)

$$q_t = (q_{t-1} + q_{t+1})/2 ,（1 < t < m） \tag{6}$$

From above, we can get normalized curve Q in accordance with the template P. At the moment, P and Q have the same starting point and the same number of feature points, which is conducive to do similarity measurement.

3.4. First-Level Matching Based on the Trajectory

The trajectory of dynamic sign language can be looked on as a kind of time series. As to classify and identify the time series, the hidden Markov models (HMM) and dynamic time warping (DTW) are two commonly used algorithms presently. In the literature

[11,12], the DTW and HMM algorithms are compared in the effect of gesture recognition, and then DTW is recommended to use in dynamic sign language recognition. However, the traditional DTW algorithm does not take into account the key frames' more important semantic contribution, in stead, it treats all frames equally. This obviously does not meet the actual situation of dynamic sign language. Accordingly, a kind of weighted DTW algorithm based on key frames is proposed here, which can improve the accuracy of dynamic sign language recognition.

Suppose there are two normalized curves P and Q, which contain M feature points. P has m key frames and the set of index of the key frames is $KP = \{kp1, kp2, ..., kpm\}$. Q includes n key frames and its set of key frames index is $KQ = \{kq1, kq2, ..., kqn\}$. In curve P, one point p_i has a distance $\delta_{KP}(i)$ to the point of key frame, and the $\delta_{KP}(i) = \min\ \{|i - e|, e \in KP\}$. In the same way, point q_j has a distance $\delta_{KQ}(j) = \min\{|j - e|, e \in KQ\}$ to the point of key frame in curve Q. Then the recurrence formula (7) can be used to compute the cumulative cost matrix D

$$D(i,j) = (|\delta_{KP}(i) - \delta_{KQ}(j)| + 1) \times \delta(i,j) + \min\begin{cases} D(i-1, j) \\ D(i-1, j-1) \\ D(i, j-1) \end{cases} \tag{7}$$

D is a $M \times M$ matrix, in which i and j are ranks number. $\delta(i,j)$ is the distance between p_i and q_j, usually the Euclidean distance. As for a matching path $W = \{w_1, w_2, ..., w_l\}$, if $w_k = (i,j), 1 \le k \le l$, it means in this matching path, p_i and q_j are matched. Compared the $\delta_{KP}(i)$ and $\delta_{KQ}(j)$, if they are equal, then there are two situations. First, p_i and q_j approach their respective key frame points from the same direction, their matching weight will be higher and the corresponding coefficient will be relatively small. Second, p_i and q_j close to their respective key frame points from the different direction, although the corresponding coefficient still be relatively small, coefficients of other matched points in this matching path will become larger. Thus the algorithm can reflect the constraint of the key frame. Since the gesture trace curve is normalized, the two curves to be matched are in the same length. If the two curves correspond to the same dynamic sign language, the positions of their corresponding key frames will be equal or near. Accordingly, the distance between such two curves will be smaller which makes recognition easier and vice versa.

Based on this optimized DTW algorithm, the gesture trace curve to be identified can be matched with all the templates, and thus got the minimum distance $d_{\min 1}$ and sub small distance $d_{\min 2}$. If $\frac{d_{\min 1}}{d_{\min 2}} > \alpha$, we can get recognition result, *i.e.* the class corresponding to the $d_{\min 1}$. Otherwise, selecting k classes corresponding to the first k minimum distance, we should go into the second level matching based on the key hand-shape. The value of α may have a great impact on the identification process. It is usually within 1.2 to 2.

3.5. Second-Level Matching Based on the Key Hand-Shape

Some dynamic sign languages have very similar trace curves, such as "wife" and "husband", "son" and "daughter" and so on. As for these dynamic sign languages, it is difficult to make accurate identification through trace matching. It is necessary to do further analysis and comparison based on their key hand-shapes. In the key frame, with the depth and skin color information, hand area can be detected and extracted features. As shown in Figure 4, each key hand-shape can be looked on as a N - dimensional feature vector H .

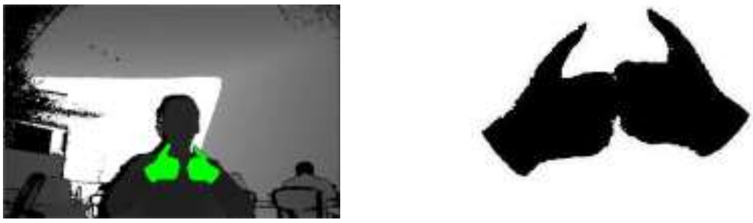


Figure 4(a) Key Frame Figure 4(b) Key Hand-shape

Figure 4. Key Hand-shape

Assuming there is a dynamic sign language S to be identified. It contains m key frames, *i.e.* m key hand-shapes. Thus S can be seen as a set of feature vectors, $S=\{H_{S1},H_{S2},...,H_{Sm}\}$. After the first –level matching, k dynamic sign language classes which are most similar with S in trace can be got, *i.e.* C_1 , C_2 , …, C_k . Now, using C_1 as an example, how to compute similarity between S and C_1 in the key hand-shape? Suppose C_1 has n key frames, *i.e.* $C_1=\{ H_{C1},H_{C2},...,H_{Cn} \}$. Theoretically, if S belongs to C_1, then m should equal to n . But in fact, It's hard to guarantee. According to the algorithm of key frame detection mentioned above, usually m is lager or equal to n . We can design a $m\times n$ matrix D , in which the element d_{ij} $(1\le i\le m,1\le j\le n)$ is just the distance between H_{Si} and H_{Cj} . S and C_1 can still be seen as time series composed of key frames, thus the DTW algorithm can be worked on them. We should find a matching path in D , *i.e.* $L=\{(x_1,1),(x_2,2),...(x_n,n)\}$, and $1\le x_1<x_2<...<x_n\le m$, in which the accumulation of d_{ij} $((i,j)\in L)$ should be smallest. Thus the distance between S and C_1 is defined as:

$$Dis(S,C_1)=\frac{\min\{\sum_{(i,j)\in L} d_{ij}\}}{\min(m,n)} \tag{8}$$

More smaller this distance is, more similar the S is with C_1, and vice versa. In the same way, we can compute distance between S and all the other classes. If the minimum distance is still smaller than rejection threshold, the class corresponding to the smallest distance should be the final recognition result. Otherwise, it can not be recognized.

4. Experimental Analysis

The experiment is divided into two parts. The first is the verification of the key frame detection algorithm, and the second is verification of dynamic sign language recognition. Five people are trained to play 60 commonly used sign language vocabularies. Among these five people, one boy and one girl are selected as templates, and remaining 3 people are just test objects.

4.1. Experiment of Key Frame Detection

First of all, the stability of the algorithm is verified, that is, whether the algorithm can give the same or roughly the same key frames for those same dynamic sign languages played by different people. For instance, we select three people to play dynamic sign language “son”. As for the first people, the point density curve and the corresponding location of candidate key frames are shown in Figure 5(a). Figure 5(b) shows the candidate key frames except the frames in which the position of hands is lower than waist. After frame subtraction method, the final set of key frames is shown in Figure 5(c). Figure 6 and Figure 7 are the results of other two people. We can see, for these three different people, although their gesture traces have obvious difference, since the traces express the same meaning, the key frames detected by our algorithm are same or roughly the same. We have detected key frames for all the 60 dynamic sign language in the same way, the experimental results show that this method has good stability and accuracy.

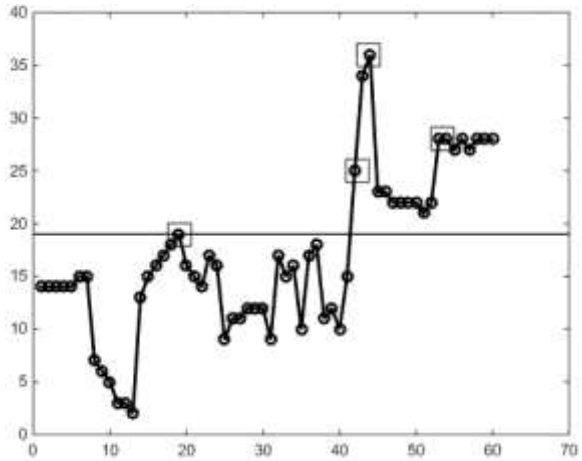


Figure 5(a) Point Density Curve of Dynamic Sign Language “Son” of the First Player

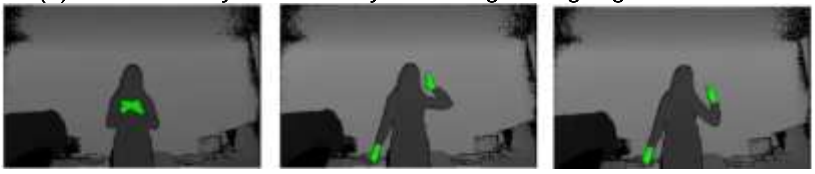


Figure 5(b) Candidate Key Frames

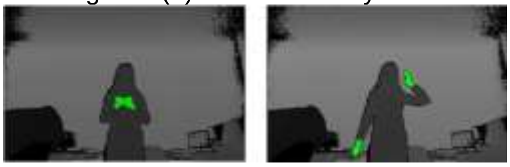


Figure 5(c) Final Key Frames

Figure 5. Key Frame Detection of the First Player

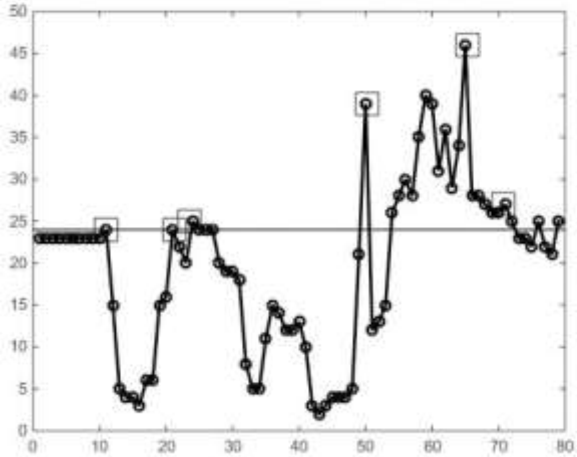


Figure 6(a) Point Density Curve of Dynamic Sign Language “Son” of the Second Player

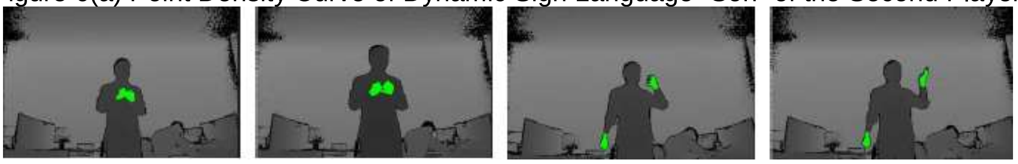


Figure 6(b) Candidate Key Frames

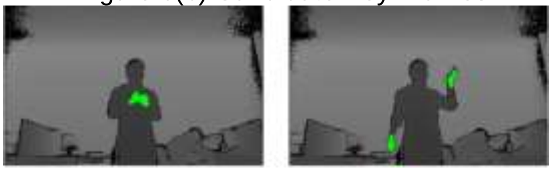


Figure 6(c) Final Key Frames

Figure 6. Key Frame Detection of the Second Player

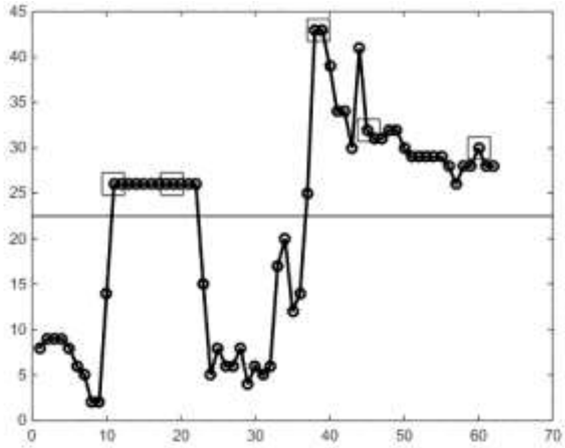


Figure 7(a) Point Density Curve of Dynamic Sign Language “Son” of the Third Player

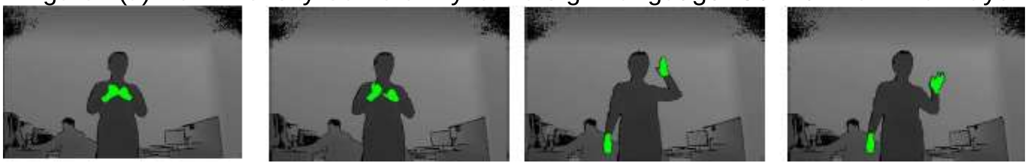


Figure 7(b) Candidate Key Frames

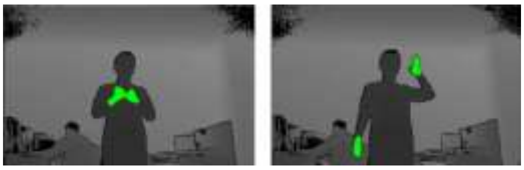


Figure 7(c) Final Key Frames

Figure 7. Key Frame Detection of the Third Player

In order to verify the effectiveness of the proposed algorithm in this paper, we use the clustering algorithm in the literature [13] to do the key frame extraction too. Table 1 shows the comparison between these two algorithms in aspects of key frame number and running time when miss rate is zero.

From the Table 1, the algorithm of this paper will undoubtedly have a great advantage. In literature [13], the algorithm used unsupervised clustering method and did not take into account the characteristics of dynamic sign language. It extracted features for each frame and did clustering, so the computation quantity is very large. What's more, in order to ensure no missing of frame, the number of clusters is also relatively high. Although the result may contain all the key frames, it includes more useless transitional frames as well. As for algorithm proposed in this paper, based on the signer's psychological sense and unique features of dynamic sign language, density of the points on the trajectory curve is used to distinguish the key frames and the transitional frames. The calculation is very small, while the accuracy rate is higher. In addition, because the extraction of the key frame depends on the density of the trajectory points, the efficiency of this method will not decrease when the sign language becomes longer and the number of the track points increases. On the contrary, with the increase of the amount of data, unsupervised clustering algorithm will spend more time, and the efficiency will be significantly decreased as well.

Table 1. Experimental Data of Key Frame Detection

	Algorithm proposed in paper	Algorithm in the literature [13]
Average number of key frames	3.4	26
Average running time(s)	0.054	12.978

4.2. Experiment of Sign Language Recognition

In order to verify the effectiveness of the hierarchical matching strategy and the weighted DTW algorithm based on the key frame, we did three sets of contrast tests.

In the first test, trajectory and hand-shape in all frames are used to describe dynamic sign language, and traditional DTW algorithm is adopt to do classification and recognition.

As for the second test, trajectory and the key hand-shape are used to represent dynamic sign language, and still the traditional DTW algorithm is to do classification and recognition.

Finally, in the third test, trajectory and the key hand-shape are used to describe dynamic sign language, and the weighted DTW algorithm based on the key frame is used instead.

In the last two experiments, the hierarchical matching strategy is used. If the first-level matching can not give recognition result, then the first 5 classes which are most similar with the sample in gesture trace will be returned to do the second-level matching subsequently.

The experimental results are shown in Table 2. From the Table 2 we can see that, the first experiment, whether in recognition efficiency or the accuracy, is lowest. Although it only matches for one time, the amount of data is large due to the features of all frames, besides, the hand-shape in transitional frame is very unstable and the number of transitional frame is far more than that of key frame, so the recognition result is not ideal. The last two kinds of experiments use hierarchical matching strategy. Although they did matching for two times, data volume in each level is not so large, their recognition efficiencies are higher. For instance, in the first-level matching, only the trace curve is to be recognized, whose feature dimension is low, only six. In the second-level matching, the number of key frame is very small. Accordingly, these two kinds of recognition have better real-time performance. The difference between the second and the third test lies in the performance of traditional DTW algorithm and the weighted DTW algorithm. Due to the full consideration of the unique characteristics of dynamic sign language, the work of key frames is highlighted. Therefore, the last experiment has improved both the efficiency and accuracy of recognition. On the other hand, since the good stability of the key frames, not only the trained signer can be recognized in high accuracy rate, those non specific populations who learn the dynamic sign language temporarily can also be recognized and the recognition rate can reach more than 80%. It should be pointed out that, for the non - specific population, the recognition accuracy is not stable because of the different standard degree of action.

Table 2. Contrast Experiments of Different Recognition Methods

	Recognition time（s）	Recognition accuracy
First set of experiments	4.2	0.7833
Second set of experiments	0.503	0.8667
Third set of experiments	0.396	0.9672

5. Concluding Remarks

In this paper, a new idea of dynamic sign language recognition is proposed. First, the dynamic sign language is described in two aspects, gesture track and the key hand-shape. Second, a kind of hierarchical matching strategy is designed thus the trace and key hand-shape are matched separately. Finally, the traditional DTW algorithm is optimized to make it more suitable to recognize dynamic sign language. Due to the full consideration of the unique characteristics of dynamic sign language, the experiment according to this new idea has achieved good results. On the basis of the recognition of dynamic sign language words, the further work of identifying more complex sign language can be carried on in the future.

Acknowledgements

Project supported by national natural science foundation of china, 61375021.
Project supported by “Shuang Chuang” foundation of Jiangsu Province.
Project supported by “Excellent PhD” foundation of Yangzhou City, Jiangsu Province.

References

[1] T. Starner and A. Pentland “Real-time american sign language recognition from video using hidden markov models”, Motion-Based Recognition. Springer Netherlands, **(1997)**, pp. 227-243.

[2] M. Maraqa, F. A. Zboun, M. Dhyabat and R. A. Zitar, “Recognition of Arabic Sign Language (ArSL) Using Recurrent Neural Networks”, Journal of Intelligent Learning Systems and Applications, vol. 4, **(2012)**, pp.41-52.

[3] A. Argyros and M. I. A. Lourakis, “Binocular hand tracking and reconstruction based on 2D shape matching”, Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. IEEE, vol. 1, **(2006)**, pp. 207-210.

[4] C. Vogler and D. Metaxas, “Parallel hidden markov models for american sign language recognition”, Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. IEEE, vol. 1

(1999), pp. 116-122.

[5] Y. Jang, "Gesture recognition using depth-based hand tracking for contactless controller application", 2012 IEEE International Conference on Consumer Electronics (ICCE), **(2012)**, pp. 297-298.

[6] X. Chai, G. Li and Y. Lin, "Sign language recognition and translation with Kinect", IEEE Conf. on AFGR, **(2013)**.

[7] G. Marin, F. Dominio and P. Zanuttigh, "Hand gesture recognition with jointly calibrated Leap Motion and depth sensor", Multimedia Tools and Applications, vol. 4, no. 2, **(2015)**, pp. 1-25.

[8] S. Yulin, "Analysis of Chinese sign language morpheme", Chinese Journal of Special Education, vol. 1, **(1993)**, pp.1-13.

[9] P. Doliotis, A. Stefan and C. McMurrough, "Comparing Gesture Recognition Accuracy Using Color and Depth Information", Proceedings of the Fourth International Conference on Pervasive Technologies Related to Assistive Environments(PETR) Crete ,Greece, **(2011)**, pp.1-7.

[10] L. Shurong, H. Yuanyuan, H. Zuojin and Daiqun, "Key Frame Detection Algorithm based on Dynamic Sign Language Video for the Non Specific Population", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 12, **(2015)**, pp. 135-148

[11] J. Carmona and J. Climent, "A performance evaluation of hmm and dtw for gesture recognition", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, vol. 12, no. 2, **(2012)**, pp. 236-243.

[12] J. L. Raheja, M. Minhas and D. Prashanth, "Robust gesture recognition using Kinect: A comparison between DTW and HMM", Optik-International Journal for Light and Electron Optics, vol. 126, no. 11, **(2015)**, pp. 1098-1104.

[13] Y. Zhuang, Y. Rui, T. S. Huang and S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering", Proc. of IEEE Int. Conf. on Image Processing, **(1998)**, pp. 866-870.

Authors

Liang Wenle, Male, Han nationality, born in 1989, Master graduate student, now studying at the Nanjing University of Aeronautics & Astronautics, college of computer science and technology, and the main research direction is pattern recognition and image processing.

Huang Yuanyuan, Female, Han nationality, born in 1975, associate professor in Nanjing University of Aeronautics & Astronautics, and now is an enterprise postdoctoral in Jiangsu Nankai Star Software Technology Co. Ltd. The main research direction is recognition and image processing.

Hu Zuojin, Male, Han nationality, born in 1965, professor, now working at the Nanjing Normal University of Special Education, and the main research direction is data processing and machine learning.