International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.12 (2015), pp.135-148
http://dx.doi.org/10.14257/ijsip.2015.8.12.14

Key Frame Detection Algorithm based on Dynamic Sign Language Video for the Non Specific Population

Li Shurong[1], Huang Yuanyuan[1], Hu Zuojin[2] and Dai Qun[1]

[1]*Institute of Computer Sci. & Tech., Nanjing Univ. of Aero.& Astronautics, Nanjing, China*
[2]*College of arts and science, Nanjing Technical College of Special Education, Nanjing, China*
li_shurong@foxmail.com, hyy_nust@sohu.com (contact email), zou.c@163.com,daiqun@nuaa.edu.cn

Abstract

The current recognition algorithms of sign language, or can only identify static gestures, or need data gloves, position sensor and other additional auxiliary equipments, which are only used for laboratory research and some special occasions. Therefore, they are not conducive to the promotion of widely use. A new idea of sign language recognition based on key frames is presented in this paper. The dynamic sign language can be looked on as a series of static gestures, which can be called the key frames. Through the key frame sequence detection and identification, the sign language can be rapidly recognized. So an algorithm of key frame detection especially for the dynamic sign language is proposed. This adaptive method uses image difference and classification theory in pattern recognition to extract key frames from video, and in addition to PC machines, the entire process requires only a camera, which is very easy to use. Experiments show that the key frames obtained by this way have good stability and accuracy, thus the real-time recognition of dynamic sign language can be realized.

Keywords: *Dynamic Sign Language, Sign Language Recognition, Key Frame Detection, Adaptive*

1. Introduction

With the rapid popularization of computer and the Internet, the human-computer interaction has become an important part of people's daily life. Gesture is a kind of natural and visualized means of communication between human and computer, so gesture recognition is one of the key technologies to realize the upgraded interaction between human and computer. At the same time, sign language is a way for deaf people to communicate and exchange ideas through finger alphabet and gestures instead of language [1]. Therefore, research on sign language recognition technology can make the communication between deaf and healthy people more convenient and accessible.

Based on computer vision, the Artificial Neural Network (ANN) and Hidden Markov Model (HMM) [2-5] are widely used in the current sign language recognition technologies. ANN is usually used to recognize the static sign language. Manar identifies Arabia static sign language through a recurrent neural network, where the signers wore highlight gloves and the recognition rate is up to 95.11% [6]. Kouichi uses back propagation neural network to recognize Japanese finger language, and the recognition rate reaches 71.4% the training sample, while 47.8% on the test sample [7]. Stergiopoulou adopts a kind of self-growth and self-organizing neural network to identify static sign language, and based on training samples, the recognition rate arrives at 90.45% [8]. Although ANN is of eminent characteristic of classification and anti-interference capability, its ability to deal

ISSN: 2005-4254 IJSIP

with time series is not very strong. Therefore, it is not suitable for dynamic sign language recognition [9]. However, HMM is a statistical model and has strong ability to describe variability of time and space. Thus it has been widely used in the recognition of character and voice, and now also plays a dominant role in the dynamic gesture recognition. Starner adopted HMM to identify sentences and the recognition rate was 92% [10]. Under the condition that the signers wore a hat equipped with a camera and the grammar in the experiment was restricted, the rate can reach 98%. Grobel and Assan used HMM to identify 262 independent American vocabularies of sign language, where signers also wore color coded gloves. In this situation, the recognition rate was up to 91.3% while it was only 56.2% to 47.6% when samples without training were used [11]. Vogler and Metaxas employed context dependent HMM to identify 486 consecutive sentences, during which the input devices are three mutually vertical cameras, and the recognition rate was 89.9% [12]. Wang and his companions extracted HOG (Histogram of Oriented Gradient) features through Adaboost algorithm and then used HMM to recognize gesture language, and in this condition the rate can reach 98.9% [13]. But this method was sensitive to light so the rate was not that stable. Though the identification efficiency of dynamic sign language is high when HMM is used, such kind of algorithm can work only when the gesture area and gesture motion have been successfully detected [14]. Thus its robustness depends on the detection and tracking of the gesture. Therefore, it is difficult to realize real-time identification without position sensors and other auxiliary equipments, which does not meet the requirements of natural interaction between human and computer, and it also restricts the promotion of sign language recognition technology.

In order to improve the efficiency of recognition, the most direct way is to reduce the amount of data, so we might as well look on the dynamic sign language as a combination of several static gestures. Therefore, we can identify the corresponding dynamic sign language through these static gestures. Moreover, the static sign language does not have any time information, so one image is enough to describe a static gesture. Through the analysis of the “China sign language”, we know that 1-5 gestures would be sufficient to fully describe a sign language. However, a sign language video generally may include 40 to 200 images. That is to say we could eliminate the majority transition frames in the video and keep only 1-5 static gestures, *i.e.*, the key frames to represent the dynamic sign language. So, by detecting the key frames, rapid recognition of sign language may be achieved.

2. Related Works

Key frame is the key action during the motions or changes of bodies, which requires not only the feature of representativeness, but also the exact sampling frequency according to different sign language actions. There are some relevant researches and applications in the field of content based on video retrieval and classification in terms of the algorithms of key frames detection in video. The simplest way is to use uniformly-interval sampling frequency upon the original motion. But this method will lead to data redundancy caused by over sampling when motions change slowly. On the contrary, it may lose details due to under sampling if motions change drastically. Therefore, researchers generally use adaptive sampling method, which can adjust sampling frequency automatically according to the motions’ characteristics, thus the problems in the uniformly-interval sampling can be solved. At present, the adaptive sampling methods can be divided into the following three categories.

(1) Algorithm of simplified curve. Lim first proposed this way based on curve simplification [15]. The movement data of the human bodies was regarded as the points on the high dimensional curve, and the key frames were extracted through curve simplification algorithm. But it is required, in this method, to employ extra magnetic

devices to acquire position information of human body and movement information of human articular.

(2) Frame subtraction method. Shen Junxing and some other people specified the first frame of the dynamic data as the key frame, and then cut the subsequent frames which had less distance from the key frame. And when the distance was greater than the threshold, this frame would be taken as the current new key frame and was then continually used to do the cut [16]. While Togawa and Li applied linear interpolation between adjacent frames to remove the frames whose interpolation error is the minimum, and then the rest frames whose number meted with the requirement are the key frames [17-18]. Liu and others computed reconstruction error based on frame subtraction algorithm, according to it, the key frames which had optimal compression ratio were extracted [19]. But these key frames may not have the least reconstruction error.

(3) Clustering method. Zhuang and others proposed an unsupervised clustering algorithm to extract key frame [20]. In the method, similar frames were classified as one class through computing distance between feature vector and clustering center. Naveed Ejaz and others extracted key frames from video by using clustering method according to the color features [21]. Clustering is a commonly used algorithm to detect key frames, but the results mostly depend on the threshold. What's more, the adaptive algorithms were generally adopted since the number of clusters cannot be determined in advance, so great computation is adopted, which will cause poor calculation efficiency. In the video retrieval and classification, such kind of method is in generally used offline which are difficult to meet the requirements of real-time.

From what has been stated above, we have learned that the existing key frame detection algorithms are not put forward to recognize dynamic sign language. Moreover, it is only fit for simple human movement, which does not make full use of the characteristic of sign language. In addition, most of these algorithms require the user to set an appropriate threshold while extracting the key frames. This will bring inconvenience to ordinary users who only focus on the recognition results. Therefore, it is necessary to design a key frame detection algorithm specifically for dynamic sign language. This algorithm can extract key frame adaptively and rapidly, and it is not necessary for users to set any important parameters.

3. Algorithm Design

Through the analysis of a large number of commonly used sign language video, we find that under normal circumstances, a sign language video continues just a few seconds. Assuming that the camera is 25 frames per second, then a 3-second video will contain 75 frames, and then the amount of frame will up to 150 if the camera is 50 frames per second. By browsing these frames one by one, it is not difficult to find that, in the vast majority of images, the human's hands are in the transitional states which do not have any description functions. The number of frames which can really describe the semantics of gestures will be less than 10. These images which can describe the semantics are called the key frames. One key frame just symbolizes a static hand gesture, and the sequence of key frames without any transition frames can be used to represent a dynamic language. This can reduce the amount of data sharply and thus make it possible to improve the efficiency of recognition dramatically. So the key frame extraction is vital for the whole recognition system.

3.1. Region Detection of the Gesture

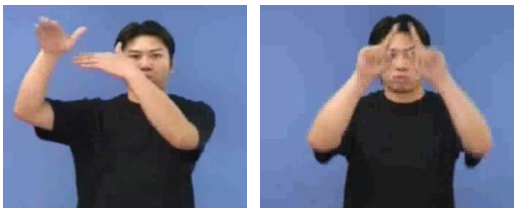

Figure 1. Dynamic Language Example

It is very important to detect the gesture area for realizing the accurate extraction of the key frame. Based on the gesture area, key frame and transition frame can be distinguished. Skin color detection and image difference are widely used in current algorithms. Skin color detection algorithm is relatively mature and easy to implement. However, in practical application, the color of face and hand is very close. So when the hand and face position dose not coincide with each other, they can be distinguished by the location relationship. But in many sign gestures, there would be a coincidence in the color of hand and face region, as shown in Figure 1(images from sign language video screenshot on the Chinese University Hong Kong website). In this case, it will be very difficult to distinguish hands and face without any other auxiliary equipment, but a camera can do some help. People use additional position sensors, or ignore the problem, or do not consider the difference in the existing algorithms, and they just detect key frames according to the characteristics of the whole image. But we must face this problem in the dynamic sign language recognition. The detection of the gesture area needs to be carried out in each frame of sign language video, so it is not appropriate for the algorithm to be very complex. Therefore, we use a simple way in this paper, *i.e.*, let user wear a pair of pure color gloves and even the ordinary gloves will work. Then no matter what kind of the sign language gestures is, the combination of color detection and image difference can eliminate interference and detect gesture area accurately and rapidly. Therefore the efficiency of the algorithm is guaranteed. Figure 2 is an example. Of course, now there is a kind of somatosensory camera which can capture the depth information. With this camera, the user can detect hands and face more easily. But now such kind of camera is not highly popular, and the algorithm in this paper can be extended to the use of somatosensory camera case.

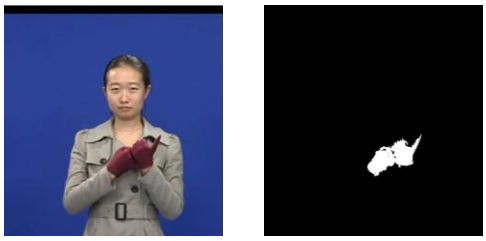

Figure 2. Gesture Region Detection Example

3.2. Detection of Key Frame

For example, the video of sign language “support ", is 3.72 seconds long, camera’s sampling frequency is 25 frame / sec, so there are totally 93 frames. Figure 3 is just a main fragment of the video, from the twenty-fifth frame to the sixty-sixth. In this fragment, the key gestures describe the different positions of two fists. The right one is clenching, and the other one’s thumb is out. The key frames may be among c34-c36, c37-c40, c41-c49 and c50-57. From Figure 3, we can see that these key frames in the video sequence are not the only ones. The successive frames in the video are very similar to one another because

of the coherence of the gesture, and thus they have strong correlation. So what we have to do is to select the key frames, and remove the associated frames as well as the transition frames. It can be achieved by the classification idea in pattern recognition theory. Firstly, since the key frames are not continuous in time, there must be several continuous transition frames between two adjacent key frames. Secondly, adjacent key frames do not resemble one another, while those continuous transition frames must be very close. Then an adaptive detection algorithm of key frame is proposed in this paper, and the process is just as follows:

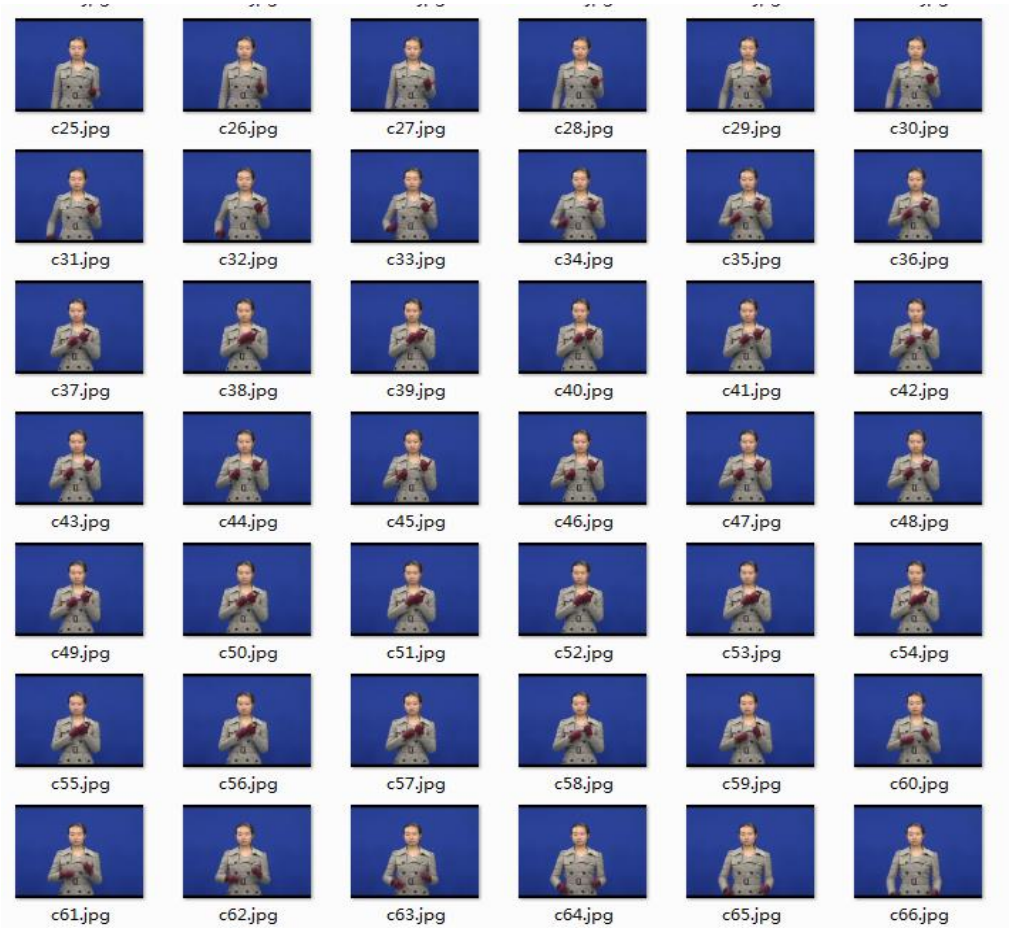

Figure 3. Video Sample if Dynamic Sign Language

(1) Assuming that after the color detection, there is an image sequence of sign language video in time order, $I = \{I_1, I_2, ..., I_n\}$. Firstly, consider I_1 as the reference. Subtracting with the reference, the subtract with it, and thus can get a difference values of all the images can be obtained. Then sorting these values, we can get an ordered sequence of the difference values, *i.e.*, $D = \{d_1, d_2, ..., d_n\}$, where $d_i < d_{i+1}$. Therefore, the image corresponding d_1 is the most similar one to I_1. On the contrary, d_n has the maximum difference from the reference image.

(2) Assuming T is the threshold. When the difference between I_i and I_1 is smaller than T, it means I_i is similar to I_1, and vice versa. Usually, the first image I_1 is not a key frame since it is just on the course of starting and has no semantic meaning. So the first key frame must be the image which is not close to I_1. According to this, the sequence

D should be divided into two types, *i.e.*, D_1 and D_2. In D_1, all the difference values are smaller than T while they are lager than T in D_2, which means D_1 and D_2 are two classes of which one is similar to I_1 and the other is different from I_1. *i.e.*, $D_1 = \{d_1, d_2, ..., d_j\}$, $D_2 = \{d_{j+1}, d_{j+2}, ..., d_n\}$, $d_j \le T < d_{j+1}$. We know that the first key frame must exist in D_2

(3) How to determine the threshold T ? In different sign language videos, the sequences may differ greatly, so the adaptive method should be used to compute the T. An objective function can be set as:

$$E(t) = \frac{t}{n}(m_1(t) - m_0)^2 + \frac{n-t}{n}(m_2(t) - m_0)^2, \quad t = 1,2,..., n \tag{1}$$

$$m_1(t) = \frac{1}{t}\sum_{i=1}^{t} d_i \tag{2}$$

$$m_2(t) = \frac{1}{n-t}\sum_{i=t+1}^{n} d_i \tag{3}$$

$$m_0 = \frac{1}{n}\sum_{i=1}^{n} d_i \tag{4}$$

When the parameter t makes the function $E(t)$ get its maximum value, the corresponding d_t will just be the threshold, *i.e.*, $T = d_t$.

(4) Based on the threshold T, we can get $D_1 = \{d_1, ..., d_t\}$, and $D_2 = \{d_{t+1}, ..., d_n\}$. In D_1, the corresponding images are similar to the first image I_1. Since I_1 is not the key frame, these images are all transition frames which can be removed from the video. On the other hand, images in D_2 are distinct from I_1 which means the first key frame must exist among them. Then how to find the exact frame from D_2 ?

(5) Let's take the sign language "support" as an example. Arrange the difference values of D_2 in time order. The abscissa represents image sequence in chronological order, and the ordinate represents d_i $(t+1 \le i \le n)$ which means the difference between an image and the reference image. Thus we can get a curve showed in Figure 4(a) in which the d_j corresponds to the earliest occurred in D_2, d_x is a local maxima value and d_{t+1} is the smallest difference. In Figure 4(b), the three images corresponding with the d_j, d_x and d_{t+1} are showed, their serial numbers are No. 33, No. 35 and No. 62 respectively. We have learned that d_{t+1} is the theoretical value calculated from formula (1). Although the corresponding image is dissimilar to I_1, it is not the key frame actually. d_j is the closest to I_1 in time, so it may be the key frame considering the continuity of time. But in fact, the corresponding image often has not reached the exact gesture while the image corresponding d_x is just the perfect key frame.

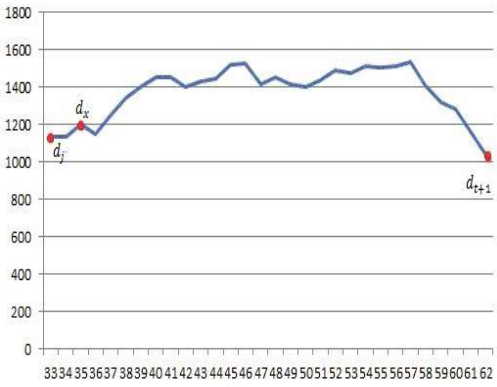

Figure 4(a) Curve of Difference in Time Series

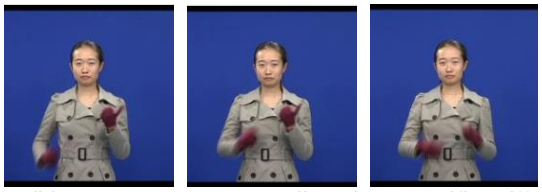

Figure 4(b) Images Corresponding the Specific Difference

Figure 4. Key Frame Selection Example 1

(6) Assuming that No. 35 is selected as the first key frame, the images before it in time can all be removed. Then this first key frame is looked on as the new reference and we can get a new sequence of difference values if all the remaining images are subtracted with the first key frame. Then after repeating step (3), (4) and (5), No. 38 is the second key frame as Figure 5 shows us. Figure 6 is the third key frame and so on. The process continues until the number of rest images are less than a certain value. A great number of experiments show that the method with local maxima is optimal.

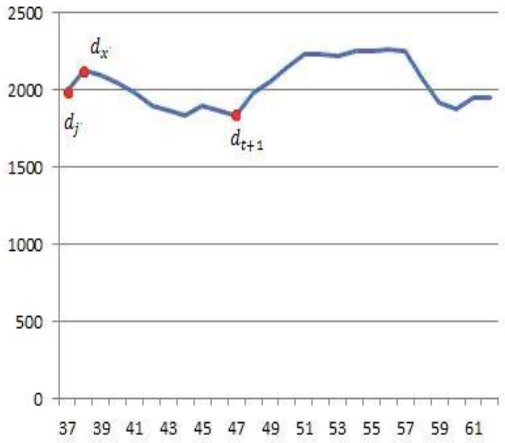

Figure 5(a) Curve of Difference in Time Series

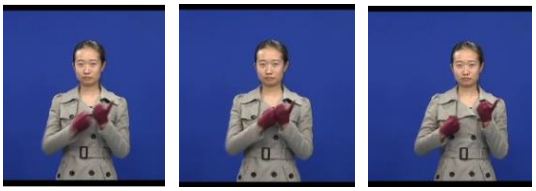

Figure 5(b) Images Corresponding the Specific Difference

Figure 5. Key Frame Selection Example 2

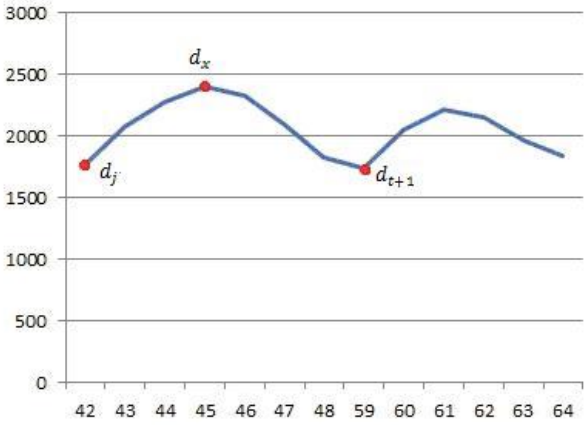

Figure 6(a) Curve of Difference in Time Series

Figure 6(b) Images Corresponding the Specific Difference

Figure 6. Key Frame Selection Example 3

4. Weighted Key Frame

In the experiment, a new problem occurs. The key frames of "support" shown in Figure 7 have two adjacent and similar images, the first one and the second one. Although the two gestures are almost the same, they locate in different positions. This is the reason why these two are both selected out as the key frames. The difference between the two is big enough to be dissimilar when they are classified through image difference. It means one of them is useless or unimportant when the other key frames are indispensable. Such condition often happens in some other signs because of the instability of people's sign gestures. Then the point is how to identify the importance of the key frames. An effective approach is to give weights to the key frames. Higher weights are placed on necessary key frames and lower on unnecessary ones. For instance, in Figure 7, one of the first two is unimportant frame and the rest frames are all important.

Figure 7. Key Frames of "Support"

Suppose the current key frame is I, and the corresponding reference is I_0, then the gesture difference between I and I_0 is defined as Δd:

$$\Delta d = \mu_1 d_x + \mu_2 d_y \quad (5)$$

The d_x is just the difference between I and I_0, and the d_y is the difference of pixels number in gesture area between I and I_0. Therefore the d_x and d_y respectively represent the gesture differences in position and shape. The two coefficients μ_1 and μ_2 represent the different importance of d_x and d_y for the gesture recognition. Thus the value of Δd can describe the difference between I and I_0 in both position and shape. This formula is trying to find out the continuous similar key frames whose hand shape is almost the same but the location changes a lot, which is different from formula (1). It doesn't care about repetition but only the change both in location and shape, which is more suitable to classify. And if we replace the formula (1) with (5), it will cause key frame missing. Many experiments in this area lead us to conclude that the unimportant key frames are often the repetitions of the important key frames. In other words, d_y of unnecessary key frames is always of a very small value. Our experiment supports our assumption that Δd of the unnecessary key frames are usually smaller, as Table 1 shows. Since the value of the second key frame is much smaller than others, the normalized Δd can be used as the weights of key frames.

Suppose S is a set of key frames, i.e. $S = \{s_1, s_2, ..., s_n\}$, their Δd is $\{\Delta d_1, \Delta d_2, ..., \Delta d_n\}$, and the corresponding weights are $W = \{w_1, w_2, ..., w_n\}$.

$$w_i = \frac{\Delta d_i}{\sum_{j=1}^{n} \Delta d_j} \quad (5)$$

From Table 1 we can see that the weight of unimportant frame is very small. Then the importance of frames can be distinguished, which is good for the latter recognition.

Table 1. Weight of Key Frames (μ_1=0.68, μ_2=0.32)

Key frames sequence	1	2	3	4	5
Δd	17590.6	1990.5	4515.8	3588.8	4281.2
Weight	0.4941	0.0559	0.1268	0.1008	0.1202

4. Experiments and Result Analysis

Experiments are performed to prove the accuracy and stability of our method. The training set consists of 80 common signs performed by a girl and a boy who are chosen at random. They are not sign language professionals, only through on-site learning to do sign language movements, *i.e.*, so-called non-specific populations. The sign videos are shot under 25 frames per second and 50 frames per second respectively and each sign is repeated three times. Then we have a total of 960 training samples. Based on the adaptive method mentioned above, these samples are used to extract key frames and the results are shown in Figure 8 and Figure 9.

Figure 8 (a) girl, sampling frequency is 25/sec

Figure 8 (b) boy, sampling frequency is 25/sec

Figure 8(c) girl, sampling frequency is 50/sec

Figure 8 (d) boy, sampling frequency is 50/sec

Figure 8. Key Frames of "Butterfly"

When people demonstrate sign languages, they must raise their arms before signing and put down them thereafter. These two parts convey nothing but they do something to our key frames detection, as shown in Figure 8. The first and the last frames are often such kind of meaningless gestures and this never happens in other frames. According to many experimentations and data comparisons we find that the center-of-gravity position of beginning and ending actions is lower and the sign languages are always begin at a certain height or above. So this problem can be solved by checking the position of the center-of-gravity. After the key frame extraction we can conduct an extra step to see if the first two frames are the right ones. As long as they are lower enough, they are regarded as a useless frame and will be removed. After the improvement, the results are shown in Figure 9.

Figure 9(a) girl, sampling frequency is 25/sec

Figure 9 (b) boy, sampling frequency is 25/sec

Figure 9 (c) girl, sampling frequency is 50/sec

Figure 9 (d) boy, sampling frequency is 50/sec

Figure 9. Key Frames of "Work"

The results in Figure 9 are more accurate. The experiments achieve our aim. The adaptive method proposed in the paper can extract key frames accurately of any video, and also it has good robustness to different camera, different rate and different person. What's more, after weighing all key frames, those unnecessary frames won't affect the final recognition any more.

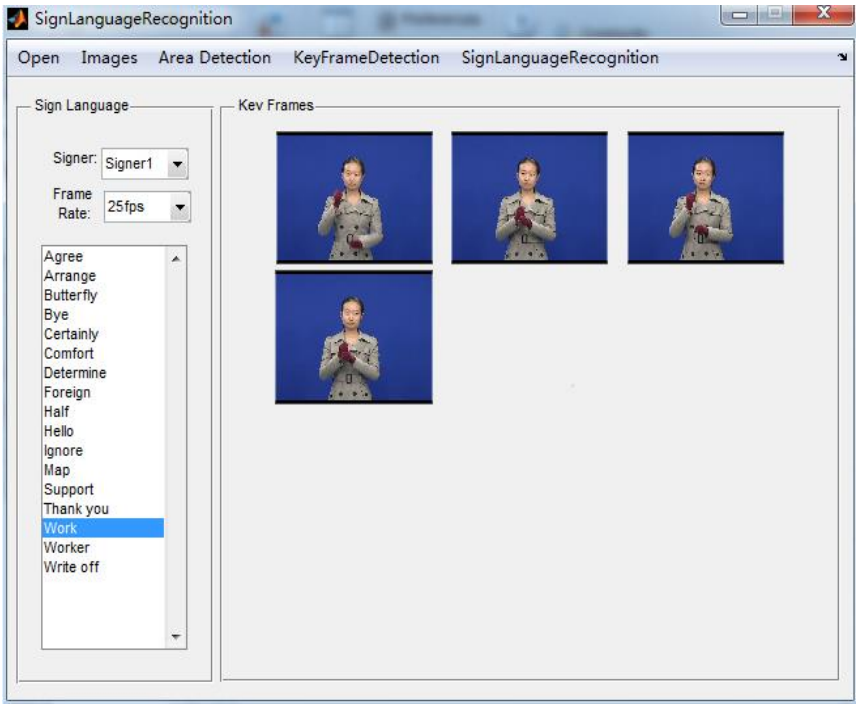

Figure 10. The System Interface

The system of key frame detection is achieved through matlab, and the interface is shown in Figure 10. This software can load original sign language video, recognize gesture area, and detect key frames. Besides, the sign language recognition function is under programming. Meanwhile, the comparative experiment is carried out using unsupervised clustering [20] with the same sample set. The result is shown in Table 2. The clustering algorithm needs a great number of computations, it has lower efficiency. Based on it, more key frames are extracted, most of which are meaningless transitional gestures, and some important key frames may be missed sometimes. The experimental results prove that our method is of higher computation efficiency, higher accuracy and

better robustness.

Table 2. The Comparison Result

		Our method	Unsupervised clustering
Rate:25 fps	Average number of key frames	4	17
	Missing detection	0	5.8%
	Average running time(s)	0.0656	1.242
Rate:50 fps	Average number of key frames	6	29
	Missing detection	0	3.4%
	Average running time(s)	0.4634	10.525

5. Conclusion

To solve the problem in the field of current sign language recognition, a new idea is proposed in this paper, *i.e.*, the dynamic sign language can be described by a sequence of key frames and then recognized by these key frames. This method focuses on the key frame extraction and can minimize the limitation to the users and the requests of equipment, which makes the interaction between people and computer more natural and realizes the comprehensive application of sign language recognition. Our future work will focus on improving algorithm and feature extraction and matching algorithm to achieve the real-time recognition system.

Acknowledgements

Project supported by national natural science foundation of china 61100108.

References

[1] The Department for Education and Employment of China Disabled Persons' Federation, China Association if the Deaf and Hard of Hearing. Chinese Sign Language, Huaxia Publishing House, Bei Jing, **(2003)**.
[2] Z. Yafei, "Research on Vision Based Hand Gesture Recognition Technology", Zhejiang University, Hang Zhou, **(2011)**.
[3] G. Cailong, "Research on Chinese Static Sign Language Recognition", Xi'an University of Architecture and Technology, Xi An, **(2009)**.
[4] A. R. Sarkar, G. Sanyal and S. Majumder, "Hand Gesture Recognition Systems: A Survey", International Journal of Computer Applications, vol. 71, no. 15, **(2013)**, pp. 26-37.
[5] A. Samantaray, S. K. Nayak and A. K. Mishra, "Hand Gesture Recognition using Computer Vision, International Journal of Scientific & Engineering Research, vol. 4, no.6, **(2013)**, pp.1602-1608.
[6] M. Maraqa, F. Al-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, no. 41, **(2012)**, pp. 41-52.

[7] K. Murakami and H. Taguchi, "Gesture Recognition using Recurrent Neural Networks", ACM Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, New Orleans, **(1991)** April 12-16.
[8] E. Stergiopoulou and N. Papamarkos, "Hand gesture recognition using a neural network shape fitting technique", Elsevier Engineering Applications of Artificial Intelligence, vol. 22, no. 8, **(2009)**, pp. 1141–1158.
[9] Ibraheem, Noor A., and R. Z. Khan, "Vision based gesture recognition using neural networks approaches: A review", International Journal of human Computer Interaction, vol. 3, no. 1, **(2012)**, pp. 1-14.
[10] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", IEEE Transaction on Pattern Analysis and Machine Intelligence, Washington D.C., **(1998)** April 3-6.
[11] K. Grobel and M. Assan, "Isolated Sign Language Recognition using Hidden Markov Models", Proc of the IEEE Int'l Conf of the system, Man and Cybernetics, Orlando, **(1997)** October 12-15.
[12] C. Vogler and D. Metaxas, "ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis", Proc of the IEEE Int l Conf on Computer Vision, India Bombay, **(1998)** January 4-7.
[13] X. Wang, M. Xia, H. Cai, Y. Gao and C. Cattani, "Hidden-Markov-models-based dynamic hand gesture recognition", Mathematical Problems in Engineering, vol. 3, no. 12, **(2012)**, pp. 1-11.
[14] X. Shen, G. Hua, L. Williams and Y. Wu, "Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields", Image and Vision Computing, vol. 30, no. 3, **(2012)**, pp. 227-235.
[15] I. S. Lim and D. Thalmann, "Key-poseture extraction out of human motion data by curve simplification", Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Istanbul, **(2001)** October 25-28.
[16] S. Junxing, S. Shouqian and P. Yunhe, "Key-frame extraction from motion capture data", Journal of Computer-Aided Design & Computer Graphics, vol. 16, no. 5, **(2004)**, pp. 719-723.
[17] H. Togawa and M. 0kuda, "Position-based keyframe selection for human motion animation", Proceedings of the 11th International Conference on Parallel and Distributed Systems-Workshops, Fukuoka, **(2005)** July 20-22.
[18] S. Y. Li, M. Okuda and S. I. Takahashi, "Embedded key-frame extraction for CG animation by frame decimation", Proceedjngs of IEEE Internatjonal Conference on Multi-media & Expo, Amsterdam, **(2005)** October 15-17.
[19] Y. Liu and J. Liu, "Keyframe extraction from motion capture data by optimal reconstruction error", Journal of Computer-Aided Design & Computer Graphics, vol. 22, no. 3, **(2010)**, pp. 670-675.
[20] Y. Zhuang, Y. Rui, S. Thomas, H. and S. Mehrotra, "Adaptive Key Frame Extraction Using Unsupervised Clustering", Proc. Of IEEE Int. Conf. on Image Processing, Nagasaka, **(1998)** October 12-14.
[21] N. Ejaz, T. B. Tariq and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism", J. Vis. Commun., Image R, vol. 23, **(2012)**, pp. 1031–1040.

Authors

Li Shurong, Female, 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 (Contact author), Female, Han nationality, born in 1975, associate professor, now working at the Nanjing University of Aeronautics & Astronautics, college of computer science and technology, and the main research direction is recognition and image processing.

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

Dai Qun, Female, Han nationality, born in 1975, associate professor, now working at the Nanjing University of Aeronautics & Astronautics, college of computer science and technology, and the main research direction is data mining and machine learning.

