A Simple and Efficient Key Frame Recognition Algorithm for Sign Language Video

Zhaosong Zhu, ShengWei Zhang, and YunLei Zhou(✉)

Nanjing Normal University of Special Education, Nanjing 210038, China

Abstract. Sign language is an important means of social communication for hearing-impaired people, and most developed countries have established their own hand language banks. Under the guidance of the National Language Commission, China has created a national sign language corpus, which is mainly composed of video. For the database, one of the most important work is to establish the index of retrieval. For sign language videos, the most important index is the hand shape displayed in the video key frame. In this paper, a simple and efficient key frame extraction algorithm is proposed based on the video library with good consistency, namely the sign language video library, to create a fast and efficient index. At the same time, it can be used as a reference for similar video libraries.

Keywords: Classification of videos · Classification of sign language · Key frame extraction · Chinese Sign Language

1 Introduction

It is an important work in the development of national language to establish a national corpus of sign language and carry out the standardization of sign language. China created its national sign language corpus in 2016. It now contains more than 60,000 sign language videos. It can be searched through the website of Nanjing Normal University for Special Education. Current retrieval methods mainly rely on Pinyin and Chinese strokes, as well as hand-shape image indexing. The indexing mainly depends on manual division, which has the disadvantages of high cost, low efficiency, high error and so on. Moreover, from the linguistic point of view, the generation of sign language video needs to rely on its own information, and cannot be disturbed by normal natural language. Therefore, the use of computers from the perspective of video itself, from the perspective of linguistics, is a very important means.

The key frame recognition technology of sign language based on graphics is a challenging subject. In a video of sign language movements, not every frame has a semantic effect on the expression. As a person who uses sign language, he will take the initiative to emphasize his sign language semantics. So there is a pause in sign language presentation, and this pause is what he emphasizes. The relatively still image produced in this pause process can be used as the key frame of sign language video. Graphics obtained

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 388, pp. 3–10, 2021.
https://doi.org/10.1007/978-3-030-82565-2_1

from this key frame can be directly used as an index, or classified to generate a secondary index.

Video key frame extraction methods are generally divided into four categories. The first category is to extract key frames according to the content of the image. The content of the video is the embodiment of the image features, and the corresponding content variation degree is the standard for selecting key frames. The second category is the motion analysis in the image. According to the optical flow field of the image, the optical flow diagram is calculated and the minimum frame is taken as the key frame. The third type is the key frame detection based on the trajectory curve density feature. The density of the trajectory density curve is used to distinguish the key frame and the non-key frame. The fourth class is a popular clustering method, which needs to set the number of clusters in advance, and then group the similar pins into one category, with each category being a key frame.

The second type of algorithm requires a large amount of computation and takes a long time, while the third type of algorithm will produce a large deviation trajectory due to inaccurate positioning. The fourth type of algorithm needs to establish clustering, which has a lot of redundancy and a high computational load. Their advantages and disadvantages are shown in the Table 1.

Table 1. The advantages and disadvantages of the four algorithms

Common key frame extraction algorithms	Advantages	Disadvantages
Based on image content [1, 2]	The feature of the bottom layer of the image can be fused	The image depth feature cannot be utilized
Based on optical flow field [3, 4]	Better expression of global movement	The image depth feature cannot be utilized
Based on curve density [5]	It can better reflect the motion trajectory	Inaccurate positioning, large deviation
Based on cluster analysis [6, 7]	Popular big data processing methods	Not being able to determine the number of clusters is prone to redundancy

The video content in the video library of sign language has good consistency, including similar background, similar light and clothes with less color difference. Therefore, it is not necessary to use a high amount of computation in the process of extracting key frames. It is only necessary to extract a small number of key frames based on content and carry out screening.

2 Key Frame Selection Process Design

This algorithm is mainly divided into the following steps: Step 1, serialization of images. Step 2, grayscale of the image. Step 3, removal of the image background, Refer to the

background image provided. Step 4. difference processing with adjacent frames. Step 5, find out the minimum frame between two maximum frames, which is the key frame. Step 6 is optional. According to the actual demand, find the best frame in the neighborhood of the smallest frame that meets the actual demand.

2.1 Serialization of Images

Complete conversion of images to frames without redundant processing. This is a relatively large amount of data, but it reduces the computation. As shown in the figure, at 25 frames per second, the three-second video is split into 75 frames, while the one-minute video has about 1,500 frames. In order to keep the video consistent and not make redundant cuts. A sequence (I_1-I_n) of images is formed shown in Fig. 1.

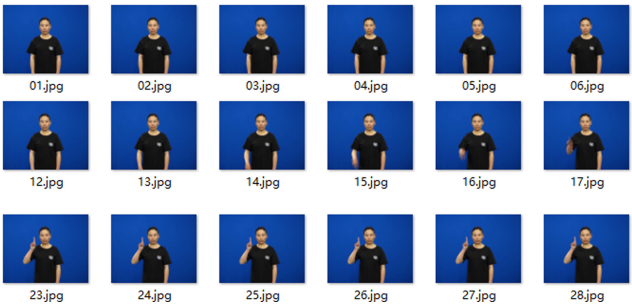

Fig. 1. Serialized picture of sign language video

2.2 Grayscale of the Image

According to the standard graying formula (1) proposed in literature [8], the color space of each image I_i is reduced from three dimensions to one dimension gray-scale image G_i, and the gray order is normalized to a range between 0 and 255. See Fig. 2 for the gray scale image.

$$gray = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{1}$$

2.3 Removal of the Image Background

Select the standardized reference image. Under normal circumstances, the first frame of image G_{first} can be used as the reference image. However, considering that the sign language presenter will have slight deviation in the process of expression, the last frame of image G_{last} also needs to be referred. The average value of the two can be taken, $G_{ref} = (G_{first} + G_{last})/2$, to generate the standard as shown in the Fig. 3(a).

Please note that if the video is longer, scenes change under the condition of larger, G_{first} and G_{last} can be replaced with $KEY_{previous}$ and KEY_{next}, This is easy to understand, the video sequence is contiguous, so the reference image of the current key frame can be selected from the previous and the next key frame. This is done only if the previous key frame and the next key frame have been found.

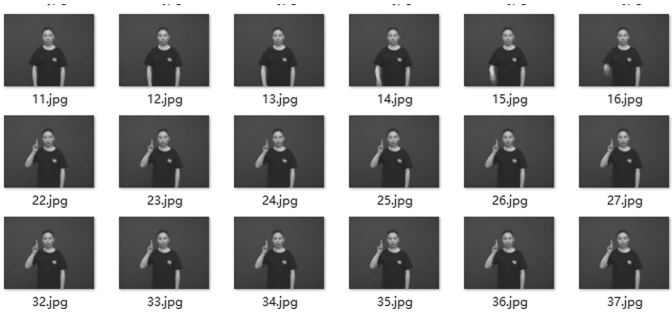

Fig. 2. A sequence of images after graying

Fig. 3. a $G_{first} + G_{last}$. b. $R_i = G_i - G_{i-1} - G_{ref}$

2.4 Difference Processing of Frames

The difference processing can be carried out with the adjacent frames, and with the before and after frames, $R_i = G_i - G_{i-1} - G_{ref}$, and the obtained part is the part formed by the gesture in the motion.

Because of the difference between the two values, there are fewer pixels, and the negative pixels are automatically set to 0, so it is more difficult to see as shown in Fig. 3(b).

2.5 Find Out the Minimum Frame Between Two Maximum Frames

All the gray-scale values of the pixels are summed up, $SUM\_GRAY_i = SUM(R_i)$, draw the line chart (Fig. 4), find the extreme value point. The frame where the extreme point is located is the time sequence region that varies greatly in the video. The middle point of the extreme point may be the image emphasized by the gesture pause, which is the key frame that the video is looking for. As shown in the figure, the extreme points are frame 17 (Fig. 5(a), value: 118745) and frame 44 (Fig. 5(b), value: 127596). This indicates that these two frames have the greatest changes, and the 30th frame (Fig. 5(c)) between them has a minimum value of 71, which is optional as a key frame.

2.6 Find the Best Frame in the Neighborhood of the Smallest Frame (Optional)

In some cases, minimum frames may not be sufficient for key frame processing. For example, it does not have clear edges and corners, which is not convenient for subsequent

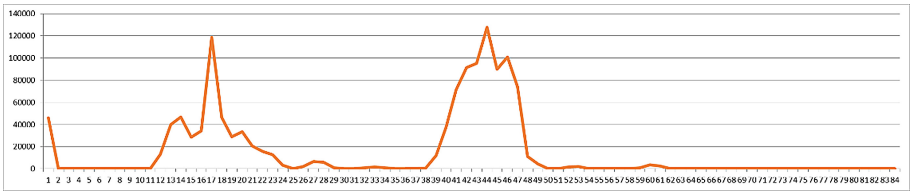

Fig. 4. A line diagram of a grayscale summation sequence

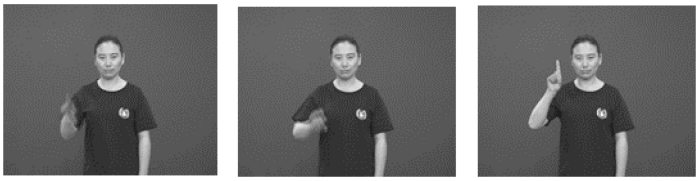

Fig. 5. a Frame 17. b. frame 44. c. frame30

image feature collection. Therefore, it is necessary to search in the area near the minimum frame. This requires defining a search area that is the neighborhood of the minima frame. The neighborhood distance formula (2) is as follows:

$$d = a^{*}diff_{NO.} + b^{*}diff_{vlaue} \tag{2}$$

Where, parameters a and b are weights, $diff_{NO.}$ is the difference of sequence number, and $diff_{value}$ is the difference of values between two frames. The ratio of a and b can be adjusted to meet actual needs. This process results in a region, as shown in Fig. 6.

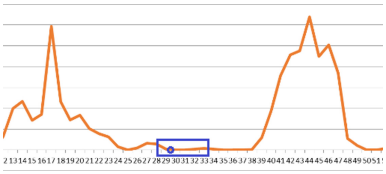

Fig. 6. The neighborhood of the minimum frame

3 Extract Key Frames from Long Videos for Verification

The selected video is an alphabet video, showing the 26 letters of English and the three initials unique to Chinese. It lasts for 1 min and 16 s, 25 frames per second, and produces a total of 1900 frames of images.

Due to the large amount of data, three segments of data (Fig. 7, Fig. 8, Fig. 9) were randomly selected to verify whether the key frame was selected properly:

In Fig. 7, the maximum frame is Frame 26 and Frame 69, the minimum frame is Frame 47, and the neighborhood is 41–53.

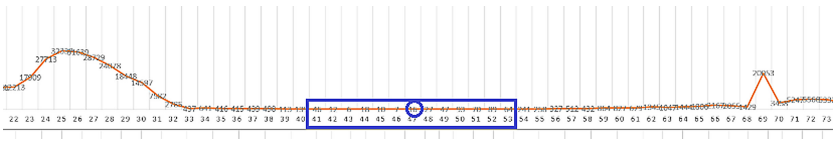

Fig. 7. The first data sequence.

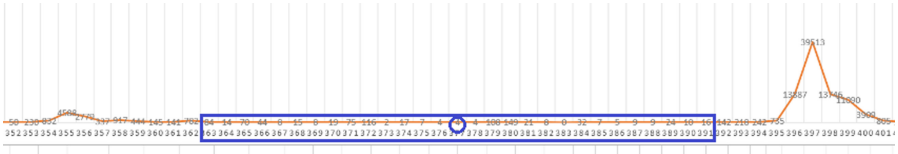

Fig. 8. The second data sequence

In Fig. 8, the maximum frame is 355 and 397, the minimum frame is 377, and the neighborhood is 363–391.

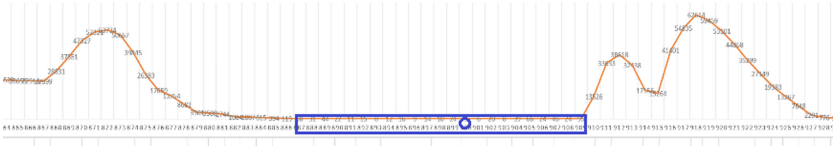

Fig. 9. The third data sequence

In Fig. 9, the maximum frame is frame 872 and frame 919, the minimum frame is frame 893, and the neighborhood is frame 887–909.

Fig. 10. The key frame found from three sequence data

The minima is found between the two maxima, resulting in three sequence images, as shown in Fig. 10.

From the results obtained, the edges of the image frames obtained by this algorithm are clear and the features are obvious, which can meet the requirements of some algorithms, such as Canny [9, 10] and SURF [11] algorithms. The key frame acquisition of sign language video is basically realized.

4 Conclusion

In this paper, a simple recognition method belonging to the key frame is essentially the application of the first derivative. When the difference between a frame sequence and its front pin is large, the frame is considered to be in motion change. And the minimum value between the two moving frames, that's the stressed and paused frame that the sign language shows. These frames play a key role in video classification retrieval and sign language recognition.

In addition, the algorithm that is not necessarily complex will have higher efficiency. If the curve density method and cluster analysis method are adopted, a large amount of data and operations will be generated, which is not a small cost for the whole sign language video library.

Therefore, in the practical application, we should analyze the specific situation and adopt the appropriate method, just like the difference method used in this paper to find the key frame for the image with good consistency, which is effective.

Acknowledgement. This work was supported by The Ministry of Education has approved a key project in the 13th Five-Year Plan for Education Science in 2017: "Research on Higher Education Teaching Support for the Disabled in the Context of Big Data". (No. DIA170367), The Major Programs of Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 19KJA310002.) and The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 17KJD520006).

References

1. Cao, J., et al.: A key frame selection algorithm based on sliding window and image features. In: 2016 International Conference on Parallel and Distributed Systems (ICPADS), pp. 956–962. IEEE, Wuhan, China (2016)
2. Chen, L., Wang, Y.: Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features. Autom. Constr. **81**, 355–368 (2017)
3. Ioannidis, A., Chasanis, V., Likas, A.: Weighted multiview key-frame extraction. Pattern Recogn. Lett. **72**, 52–61 (2016)
4. Devanne, M., et al.: 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold. IEEE Trans. Cybern. **45**(7), 1340–1352 (2015)
5. Guo, X.P., Huang, Y.Y., Hu, Z.J.: Research on recognition algorithm of continuous sign language statement based on Key frame. Comput. Sci. **44**(2), 188–193 (2017). (in Chinese)
6. Nasreen, A., et al.: Key frame extraction and foreground modeling using K-means clustering. In: 2015 7th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), vol. 34, pp. 141–145. IEEE, USA, (2015)
7. Gharbi, H., et al.: Key frames extraction using graph modularity clustering for efficient video summarization. In: IEEE International Conference on Acoustics, pp. 1502–1506. IEEE, USA (2017)
8. Rubner, Y.I., Tomasi, C., Guibas, L.J.: Mover's distance as a metric for image retrieval. Int. J. Comput. Vision **40**(2), 99–121 (2000)
9. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Image Understand. **18**(6), 679–698 (1986)

10. Wang, X., Liu, X., Guan, Y.: Image edge detection algorithm based on improved Canny operator. Comput. Eng. **34**(14), 196–198 (2012). (in Chinese)
11. Bay, H., Tuytelaars, T., Cool, L.V.: SURF: speeded up robust features. In: Proceedings of the 9th European Conference on Computer Vision, pp. 404–417. Springer-Verlag, Berlin, Germany (2006)