Sign Language Video Classification Based on Image Recognition of Specified Key Frames

Zhaosong Zhu[1], Xianwei Jiang[1,2], and Juxiao Zhang[1(✉)]

[1] Nanjing Normal University of Special Education, Nanjing 210038, China
zzs2019@foxmail.com, 330161l@qq.com
[2] Department of Informatics, University of Leicester, Leicester LE1 7RH, UK

Abstract. This paper is based on the Chinese sign language video library, and discusses the algorithm design of video classification based on handshape recognition of key frames in video. Video classification in sign language video library is an important part of sign language arrangement and is also the premise of video feature retrieval. At present, sign language video's handshape classification work is done manually. The accuracy and correctness of the results are quite erroneous and erroneous. In this paper, from the angle of computer image analysis, the definition and extraction of key frames are carried out, and then the region of interest is identified. Finally, an improved SURF algorithm is used to match the area of interest and the existing hand image, and the classification of the video is completed. The entire process is based on the actual development environment, and it can be used for reference based on the classification of video image features.

Keywords: Classification of videos · Classification of sign language · Key frame extraction · Image matching · Handshape matching

1 Introduction

With the development of information technology, Internet technology and multimedia technology have been greatly improved, In particular, the emergence of "we media", such as YouTube, Facebook, tiktok and so on, has led to the explosive growth of videos on the Internet, in this case, the manual annotation has become impossible, and the subjectivity of artificial tagging cannot meet the needs of users. In order to facilitate the management and retrieval of massive video, automatic video classification is particularly important, Automatic video classification is also widely used in video monitoring, network supervision, medicine and other fields: For example, Johnson et al. proposed a multi-mode monitoring method, which extracts the static features of human body from multiple angles to realize the detection, separation and recognition of human beings at a distance [1]; Through video classification, video on the Internet is regulated to filter out undesirable videos (pornography, violence, etc.) [2]. The video classification method was applied to the video library obtained by the wireless endoscope, and all the videos was classified according to the different organs diagnosed [3]. In addition to specific fields, there are also some general video classification methods: Fischer et al. proposed in 1995 that video could be divided into news, sports, business, advertising,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2020
Published by Springer Nature Switzerland AG 2020. All Rights Reserved
Y.-D. Zhang et al. (Eds.): ICMTEL 2020, LNICST 327, pp. 371–381, 2020.
https://doi.org/10.1007/978-3-030-51103-6_33

cartoons, etc. [4]; Huang et al. proposed a classification algorithm based on text features, extracted user-generated text features, and used a classifier for classification [5]. Jiang et al. proposed a method of video classification using support vector machine (SVM) based on visual features (color, motion, edge, etc.) [6]; Subashini et al. proposed a machine learning algorithm based on audio features and image histograms [7].

However, the current video classification methods generally have two problems:

1) Insufficient universality. Some prior knowledge is needed to design the classification rules, which can only be targeted at certain fields;
2) Complex algorithms. It needs a lot of computing resources to use multi-level deep learning algorithm to deal with video library with a large amount of video.

In order to solve the above problems, this paper proposes a classification method for key frame images of sign language video. The main steps are as follows:

1) Extraction of key frames from sign language video;
2) Image visual feature preprocessing and hotspot extraction;
3) Feature matching with the designated image to achieve video classification.

2 Datasets and Problems

The object of this paper is video library of Chinese sign language (csl-lib, Project NO. zda125-8, 2016). The library contains 57,531 sign language vocabulary videos from nine specific regions in China. At present, the copyright of this dataset belongs to the Chinese language and script commission, and some contents will be released later. In the retrieval operation of video library, the handshape index of video is an important retrieval method, which is also the only retrieval method based on video image features, and has important video analysis and research value. The classification is mainly based on the 60 hand shapes in sign language (see Fig. 1). It will consume a lot of human, material and time resources by manual classification. In view of this problem, a set of practical classification methods is proposed from the perspective of computer image processing, Furthermore, SURF algorithm [9] which is based on Lowe D G's SIFT algorithm [8], is improved by plane angle rotation. Meanwhile, the key frame extraction algorithm of literature [10, 11] is applied. Finally, a classification method based on key frame matching is proposed which can be applied to sign language handshape classification and has the characteristics of batch and high efficiency.

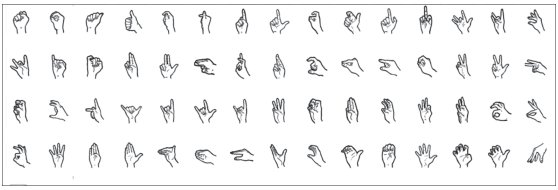

Fig. 1. The handshape index for Chinese sign language

3 Extraction of Video Key Frames

The key frame of video retrieval is defined as the image of the handshape used by the gesture in the sequence formed by the video stream. Take the sign language video 'lightning/electricity' as an example. This sign language is the standard sign language. In most parts of mainland China, it has the same or similar stroke (the left hand does not move, and the right hand draws the shape of lightning in the air), so it is typical of cases.

Key frame extraction is mainly divided into two steps: Firstly, video serialization and graying. The grayscale processing adopts the general formula (1) proposed in literature [12] to form the video sequence as shown in Fig. 2. Secondly, extract key frames according to the algorithm.

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

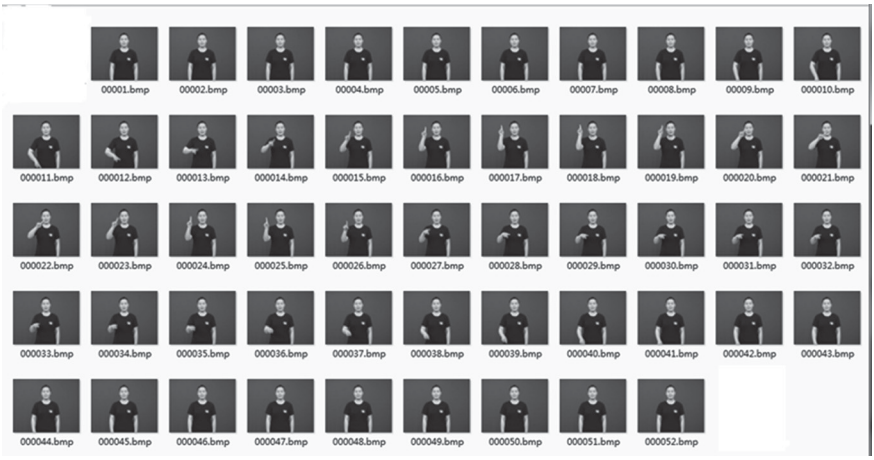

Fig. 2. The grayscale sequence of sign language video 'lightning/electricity'

The extracted key frame image shoulsd have two characteristics: 1) it should have a certain duration of stability; 2) clear edges can be preprocessed into hand recognition materials.

According to the two characteristics, the key frames are described as follows during the processing: The current frame has a small difference with the preceding m frames and the following m frames, The value of m is a natural number, so it should not exceed the number of frames, In the method of this paper, the efficiency of video library processing is considered. So m is equal to 2, that is, the difference is calculated with the first two frames and the second two frames. The difference coefficient is calculated by the gaussian function, and the difference is summarized at last.

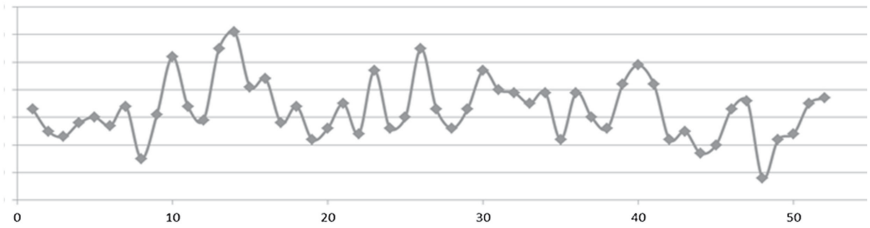

Fig. 3. The mean grayscale image sequence of sign language 'lightning/electricity', the X-axis is the video image sequence, and the Y-axis shows the change of grayscale mean

Assuming that the number of frames in the video is n, and the key frame is i, V_i is defined as the grayscale matrix of the sequence image i, The vector (i, V_i) forms a discrete function: $V_i = f(i)$, Fig. 3 shows the mean value function of gray square of "lightning" sign language video sequence. Key frame image extraction is completed according to the following formula:

$$V_{key} = f\left(\int Min_p(|f'(x)|)\right) \tag{2}$$

But it's hard to compute the derivative of the function formed by the discrete sequence directly, So $f'(x)$ is calculated using the series of coefficient operators S. S is normalized by the gaussian function formula (3), (4), (5) (A = 1, σ = 1.5)

$$G(x) = Ae^{-\frac{x^2}{2\sigma^2}} \tag{3}$$

$$S = |-0.135, -0.365, 1, -0.365, -0.135| \tag{4}$$

$$f'(x) = |\mathrm{V}_{i-2}, \mathrm{V}_{i-1}, \mathrm{V}_i, \mathrm{V}_{i+1}, \mathrm{V}_{i+2}| \times S^T \tag{5}$$

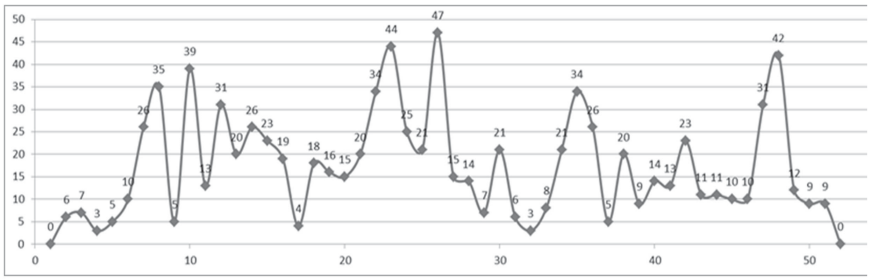

Fig. 4. The change rate of the overall grayscale pixel in the image sequence

Generate $|f'(x)|$ data sequence (Fig. 4) to reflect the changes between images. According to the actual situation of the video, the head subsequence and the tail

subsequence are removed, because these two sub-sequences may contain useless information. In this project, the front part is about 10 frames for the video model gesture preparation stage, and the back part is about 10 frames for her gesture homing stage, which needs to be removed. Then find the p frames in the sequence where the rate of change is from small to large, $Min_p(|f'(x)|)$, When $p = 3$, three Ordinal Numbers conforming to the conditions are obtained, which are 17, 32 and 37 respectively. The corresponding images are frame 17, 32 and 37 (Fig. 5).

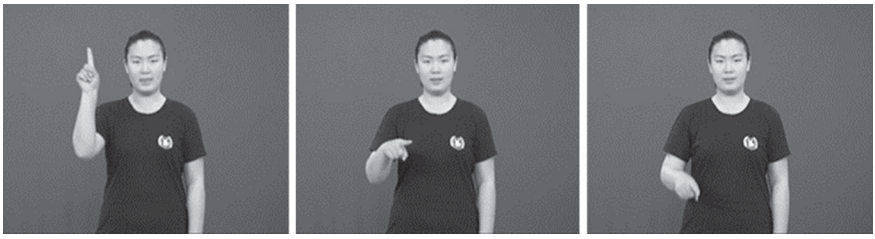

Fig. 5. The three images with the lowest rate of change in the video sequence

The video key frame is extracted, and then the hand shape matching recognition is carried out.

4 Image Visual Feature Preprocessing and Thermal Region Extraction

The interference of hand image matching mainly comes from the following two points: 1) Differences between the matched images. That is, the source of the key frame image and the matching handshape image are different, and the difference of the image feature is relatively large, so it is not easy to form a match. 2) Spatial difference. Although it is the same handshape, the spatial position transformation gap is large and it is not easy to match.

For the second point, there is no valid three-dimensional transpose algorithm for plane space, Therefore, in order to reduce the first kind of interference and form an effective matching feature, the key frame image is preprocessed before matching to improve the matching accuracy. Preprocessing is divided into two steps: 1) Image marginalization and binarization to reduce pixel interference; 2) image hotspot extraction to reduce the interference of non-hot region.

4.1 Image Marginalization and Binarization

High pass filter is used to process the image. The improved regional edge detection algorithm proposed by J Canny [13] and wang [14] et al. was referred to and simplified to meet the needs of this project. The kernel matrix of Sobel operator is used as filter (Fig. 6).

-1	0	1		-1	-2	-1
-2	0	2		0	0	0
-1	0	1		1	2	1

Fig. 6. Sobel operator for edge detection

The horizontal filtering and vertical filtering were performed respectively, and after filtering, the matrices sobel_X and sobel_Y were formed, Then the L1 normal form is used to obtain filtering results.

$$L1 = |sobel_X| + |sobel_Y| \tag{6}$$

Finally, the low threshold is used for binarization processing to retain the basic contour information. In the case of this project, the gray intermediate value 128 is adopted as the threshold. The result is shown in Fig. 7. The selection of threshold value should be optimized according to the specific situation of the image.

Fig. 7. key frame images that has been marginalized

4.2 Extraction of Image Hot Area

The extraction process of hot area is as follows: Suppose the key frame image is Img_i. It's previous frame is Img_{i-1}, and next frame is Img_{i+1}, The extraction formula is as follows:

$$Img_{key} = |Img_i - Img_{i-1}| + |Img_i - Img_{i+1}| \tag{7}$$

By using the difference of adjacent images to extract the hot areas, the background interference can be minimized and only the dynamic region can be concerned, Finally, the results are filtered by low threshold, and the remaining part is the hot areas, The final results are shown in Fig. 8, and the position of hot areas is shown in Table 1.

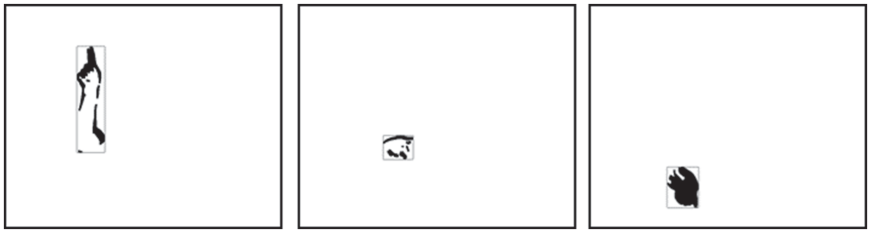

Fig. 8. Sign language video 'lightning/electricity' key frame image hot areas

Table 1. Coordinates of hot areas in three images of Fig. 8

Upper left and lower right coordinates	X_0	Y_0	X_1	Y_1
Left image	177	104	275	104
Middle image	218	337	297	337
Right image	200	419	283	419

5 Matching the Hot Areas of Key Frame with Handshape Image

In the process of matching the key frame hot areas with the feature handshape images, the scale-invariant feature matching is needed. The accelerated version of SIFT algorithm [8], SURF algorithm [9], was used and the results were further filtered, SURF algorithm has the following 6 steps:

(1) Construct the Hessian matrix of the image through formula (8) and calculate the eigenvalue. The convolution window of gaussian filter is used in the calculation process, and the simplified matrix of 3 * 3 is adopted $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$.

(2) Construct gaussian difference scale space of the image. It is generated by convolution of gaussian difference kernel with image at different scales. Core formula (8), (9) and (10) are as follows:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{8}$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{9}$$

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma))I(x, y) = L(x, y, k\sigma)L(x, y, \sigma) \tag{10}$$

Since the image size is small, the space is set as 3 scales. Each scale contains five levels. The template size of the filter increases gradually from 1 to 3 at different scales. In 1 scale, the filter's fuzzy coefficient increases gradually, forming a 3 * 5 scale space.

(3) Feature point location. Compare each pixel processed by Hessian matrix with 26 points in the neighborhood of 2d image space and scale space. The key points were preliminarily located, and then the weak key points and the wrong key points were filtered out to screen out the final stable characteristic points.
(4) Main direction distribution of feature points. The harr wavelet transform in the circular neighborhood of the statistical feature points is used. The sum of the horizontal and vertical Haar-like features of all points in the 60-degree sector was calculated. After the rotation was conducted at an interval of 0.2 radians and the harr small baud value in the region was calculated again, the direction of the sector with the largest value was finally taken as the main direction of the feature point.
(5) generate feature point descriptors. Take a rectangle of 4 * 4 around the feature point. The direction of the rectangle region is the main direction along the feature point. Each subregion counts the horizontal and vertical Haar-like features of 25 pixels. The Haar-like features contains four sum operations, that is, the sum of the horizontal value, the sum of the vertical value, the sum of the horizontal absolute value, and the sum of the vertical absolute value.
(6) Feature point matching The Euclidean distance between two feature points is calculated to determine the matching degree. The shorter the Euclidean distance is, the better the matching degree of the two feature points is. At the same time, the Hessian matrix trace was judged. The matrix trace of the feature points had the same sign, which represented the contrast change in the same direction. If different, it was the opposite, even if the Euclidean distance was 0, it was directly excluded.

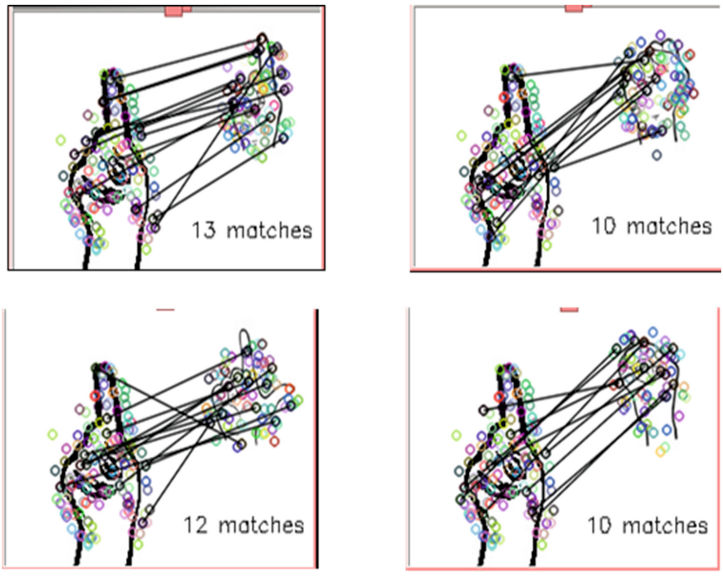

Fig. 9. The matching result of the key frame hot area and handshape image

This algorithm is used to match the hot areas of the key frames with each handshape image in turn, Finally, the percentage matching value is formed, Not all of the 60 handshape feature matching results are listed, but some typical matching results are listed as shown in Fig. 9.

As can be seen from Fig. 9, although the handshape with the highest matching degree can be analyzed as the top left figure, with the matching degree of about 14.44% among the 60 hand shapes, the difference between the matching degree of other handshapes is not large enough to form matching results, such as figure 13.33% in the bottom left figure, figure 11.11% in the top right figure, and figure 11.11% in the bottom right figure.

From the matching results, it can be seen that the matched feature point pairs need to be filtered due to more interference. Inspired by the geometric characteristics proposed by Liu et al. [15]. The filtering method is as follows:

If the matching image is placed in the same plane, the two images will have relatively fixed positions in the plane coordinate system, as shown in Fig. 10.

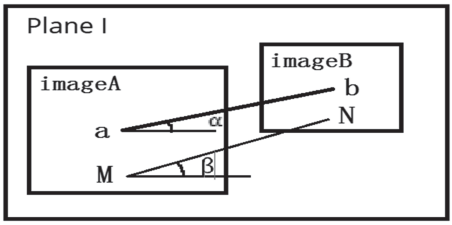

Fig. 10. Image A is the hot area of key frame, and image B is the handshape image. They are in the same plane I.

The connecting line segment of the center point of imageA and imageB is ab, The Angle between ab and the horizontal line is α. Let the Angle increment be $\Delta\alpha$, and assume that the points M and N are a pair of matching points. The Angle between the line MN and the horizontal line is β, If β satisfies formula (11), then MN is the matching point, otherwise it is deleted from the matching point pairs. You can make some adjustments by setting the size of the $\Delta\alpha$.

$$\beta\varepsilon\,[\alpha - \Delta\alpha,\ \alpha + \Delta\alpha] \tag{11}$$

The matching point pairs formed after filtering are shown in Fig. 11. The matching degrees of handshapes in the four cases were 21.43% in the top left, 17.86% in the bottom left, 5.36% in the top right, and 5.36% in the bottom right. The difference is large enough for classification, and A good result distribution was formed in 60 handshapes matches.

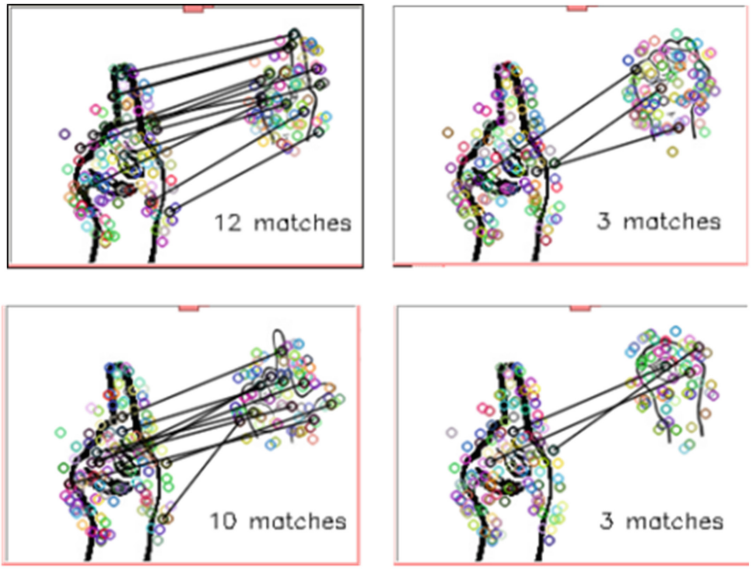

Fig. 11. The result of the Angle operation on matching point pairs

Based on the above results, we can find that the handshapes in the sign language video "lightning" have a good matching degree with the seventh and eighth hand shapes in Fig. 1.

6 Conclusion

In order to classify the videos in the sign language database by specified handshapes, this paper proposes an algorithm combination flow. Firstly, extract the key frames of digital video. The video image sequence matrix difference is used to calculate the sequence change rate and the smaller frame is taken as the key frame; Secondly, the feature extraction and hot areas extraction of the key frame images are carried out. Feature extraction uses Sobel operator to extract the contour and conduct binarization, and hot areas extraction is to use the difference between adjacent images to separate the image area with larger values. Finally, handshapes matching was performed, mainly using SURF algorithm. In addition, the point pairs generated by SURF algorithm were filtered into the plane at an Angle to form a handshape-matching distribution satisfying the requirements. Compared with the supervised learning algorithm, this process avoids the stage of sample learning and the complicated classification calculation, saves the computing resources, and has a certain efficiency and practicability. But because the video key frame image has spatial transformation, it can't match the specified handshape completely. The next research direction focuses to solve this problem with depth information [16], and at the same time, deep learning algorithm and more graphical features [17] is introduced to apply this algorithm flow to video classification and other practical applications.

Acknowledgement. This work was supported by Surface Project of Natural Science Research in Colleges and Universities of Jiangsu China (No.16KJB520029), 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. Johnson, A.Y., Bobick, A.F.: A multi-view method for gait recognition using static body parameters. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 301–311. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45344-X_44
2. Jiang, C., Jiang, X., Sun, T.: Video filtration for content security based on multimodal features. Inf. Secur. Communi. Privacy **3**, 76–77 (2012). (in Chinese)
3. Feng, W., Gao, J., Bill, P.B., et al.: Wireless capsule endoscopy video classification using an unsupervised learning approach. J. Image Graph. **16**(11), 2041–2046 (2011). (in Chinese)
4. Fischer, S., Lienhart, R., Effelsberg, W.: Automatic recognition of film genres. In: Proceedings of the 3rd ACM International Conference on Multimedia, pp. 295–304. ACM Press, New York (1995)
5. Huang, C.N., Fu, T.J., Chen, H.C.: Text-based video content classification for online video-sharing sites. J. Am. Soc. Inf. Sci. Technol. **61**(5), 891–906 (2010)
6. Jiang, X.H., Sun, T.F., Wang, S.L.: An automatic video content classification scheme based on combined visual features model with modified DAGSVM. Multimedia Tools Appl. **52** (1), 105–120 (2011)
7. Subashini, K., Palanivel, S., Ramalingam, V.: Audio-video based classification using SVM. IUP J. Sci. Technol. **7**(1), 44–53 (2011)
8. Lowe, D.G.: Distinctive image features from scaleinvariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
9. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_32
10. Zhen, E., Lin, J.: Unordered image key frame extraction based on image quality constraint. Comput. Eng. **43**(11), 210–215 (2017)
11. Wang, Y., Sun, S., Ding, X.: A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition. J. Vis. Commun. Image Representation **33** (3), 193–202 (2015)
12. Yi, R., Tomasi, C., Guibas, L.J.: Mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
13. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Image Understand. **18**(6), 679–698 (1986)
14. Wang, X., Liu, X., Guan, Y.: Image edge detection algorithm based on improved Canny operator. Comput. Eng. **34**(14), 196–198 (2012)
15. Wu, L., Yadong, C.Y.: Gesture recognition based on geometric features. Comput. Eng. Des. **35**(2), 636–640 (2014). (in Chinese)
16. Binjue, Zhang, Liaoyin, Zhao, Yixuan, Wang: Fingeritip detection and gesture recognition based on kinect depth data. IEEE Trans. Comput. Sci. Technol. **3**(1), 9–14 (2014)
17. Mitra, S., Acharya, T.: Gesture recognition: a servey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **37**(3), 311–324 (2007)