

Key Frame Extraction Based on Improved Frame Blocks Features and Second Extraction

Huayong Liu, Tao Li

Department of Computer Science, Central China Normal University

Hubei Wuhan, China

Abstract—An improved key frame extraction algorithm based on the low-level features of image is proposed in this paper. Firstly, each frame is divided into equal area of rectangular ring. Secondly, sub-block accumulative color histogram is extracted as color features and different weight is set for different rectangle rings in order to highlight the central part of frame. Thirdly, key frames are selected according to the significant change of frames. Lastly, the algorithm is optimized and key frames are selected in accordance with the frames of location in the video. The experimental results show that the proposed algorithm has good adaptability and can effectively reduce the redundant key frames when the shot has a sudden flash or the object moves fast.

Keywords- key frame; color histogram; equal-area of rectangular ring; second extraction

I. INTRODUCTION

As the advancement of network information and information technology rapidly, people mainly obtain knowledge by network. Videos are viewed as an important information media, and it gradually becomes the most important source of information resources. The amount of multimedia data is very large and the structure of video is very complex, therefore, the processing of the video is very difficult. It has become an important problem about how to fast save, scan and search these multimedia data. Key frames are defined as using some or less frames to express the theme of video; it can effectively describe the main contents of the video, but the amount of it should be little as much as possible. Consequently, it is more effective to be used to save, scan and search multimedia data than original video streams. However, on the one hand, the extracted key frames must express the theme of video more accurately and comprehensive; additional, the amount of the key frame ought to be little as much as possible, moreover, the method should be simple as much as possible.

In recent years, key frame extraction technology has attracted many domestic and foreign scholars' attentions, and some key frame extraction algorithms are proposed. At present, key frames extraction method is mainly divided into the following five categories.

A. The Method based on Video Shots

This method firstly divides video stream into several shots^[1-2], and then it selects some key frames from the stipulate seat of the shots, such as, the initial frame, the center frame and the final frame. This kind of approach includes the histogram averaging approach and the averaging frame approach. The histogram averaging approach is first to calculate the mean

values of the histogram of the frames, after that choose the frames as the key frames that closest to the mean histogram. The averaging frame approach is first to calculate the mean value at particular position from video shots, after that choose the frames as key frames whose pixel values closest to the mean value. Based on the above approaches, Zhang et al.^[3] used color histogram for extracting key frames, and it is an effective approach. Yeung et al.^[4] used the method of calculating the maximum distance in the features space for extracting key frames. The benefit of such an approach is that it has low computation complexity. However, the disadvantages of these methods are that they don't consider the variety of content about current video shots, and the amount of key frame is fixed in determined value, furthermore, the motion content cannot be efficiently described about the videos. As a result, the key frames can't appropriately represent contents of video.

B. The Method based on Motion Analysis

Wolf^[5] presented an algorithm by means of analyzing the light flow and calculating the number of the movement of the video shots, and then selected the key frames which the number of movement is least. The exercise of the shot delimiting as the following formulary (1):

$$M(k) = \sum_i \sum_j |O_x(i, j, k)| + |O_y(i, j, k)| \quad (1)$$

where $O_x(i, j, k)$ represents the x component of optical flow at pixel (i, j) in frame k and the same for the y component. This approach could extract the suitable amount of key frame from the most of video shots and the key frame also can express the video's motion effectively. However, the disadvantages of this algorithm has a weaker robustness due to it depends on the local features, and the calculation is also expensive.

C. The Method based on Video Clustering

The extracted key frames could express the theme of the video by means of clustering, the frame sequences are classified into several clusters by clustering, and then the frame is chosen as key frame at every cluster^[6-11]. This idea of algorithm is described as follow: firstly, initializing a clustering center. Secondly, determining a reference frame which is classified as the class or as a new cluster center of class by means of computing the range between cluster center and current frame. Lastly, selecting key frame which is closest the cluster center. Zhang and Hanjalic^[6] extracted key frames by means of clustering efficiency analysis of clustering segmentation algorithm; frame is set as key frame which is

closest to the cluster center. Joshi et al. [11] used the fuzzy cluster approach to a brief video sequences with only gradient video shot, for every video shots, selected the frame as key frame which is in the center. But this method needs a predetermined number of clusters before clustering, the calculation is also expensive, so it has a limit in some degree.

D. The Approach based on Visual Attention Model

In order to reduce the semantic gap between low-level feature and semantic meaning, some researchers use visual attention model to extract key frames [12-14]. Xiaolin, Peng [12] used color histogram for clustering the frame initially, after that chosen the frame as key frame which the vision is more outstanding in all clusters, Due to the use of k -means algorithm to cluster, the order sequence of key frames maybe change. Ejaz and Mehmood [14] proposed an algorithm based on an efficient visual attention model for extracting key frames. This method uses the time gradient depend on dynamic visual saliency detection, rather than the traditional method of optical flow. However, these methods' disadvantages are that the calculation is also expensive.

E. The Method based on Video Content Analysis

This method uses significant content change of the frames as a reference of selecting the key frames [15-17]. When the contents of frames changes significantly, choosing the current frame as key frame. Wu, Zhang et al. [15] adopted this approach, this approach firstly chooses the first frame as key frame, and it is regarded as a reference frame, and then calculation the distance between reference frame and back frame according to priority, the k -th frame is selected as new current key frame until the difference between the k -th and $(k-1)$ -th frame over a certain threshold. This approach could choose key frame depend on the significant content change of frames in the video shots. However, the key frames maybe not be typical, and maybe choose too many key frames when shot moves fast.

In this paper, based on the methods mentioned above, one key frame extraction approach is proposed depend on improved block color features and second extraction. Experiment shows that the key frames by this approach have a better representative and completeness, besides, could express the mainly contents of the video plenary.

The remainder of this paper is arranged as follow. In section II, the key frames extraction algorithm is discussed which is proposed in this paper detailed; experimental results are presented and compared with other algorithms in section III; finally, concludes are drawn, and some future work is also discussed in section IV.

II. THE APPROACH KEY FRAMES EXTRACTED BASED ON IMPROVED BLOCK FEATURES AND SECOND EXTRACTION

A. Improved Block Color Features Extraction

HSV color space is directly corresponding to the three elements of the visual characteristics of human, and each channel is independent, so it can be independent of the changes of the color components. We must quantize it into several bins; Hue panel is quantized into seven non-uniform colors,

Saturation panel is quantized into two non-uniform colors, and Value panel is quantized into two non-uniform colors too. Then the three color components are converted into one component according to the formula (2). As a result, we could compute the color histograms of 36 bins depend on this quantization method, because the quantization result has only 36 bins, the computational complexity is decreased tremendously.

$$K = 4H + 2S + V + 8 \quad (2)$$

According to the HSV, the central region of an image is usually more conspicuous. Color histogram is just the color statistics of the image; it ignores the information of the color spatial distribution. For the sake of reflecting the spatial information of image, traditional approach of dividing is to divide image into $m \times n$ blocks, but the central region of image is not highlighted. In order to highlight the important of the central region and spatial information, we adopt an approach of partitioning depend on equal area of rectangular ring in this paper, and then extracts the color feature from each rectangular ring. The specific steps are described as follows:

Step1: Determine the center point O of image, calculate the length of each rectangular side a and b . Assuming that each rectangular ring is denoted as R_1, R_2, \dots, R_N according to the order from the center to outside, as Figure 1 shows, we can calculate the length of each rectangular side a and b , according to the formula (3):

$$\begin{cases} a = m \times \sqrt{\frac{i}{N}} \\ b = n \times \sqrt{\frac{i}{N}} \end{cases} \quad (i = 1, 2, \dots, N) \quad (3)$$

where m, n represents the size of side of the image, N denotes the number of rectangular rings.

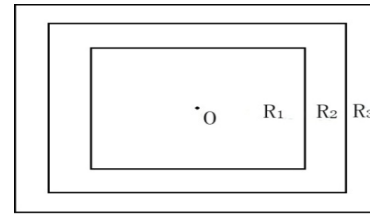


Figure 1. Equal area of rectangular ring

Step2: Calculate the color histogram of 36 bins of each rectangular ring based on above mentioned quantization scheme, and denote as $L_i (i=1, 2, \dots, N)$ from the center to outside.

Step3: Give each rectangular ring a different weight ω_i , in order to highlight the central portion and reduce the influence of edge portion of frame, the value of weight gradually decrease from center to outside, ω_i satisfy the formula (4):

$$\sum_{i=1}^N \omega_i = 1 \quad (4)$$

where N represents the quantity of rectangular rings.

Step4: Consequently the feature vector of each image can be defined as formula (5):

$$F = \sum_{i=1}^N \omega_i L_i \quad (5)$$

where L_i denotes the color histogram of 36 bins of each rectangular ring, F denotes the feature vectors of image.

B. Similarity Measure of the Inter-Frame

The measurement of frames similarity is important. We adopt Euclidean distance to calculation the distance of Inter-Frames in our paper. The smaller Euclidean distance reflects the higher similarity measure between frames. The frames similarity is defined as formula (6):

$$S(H_i, H_j) = 1 - \sqrt{\sum_{k=0}^{35} (F_{i,k} - F_{j,k})^2} \quad (6)$$

where H_i and H_j denotes the i -th frame and j -th frame of the video, $F_{i,k}$ and $F_{j,k}$ denote k -th dimensional feature value of the i -th frame and j -th frame.

C. Initial Key Frame Extraction

This approach extracts key frame depend on the significant color features changes of frames. This approach firstly chooses the first frame as key frame, and it is regarded as a reference frame, and then calculation the distance between reference frame and back frame according to priority, the k -th frame is selected as new current key frame until the similarity between the k -th and current key frame less than an adaptive threshold. The specific steps are described as follows:

Step1: Extract the feature vectors of all frames according to the method mentioned above.

Step2: Select different threshold could gain different key frame. We adopt an adaptive approach to determine the threshold in this paper; the threshold δ is defined as formula (7):

$$\delta = \frac{1}{N} \sum_{i=1}^{N-1} S(H_{i+1}, H_i) - \Delta \quad (7)$$

where N denotes the total number of frames in the video, H_{i+1} and H_i denote the feature vector of the $(i+1)$ -th frame and i -th frame, $S(H_{i+1}, H_i)$ denotes the similarity of the inter-frames, Δ denotes the adjust value of the threshold δ .

Step3: The first frame is selected as current key frame.

Step4: Computer the similarity S between current key frame and next frame by formula (6), if $S < \delta$, and then select this frame as key frame, and update it as new current key frame. Otherwise, continue to calculate the similarity with the next frame. Where δ denotes the adaptive threshold.

Step 5: Repeat step 4 until all the key frames are selected.

We take a video for example. The content of video is that a small plane is taking off, then a car catch up with and overtakes the plane, finally, they all disappear. The number of frames is 994 in this video; the adjusted value Δ of threshold δ is set as 0.15, the number of rectangular rings N is set as 3. The Figure 2 displays the key frames in initial extracted.

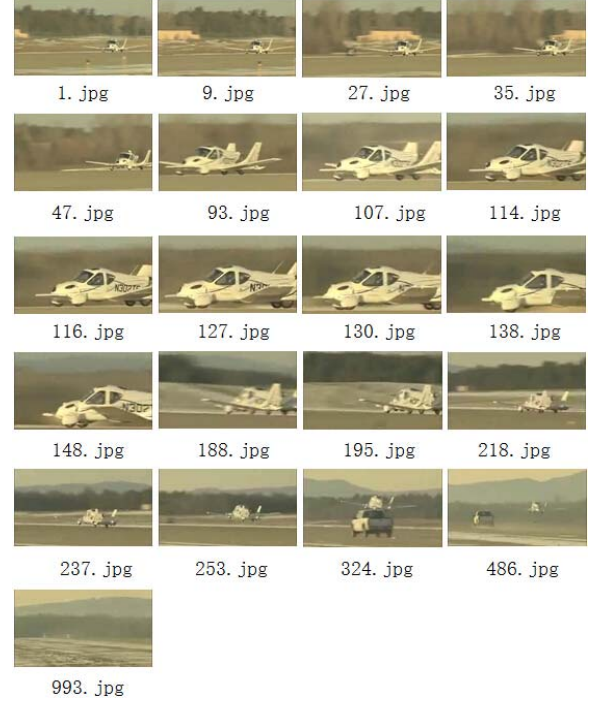


Figure 2. The initial extracted key frames

D. Secondary Extracted Key Frame

Figure 2 shows that the extracted key frame has some redundancies. In comparison, we can find that the redundant key frames have fast moving objects or a sudden flash in the video. As a result, the similarity is small between the two frames of similar content, and one frame is mistaken for key frame. For example, there is an obvious brightness change between frame 93 and 107. In addition, the position of repeated key frames is adjacent. In order to solve this issue, an approach of secondary extraction depend on the position of frames is proposed in this paper. The detailed steps are described as follows:

Step1: Assuming that the amount of key frame is M at initial extracted. f defined as the video frame rate, μ is the multiple coefficient of frame rate, the adaptive threshold of inter-frames is defined as $d = \mu f$. Mark the sequence p of key frames extracted at the first time. The array P of key frames sequence is defined as $P = \{p_i | i=1, 2, \dots, M\}$.

Step2: Calculate the sequence number difference between two adjacent key frames in order, and the difference array is defined as $A = \{a_i | i=1, 2, \dots, M-1\}$.

Step3: If $a_i < d$, remove the key frame of the sequence is p_{i+1} (the later key frame of the two adjacent redundant key frames). Otherwise, they all are selected as key frame. Where a_i denotes the sequence number between $(i+1)$ -th key frame and i -th key frame, d denotes the adaptive threshold of inter-frames.

We optimize the key frames which have been extracted at the first time. In the experiment, $\mu=1.4$, $N=3$, $\Delta=0.15$, finally, the result displayed at Figure 3:

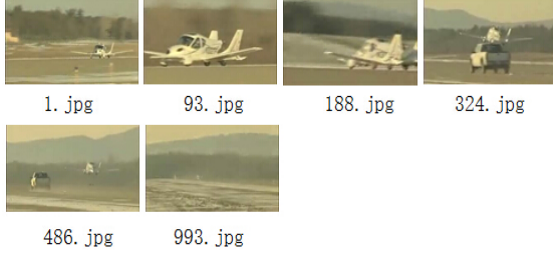


Figure 3. The secondary extracted key frames

As is shown in Figure 3, the key frame could repress the theme of the video effectively. Furthermore, the redundancies of this algorithm are relative low. We know that selecting different threshold parameters could gain different key frames. However, some methods only select a fixed threshold. This paper adopts adaptive threshold parameters according to the different characteristics of video. This method can be applied to different video.

III. EXPERIMENTAL RESULTS AND ANALYSIS

For the sake of checking the validity of the proposed algorithm, a wide range of experiments have been done in lots of videos. These videos include lots of various types of content: sports, cartoons, news, movies and so on. At the same time, we compare the performance of this method with existing methods: traditional HSV color histogram method, k -means method. In order to estimate the results of this method, we use Precision Ratio (P) and Recall Ratio (R) which can be defined as the following formulas:

$$P = \frac{N_c}{N_c + N_f} \times 100\% \quad (8)$$

$$R = \frac{N_c}{N_c + N_m} \times 100\% \quad (9)$$

where N_c represents the amount of the correct extracted key frames, N_f represents the amount of the false key frames, and N_m represents the amount of missing key frames.

For the sake of gaining a combination measure, we choose a measure standard of the F-measure, the average value of F-measure can be defined as following formula:

$$F = \frac{2 \times R \times P}{R + P} \quad (10)$$

The higher value of F -measure thus indicates the higher value for both P and R . The experimental results of different methods are shown in Table I, II, and III:

TABLE I. THE EXPERIMENTAL RESULTS BY K -MEANS METHOD

Video ID	No. Frames	No. Key-Frames	No. Extraction	N_c	N_m	N_f	P	R	F
1	994	7	11	6	1	5	54.5%	85.7%	0.666
2	3581	18	15	13	5	2	86.7%	72.2%	0.788
3	410	10	6	6	4	0	100.0%	60.0%	0.750
4	283	5	7	5	0	1	71.4%	100.0%	0.833
5	1988	9	8	7	2	1	87.5%	77.8%	0.824

TABLE II. THE EXPERIMENTAL RESULTS BY TRADITIONAL HSV COLOR HISTOGRAM METHOD

Video ID	No. Frames	No. Key-Frames	No. Extraction	N_c	N_m	N_f	P	R	F
1	994	7	5	5	2	0	100.0%	71.4%	0.833
2	3581	18	13	11	7	2	84.6%	61.1%	0.710
3	410	10	12	8	2	4	66.7%	80.0%	0.728
4	283	5	5	4	1	1	80.0%	80.0%	0.800
5	1988	9	17	9	0	8	52.9%	100.0%	0.692

TABLE III. THE EXPERIMENTAL RESULTS BY OUR PROPOSED METHOD

Video ID	No. Frames	No. Key-Frames	No. Extraction	N_c	N_m	N_f	P	R	F
1	994	7	6	6	1	0	100.0%	85.7%	0.923
2	3581	18	17	15	3	2	88.2%	83.3%	0.857
3	410	10	10	9	1	1	90.0%	90.0%	0.900
4	283	5	4	4	1	0	100.0%	80.0%	0.889
5	1988	9	11	9	0	2	81.8%	100.0%	0.900

According to above experimental results, we can find that the presented approach in this paper is more effective than Traditional HSV color histogram method and k -means method. However, sometimes the Recall or Precision of key frame extracted in ways of approach is proposed in this paper is lower than other two approaches in different videos. For example, the Recall of this method and k -means method all can reach 85.7% in the video 1. But too many key frames are extracted by k -means method, as a result, the Precision is too small, finally, the value of F -measure is less than the proposed method in this paper. The Precision of this method and traditional HSV color histogram method all can reach 100.0% in the video 1. However, less key frames are extracted by traditional HSV color histogram method, lead to the Recall is too small, finally, the value of F -measure is less than the proposed method in this paper.

For the sake of illustrating the effectiveness of this approach and the representative of the extracted key frames intuitively, we use another video as example. In this video, the backdrop is not moving but the person is moving. Its' content is two man meet in a room, talking with each other for a while and then they all leave. The total number of frames is 283. We extract key frame by means of traditional HSV color histogram. The result is displayed in Figure 4:



Figure 4. Key frames extracted by traditional color histogram method

Then, we extract key frames by means of approach is proposed in this paper. In this experiment: $\Delta=0.15$, $\mu=1.4$. The proposed method extracts 4 frames as key frame; the results are shown in Figure 5. It can be seen that these key frames could express the main dynamic process in Figure 5.





























































Figure 5. Key frames extracted by our proposed method

Figure 5 compares with Figure 4. The proposed approach could extract the optimal key frame which could represent the theme of the video adequately in this paper, and it has less redundancy. Meanwhile, this method considers the changes of lighting condition. Furthermore, in order to highlight the central part of the frame, we adopts the approach of partition depend on equal area of rectangular ring in this paper.

Finally, we compared the proposed approach with [12], [13] and [14]. The key frame extracted by different methods is presented for the documentary video 'Hurricane Force-A Coastal Perspective, segment 03'. The standard key frame and the extracted key frame by different methods for this video are displayed in Table IV. It could be found that there are some important frames missing from the [12], [13] and [14]. Moreover, the extracted key frames have some redundancies in [13]. However, it can be observed that the key frame extracted by our proposed method is closest to the standard key frame compared with the other methods.

TABLE IV. THE KEY FRAME EXTRACTED BY DIFFERENT METHODS

Method	Extracted key frames											
Standard key frames												
[12]												
[13]												
[14]												
Proposed												

IV. CONCLUSIONS

This paper proposes an effective key frames extraction approach in accordance with the weakness in the traditional key frames extraction methods, this approach depend on

improved block color features and second extraction. In this method we divide the frame into equal-area rectangular rings, and then extract the color features from each rectangular ring. In order to highlight the central region and reduce the influence of edge portion of frame, the value of weight gradually

decrease from center to outside, meanwhile, spatial information of the frame is also considered. Selecting key frame in accordance with the significant variety of video frames, and then secondary extract and optimize key frames according to the initial key frames' position in the video. The experiment result shows that the key frame extracted by this method is well in expressing the primary theme of video. Moreover, it has a strong robustness and is suitable for various types of video. Of course, this method has some shortcomings. For example, maybe lose some key frames in second extraction process, since the interval of first extracted key frames is too small. The frames are divided rectangular ring, leading to increase computational complexity in some ways. In the future, we will try to solve these issues.

ACKNOWLEDGMENT

This work is financially supported by self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No. CCNU10A01012).

REFERENCES

- [1] E. Mendi, B. Coskun, "Shot boundary detection and key frame extraction using salient region detection and structural similarity," *Proceedings of the 48th Annual Southeast Regional Conference*, 2010.
- [2] Z. Qu, T. F. Gao, and Q. Q. Zhang, "Study on an Improved Algorithm of Video Keyframe Extraction," *Computer Science*, 2012, 39(8), pp. 300-303.
- [3] H. Zhang, J. Wu, D. Zhong, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, 1997, 30(4), pp. 643-658.
- [4] M. M. Yeung, B. Liu, "Efficient matching and clustering of video shots," *Proceedings of IEEE ICIP*, 1995, pp. 338-341.
- [5] W. Worf, "Key frame selection by motion analysis," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, 2, pp. 1228-1231.
- [6] A. Hanjalic, H. J. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 1999, 9(8), pp. 1280-1289.
- [7] K. Sanjay K, P. Rameswar, and C. Ananda S, "Video key frame extraction through dynamic Delaunay clustering with a structural constraint," *Journal of Visual Communication and Image Representation*, 2013, 24(7), pp. 1212-1227.
- [8] M. Furini, F. Geraci, and M. Montangero, "STIMO: STILL and MOving video storyboard for the web scenario," *Multimedia Tools and Applications*, 2010, 46(1), pp. 47-69.
- [9] S. Avila, A. B. P. Lopes, and L. J. Antonio, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, 2011, 32(1), pp. 56-68.
- [10] F. S. Wang, D. Xu, and W. X. Wu, "A Cluster Algorithm of Automatic Key Frame Extraction Based on Adaptive Threshold," *Journal of Computer Research and Development*, 2005, 42(10), pp. 1752-1757.
- [11] A. Joshi, S. Auephanwiriyakul, and R. Krishnapuram, "On fuzzy clustering and content based access to networked video databases," *IEEE conference, Eighth International Workshop on Continuous-Media Databases and Applications*, 1998, pp. 42-49.
- [12] P. Jiang, X. L. Qin, "Key frame based video summary using visual attention clues," *IEEE Transactions on Multimedia* 2010, 17(2), pp. 64-73.
- [13] J. L. Lai, Y. Yi, "Key frame extraction based on visual attention model," *Journal of Visual Communication and Image Representation*, 2012, 23(1), pp. 114-125.
- [14] N. Ejaz, I. Mehmood, S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, 2013, 28(1), pp. 34-44.
- [15] H. J. Zhang, J. H. Wu, and D. Zhong D, "An Integrated System for Content-based Video Retrieval and Browsing," *Pattern Recognition*, 1997, 30(4), pp. 643-658.
- [16] H. L. Ding, H. X. Chen, "Key frame extraction algorithm based on shot content change ratio," *Computer Engineering*, 2009, 35(13), pp. 225-227.
- [17] Y. J. Gu, Y. Xie, T. Xia, "Keyframe Extraction Based on Representative Evaluation of Contents," *Computer Science*, 2014, 41(8), pp. 286-288.