An Integrated Scheme for Video Key Frame Extraction

Mei Huang, Ling Xia, Jin Zhang, Hui Dong
School of Electrical and Information Engineering
Xihua University
Chengdu, China
hmalmc@163.com, xialing.cd@gmail.com

***Abstract*—Key frame extraction is an essential technology of content-based video retrieval and directly influences the retrieval efficiency. In this paper, an integrated key frame extraction method is presented. Firstly, we get three candidate key frames by shot boundary and visual content based method . Then, compare the similarity of histogram between candidate forums. Finally, we extract key frames by proposing rules. Experimental results illustrate the proposed method is feasible and efficient.**

Keywords-Video retrieval; Key frame extraction; Color histogram; Threshold value.

I. INTRODUCTION

In recent years, with the rapid development of network and multimedia technology, digital video applications have been involved in every aspect of science technology and daily life. People can sit at home through the Internet to surf remote multimedia database, such as video on demand, electronic shopping, etc. These aspects with broad prospect of business make the study of video retrieval technology increasingly widespread attention. Video integrate other media information (such as text, graphics, images, audio, etc.) into a single data stream, so contain more abundant information than other media, but it doesn't like text which gives its content or compares the content directly. In this case the content-based Video Retrieval (CBVR) comes forward and becomes a research hotspot. How to efficiently retrieve the extremely large amount of video data to satisfy users' need is the key problem. To manage and retrieve the video data， an effective method of key frame extraction is necessary.

Generally, the workflow of content-based video retrieval system is shown in figure 1. The first step is to segment the video series into basic units called shots that are composed of a sequence of frames taken by one camera without interruption. And then extracts key frames, also known as the representative frames. Comparing with the original video, key frame sequence is only part of the video frame. Therefore the required storage space is much smaller, is better for video database to quickly query, searching and browsing effectively, and can greatly reduce the amount of video data, improve the efficiency of video retrieval and manage at low cost.

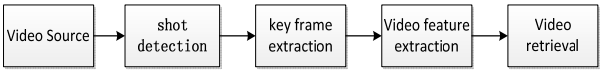


Figure1. Content based video retrieval system

Key frame extraction is an important technology in video retrieval system. According to the complexity of the different shots, one or several key frames can be extracted from a lens. Key frames, also known as the representative frames [1] which reduces the amount of data required in video indexing and reflects the main event of shots. Generally, the principle to select key frames is conservative, namely "rather wrong not less''. We would rather choose overmuch, also won't miss important content. At present key frame algorithms basically can be categorized into following six classes:

Shot boundary based approach. This method is to extract frames from a fixed position in a lens as key frames, which mainly chooses the first frame, middle frame and the last frame as key frames. The advantages are simple, on the contrary, the disadvantages are the number of key frames in each shot is stationary. For short and high similarity of shots will cause redundancy, it has not enough key frames to generalize the content for long and change greatly.

Visual content based approaches. The key frames are detected using low-level features like edge, color and texture changed significantly. Reference [2] selects the first frame of the lens as key frames, and then calculate histogram distance between the current frame and the last key frame. If it is above a certain threshold, the frame will be selected as the key frames. Reference [3] Calculate the distance between the histogram of the current frame and the average histogram of N frames to determine the key frames. Reference [4] first create a reference frame (the creation method of the reference frame refers to [5]), then compare the histogram distance of the current frame and the reference frame, and then select key frames by comparing with a constant threshold. The method is that the number of key frames extraction changes automatically with time of shot content. But for video content changes rapidly, the number of the key frames is often too much and redundancy.

Clustering based approaches. This algorithm is mainly to calculate the distance from each frame to several existing clustering center, if the distance is less than the threshold value which set before is classified into the smallest distance cluster, or create a new cluster. The advantages of it eliminate correlation while can't guarantee the time sequence and dynamic of the image.

Motion analysis based approache. The method exploits movement information of objects in the video or camera to extract key frames. Reference [6] compute the amount of exercise of the lens by optical flow analysis, and select key frames at the local minimum point. In addition, there were many scholars to analyze objects in the video and camera movement through the motion vectors in MPEG video streaming to extract key frames. According to the size of movement of objects or camera to extract key frame, however, the calculation of this method is considerable.

Compressed video stream extraction based approaches. This key frame extraction method directly analyzes and processes some feathers of the compressed video data, thus greatly reduces the complexity of the calculation. Another algorithm for MPEG compressed streams to detect extract key frames with DCT DC coefficients and motion vectors, so there is no need for the full decompression.

The focus of this work is based on content-based research of key frame extraction technology in video retrieval. It is a fundamental process for content-based video retrieval and video analysis and directly affects the efficiency of retrieval. Among them, as a crucial step in content-based video indexing and retrieval system, shot boundary detection is one of premise processes. On the basis of shot segmentation, shot decomposes into a successive image frames, and extract key frame through comparison of the gray-level histogram of these image frames, then present constraint rules. This paper puts forward an improved key frame extraction algorithm, which can improve the accuracy and efficiency of the algorithm.

II. IMPROVED KEY FRAME EXTRACTION ALGORITHM

This paper utilizes OpenCV to select representative frames based on Shot boundary and Visual content frames so as to make integration and improvement.

A. Similarity comparison based on the OpenCV

OpenCV provides abundant visual processing algorithms, and written by C++ language. The main functions are including image data operation, image/video input and output, basic image processing, etc.

To compare the two histograms (H_1 , H_2) [7], first we must choose a comparative measure of histogram similarity. OpenCV function 'cvCompareHist' execute specific histogram comparison. This function provides four kinds of contrast standard to calculate the similarity:

Correlation (method = CV_COMP_CORREL) :

$$d_c(H_1,H_2)=\frac{\sum_i (H_1'(i)\cdot H_2'(i))}{\sqrt{\sum_1 H_1'^2(i)\cdot H_1'^2(i)}} \quad (1)$$

Where $H_k'(i)=H_k(i)-\frac{1}{N}\sum_j H_k(J)$ and tbinsN equals the number of bin in the histogram.

For COREEL, the larger value means better match. Exactly match value is 1, or it 'll be -1, and 0 indicates no correlation (random combination).

Chi-square(method=CV_COMP_CHISQR):

$$d_{chi}(H_1,H_2)=\sum_i \frac{(H_1(i)-H_2(i))^2}{H_1(i)+H_2(i)} \quad (2)$$

For chi-square, the match degree of low score is better than the high. Exactly match value is zero, completely not match for the infinite values (depends on the size of the histogram).

Histogram intersection (method = CV_COMP_ INTERSECT):

$$d_h(H_1,H_2)=\sum_i (H_1(i),H_2(i)) \quad (3)$$

For histogram intersection, a higher score shows better matching, or it'll represent bad matching. The histogram comparison first is normalizing their statistical histogram to 1 by using 'cvNormalizeHist' function, afterwards comparing the statistical distribution of the histogram normalized. If both histograms normalized to 1, the perfect matching is 1 and incompletely is 0.

Bhattacharyya (method = CV_COMP_ BHATTACHARYYA):

$$d_b(H_1,H_2)=\sqrt{1-\sum_i \frac{\sqrt{H_1(i)\cdot H_2(i)}}{\sum_i H_1(i)\sum_i H_2(i)}} \quad (4)$$

For Bhattacharyya distance, low score shows good matching, and high marks express bad matching. The match exactly is zero, imperfect is 1.

This paper used the histogram correlation similarity to compare two images, and proposed an improved key frame selection principle. Through many experiments we choose a suitable threshold to obtain more representative key frames.

B. Proposed algorithm

According to different criteria of key frame extraction algorithms and the specific situation of different videos, we can choose the appropriate decision. The specific method statement is as follows:

1) Candidate key frame selection. According to the information theory, different (or less correlationship) frame image carry more information than similar frame. So when extract more key frames, criteria is mainly consider the dissimilarity between frames.

If f represent a frame image, let f_i (f_i, 0 < i <= N) denotes ith frame of a shot. First we selected the first frame (f_1), the last frame (f_N) by adopting the method based on shot boundary as a candidate key frame, then the average histogram ()of all the frames in the shot, finally select the frame (fi) which is the most close to the average frame to be a candidate key frame. So we get three candidate key frames.

2) Key frame extraction. When extract key frames, we use visual content based approaches to calculate the similarity ($d_c(H_1,H_N)$) between the first and last key frames, and compare them with predetermined thresholds T1, T2 (They are based on a lot of experiments, different types of video data to set two suitable thresholds). Detailed steps are as follows:

- If $d_c(H_1,H_N)$ is bigger than T2, they are similar and take fi to be the key frame;

- If $d_c(H_1, H_N)$ is lower than the T1, the difference between them is large, so the three frames are counted as the key frames;
- If $d_c(H_1, H_N)$ is concludeT1,), we will canculate $d_c(H_1, H_N)$, $d_c(H_1, H_i)$, $d_c(H_i, H_N)$, that is we will put the smallest similarity of two frames as the key frames.

III. EXPERIMENTAL RESULTS AND ANALYSIS

This paper utilizes Microsoft Visual Studio 2010 and OpenCV2.4.3 programming environment to realize the contrast experiment between the traditional method of key frame extraction based on the lens and the improvement of key frame extraction method.

Experiment: We adopt a music and a football video. They are AVI file format. The music video can convert into 388 frame images through the program. Through shot segmentation procedure, the video was divided into four shots. It describes the performance of singers. The shots with slow transformation, has an obvious gradual process, and the way is fading out. Football video was converted into 74 frames and was divided into 2 shots. It represents a football campaign. It is cut shot obviously.

By the procedure described before, we choosed a comparative measure to calculate the distance $d_c(H_1, H_N)$ and the average $\overline{H}$. We got music video data in TABLEI and football video data in TABLE II.

TABLEI. MUSIC VIDEO DATA

	Shots			
	1 (1-92f)	***2 (93-206f)***	***3(207-350f)***	***4(351-388f)***
$d_c(H_1, H_N)$	0.660566	0.415264	0.337408	0.788409
$\overline{H}$	$5.042*10^7$	$4.346*10^7$	$4.726*10^7$	$2.876*10^7$

TABLEII. FOOTBALL VIDEO DATA

	Shots	
	1 (1-63f)	***2 (64-74f)***
$d_c(H_1, H_N)$	0.751694	0.94004
$\overline{H}$	$2.02682*10^7$	$2.27532*10^7$

By TABLEI and TABLEII, using the traditional method elects representation frames based on Shot boundary and Visual content frames. And we got three Candidate frames. Then we compare $d_c(H_1, H_N)$ with (T1, T2), and extract key frames by the rules provided. The threshold of the music video is 0.4 and 0.6. And the threshold of football video is 0.80 and 0.94. The results as TABLEIII and TABLEIV shows.

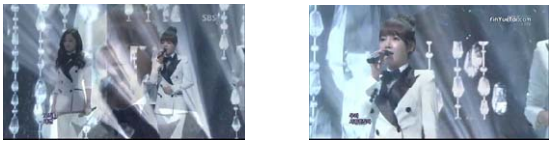
TABLEIII.KEY FRAME EXTRANCTION IN MUSIC VIDEO

	Shots			
	1 (1-92f)	***2 (93-206f)***	***3 (207-350f)***	***4 (351-388f)***
Candidate frames	1, 77, 92	93,149, 206	207,211,350	351,354,388
Key frames	77	149,206	207,211,350	354

TABLEIV. KEY FRAME EXTRACTION IN FOOTBALL VIDEO

	Football	
	1(1-63f)	***2(64-74f)***
Candidate frames	1, 41, 63	64, 74
Key frames	1, 41, 63	64

Take shot 2 of music video for example. Shown as Figure 2, (a) is taking the first frame as key frames by shot boundary method. (b) get the last frame as key frame. Figure 3 shows the key frames by proposed method.



(a) The first frame (b)The last frame

Figure2. The key frames of shot 2 in music video based on shot boundary

Figure3. The key frames of music video extracted by proposed method

As the show from image frames above, it is a music video. Comparing the three results, and it can be seen that the traditional method to extract the first frame as key frames can't discern the movement of the lens, and to take the last frame method as key frames is unable to react the condition before moving. In brief, the improved method reflects not only lens transformation, but also the main content of the shots. It is more representative than traditional methods.

Take football video for example. Shown as Figure 4, (a) is taking the first frame as key frames by shot boundary method. (b) get the last frame as key frame. Figure 5 shows the key frames extracted by the proposed method.

(a)The first frame (b)The last frame

Figure4. The key frame of football video based on shot boundary method

Figure5. The key frames of football video extracted by proposed method

It is obvious that is a video of the football game from the frame images, and there are great changes in the content of shot. Comparing the three key frame extraction results, it can be seen that the traditional method to extract the first frame as key frames can't response the content of the lens, only a frame of the midfield passing or goalkeeper holding. In comparison, the improved key frame extraction method to extract key frames reflect the main content of the lens better, including midfield pass, a front grab the ball and goalkeeper holding. It is more representative than traditional methods.

IV. CONCLUSION

This paper proposed an effectively key frame extraction method. The algrithm selects three candidate key frames based on two classical approaches, then compare the similarity of histogram between candidate frames, and determine the key frames by proposed strategy. It extracts one to three key frames according to shot content automatically. Experiments prove that the algorithm requires a smaller calculation and has a better description ability of time domain changes for video frames.

ACKNOWLEDGMENT

This work was financially supported by the Sichuan Provincial Key Lay on Signal and Information Processing (S2jj20110010), the key research project of Xihua University (Z1120945), and 2011 Chunhui program, Chinese Ministry of Education, Z2011090.

REFERENCES

[1] Chan T F, Vese L A. Active contours without edges[J].IEEE Transanction. 2001, 10(2),pp.266–277

[2] Xiaomu Song, Guoliang Fan. Joint Key-Frame Extraction and Object Segmentation for Content-Based Video Analysis [J]. IEEE TRANSACTIONS, 2006, 7(16), pp.904–914

[3] Z. Sun , K, Jia, and H. Chen. Video Key Frame Extranction Based on Spatial-Temporal Color Distribution[C]. Proc. Conference on Intelligent Information Hiding and Multimedia Singnal Processing, 2008, pp.196–199..

[4] Sze K W, Lam K M, Qiu G P. A New Key Frame Representation for Video Segment retrival[J] . IEEE Transaction on Circuits and Systems for Video Technology, 2005 ,15(9) , pp. 1148–1155. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Senlin Luo, Shu jie Ma, Jing Liang , etc. Based on the sub-shots clustering method of key frame extraction technology [J]. Journal of Beijing institute of technology, 2011, 31(33),pp.348–352.

[6] Ling Shao, Ling Ji. Motion Histogram Analysis Based Key Frame Extraction[J]. IEEE Canadian Conference on Computer and Robot Vision, 2009, pp,88-92

[7] XiaoMeng Xie , Shaofa Li. A new key frame redundancy removal algorithm [J]. Journal of south China university of institute of computer science and engineering,2012, 36（S1）, pp. 53-56.