Proceedings of the 2017 IEEE
International Conference on Information and Automation (ICIA)
Macau SAR, China, July 2017

Key Frame Extraction Based on Dynamic Color Histogram and Fast Wavelet Histogram

Zhikai Zong[1st]
School of Mechanical, Electrical and Information Engineering
Shandong University
Weihai 264209, China
zzksdu@163.com

Qing Gong[2nd]
School of Mechanical, Electrical and Information Engineering
Shandong University
Weihai 264209, China
gongqing_mail@163.com

***Abstract* - There has been a dramatic increase in video data on the Internet in recent years. The abundant video data result in the request for content-based video retrieval (CBVR), which uses efficient techniques to index, retrieve or browse these data. For this purpose, we combine color details derived from dynamic color histogram and texture details derived from fast wavelet histogram to select key frames. Then, the number of key frames is determined by optimized k-means algorithm. Finally, we remove the redundant key frames and get the final key frames by way of mutual information. The effective gradual shot transition is obtained by means of variance curve of dynamic color histogram. Three test videos including movie, sport and lecture are adopted to evaluate the proposed method on performance of key frame extraction and gradual shot detection.**

Index Terms* - *dynamic color histogram. fast wavelet transform. gradual shot detection. key frame extraction. content-based video retrieval.

I. INTRODUCTION

With the rapid growth of video data, retrieving these video data becomes extraordinarily important, and has been widely studied. Selecting key frames is a key step on indexing and retrieving this video data. The first step in video retrieval is to divide the video into smaller units in order to further index and browse. The smallest unit of video is the shot which is defined as continuous action of target in space and time. Shot transition can be divided into abrupt and gradual transition superficially. The abrupt shot transition occurs in a particular frame. While the gradual shot transition occurs among multiple frames and is caused by fading out, fading in and dissolving. The earlier work is focused on the abrupt shot transition. However, related work turns the focus on the gradual shot transition because of its complexity and occurrence over successive frames only. The algorithm of color histogram difference is a typical method for gradual shot transition detection and it is based on the color histogram difference of adjacent frames. On account of the algorithm only counts the pixel brightness and color information, it has strong anti-interference capability. Extracting key frames within each shot is to represent the content of whole shot. In order for video retrieval to be effective, the selected key frames should summarize the entire video without missing any important information.

Up to now, numerous algorithms have been proposed to accomplish the video summary task. Guan et al. [1] selected key frames based on key points which are extracted from all frames using SIFT descriptors [2], and those selected key frames can perfectly cover the global key points. The algorithm has some drawbacks, e.g., key frames may be same if adjacent frames' key points are similar. Jiang et al. [3] applied various low level features to select key frames, including frame correlations, color histogram, and edge histogram. Liu et al. [4,5] proposed a novel model which combined geometry and color features to scene recognition. In [6], a key frame detector based on spectral clustering was presented which combined quantify performance of frame content and evidence of temporal accumulation. In the work of Chen et al. [7], the key frames was selected by multiple features, such as cameral viewpoint, scene analysis and adaptive streaming. However, such method relies on heuristic rules which are extracted from limited data set. Consequently, the algorithm may fail to finish initial test when videos' motion pattern is complex and irregular. Bogdan et al. [8] proposed an intensity-based dissolve detection algorithm which used a twin-thresholding method to reduce the false shot detection. The deficiency of this technique is that it is less efficient dealing with complex shot transitions. The mini number selection of key frames based on sparse dictionary was presented in [9], where the selection adopts a real sparse constraint based on L_0 norm.

Some techniques used clustering for an attempt to extract key frames. For instance, Yeung and Yeo [10] presented a technique to generate key frames of video sequence which consists in each representative scene in the sequence. Therefore, the method is time-constrained when it takes into account both temporal locality and visual properties of the shots. Zhuang et al. [11] proposed a method for key frame extraction using a color histogram based on unsupervised clustering. Doulamis et al. [12] introduced multidimensional fuzzy classification to summarize the video based on segment features which extracted from stereoscopic videos.

In this paper, we first apply the dynamic quantization interval instead of traditional HSV color quantization interval to detect gradual shot transition. The more color details are found out to achieve efficient gradual detection. And we first combine the main color descriptor derived from the dynamic quantization interval and texture descriptor derived from Discrete Wavelet Transform (DWT) to select key frames. The optimized k-means algorithm is used to automatically search the key frames in every shot. Then, mutual information is applied to tackle the problem, similarity of these key frames, because these key frames are derived from different feature descriptors. The mutual information reflects the relevance of content of different frames and removes the spare key frames to improve the selection precision. Finally, the left frames are selected as key frames.

The rest of this paper is organized as follows. Section II briefly summarizes related work of key frames selection. The proposed key frame selection algorithm is described in Section III. Section IV exhibits simulation experiments and analysis of proposed method. Conclusion and further work are described finally in Section V.

II. RELATED WORK

In order to effectively detect the gradual shot transition, Yang et.al [13] calculated the variance of the frame image histogram in the video sequence, and obtained the relative effective gradual shot transitions. In the extraction of key frames, Wolf [14] calculated the optical flow of each frame and the motion metric based on optical flow, and selected key frames at minima of motion.

A. The Gradual Shot Transition Based on HSV Color Histogram

The HSV color space [13] is adopted because the RGB color space characteristics and the perception of the human eye have a great color deviation. HSV color space has two merits: firstly, the color information of the image has nothing to do with the luminance component; secondly, saturation and hue components are very close to the way people perceive color. The quantization scheme is shown as:

$$H=\begin{cases}1 & h\in[0,0.1)\text{or}(0.9,1.0]\\ 2 & h\in[0.1,0.2)\\ 3 & h\in[0.2,0.3)\\ 4 & h\in[0.3,0.4)\\ 5 & h\in[0.4,0.5)\\ 6 & h\in[0.5,0.6)\\ 7 & h\in[0.6,0.7)\\ 8 & h\in[0.7,0.8)\\ 9 & h\in[0.8,0.9)\end{cases}\tag{1}$$

$$S=\begin{cases}1 & s\in[0,0.25)\\ 2 & s\in[0.25,0.5)\\ 3 & s\in[0.5,0.75)\\ 4 & s\in[0.7,1.0)\end{cases}\tag{2}$$

$$V=\begin{cases}1 & \mathrm{v}\in[0,0.2)\\ 2 & \mathrm{v}\in[0.2,0.4)\\ 3 & \mathrm{v}\in[0.4,0.6)\\ 4 & \mathrm{v}\in[0.6,0.8)\\ 5 & \mathrm{v}\in[0.8,1.0)\end{cases}\tag{3}$$

where *H*, *S* and *V* stand for three channels of the HSV color space. After quantization, HSV color space is divided into $9\times4\times5=180$ stems and the three channels merge into one dimension vector:

$$L=20H+5S+V,\tag{4}$$

We can get the change curve of frame image histogram variance for video sequence. The curve shows obvious monotonicity in the gradual changing part of the shots. However, one major problem with this approach is the expensive computation cost caused by high quantization dimension.

B. Key Frame Extraction Based on Motion Analysis

Wolf et al. [14] proposed optical flow analysis method. The total number of optical component of each pixel is calculated firstly as motion metric $M(k)$ of the kth frame. Then, the local minimum of $M(k)$ is selected as the key frame and the amount of motion of the kth frame $M(k)$ is shown as:

$$M(k)=\sum_i\sum_j\left|O_x(i,j,k)\right|+\left|O_y(i,j,k)\right|,\tag{5}$$

where $O_x(i,j,k)$ is the x component of optical flow in the kth frame at location of (i,j) and similarity for y component. Starting from $k=0$, the $M(k)\sim k$ curve is scanned and two local maximums $M(k_1)$ and $M(k_2)$ are obtained. If $M(k_3)=\min\left[M(k)\right],k_1<k<k_2$, the k_3th frame is set as key frame. Then, we use k_2 to reset k_1 and continue looking for the next k_2 .

III. THE PROPOSED ALGORITHM

One characteristic of the gradual shot transition is that there is small changes over continuous frames. Therefore, we have to find more refined color and texture details. We introduce dynamic color space algorithm and fast wavelet histogram algorithm to select key frame. In Wang et al. [15], grading image retrieval was proposed on compressed domain of Discrete Cosine Transformation (DCT) and Discrete Wavelet Transform (DWT) and they obtained a better performance of image retrieval by a dynamic color space and fast wavelet histogram. To select key frames, we make full use of both the color details derived from dynamic color histogram and the texture details derived from fast wavelet histogram.

A. Key Frame Detection Structure

In the process of selecting key frames, we firstly segment the video sequence into a number of segments by dynamic color quantization histogram. And then we select key frames in every shot by the means of two attribute feature respectively. So the number of key frames can be automatically selected by using the optimized k-means[16] algorithm. Therefore, the two different key frame sequences are obtained. Finally, we compare the two different key frame sequence by the mutual information and remove the redundant key frames. A specific detection structure is illustrated in Fig. 1. It includes four main stages:

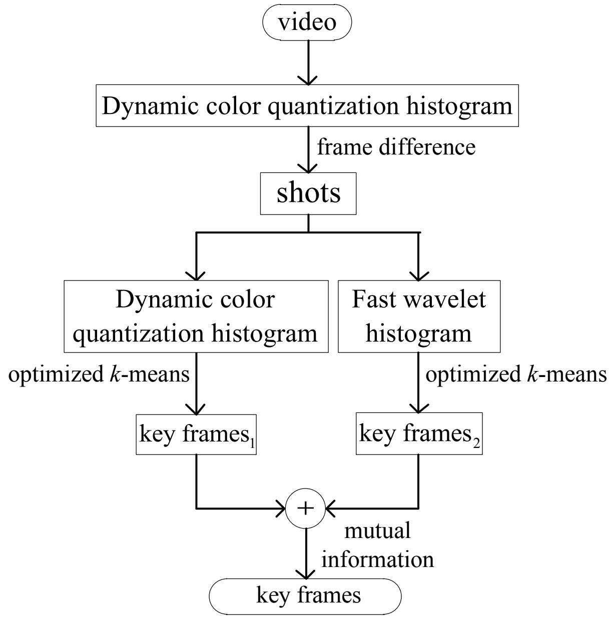


Fig. 1 Flow chart of the proposed algorithm.

*S*1: Shot detection. Generate a set of shots including abrupt shot and gradual shot using dynamic color histogram.

*S*2: Obtain a set of key frames. In every shot, key frames are obtained through dynamic color quantization histogram by way of optimized k-means[16].

*S*3: Obtain another set of key frames. Key frames are obtained through fast wavelet histogram by means of optimized k-means[14].

*S*4: Obtain final key frames. remove redundant key frames using mutual information.

B. *Shot Boundary Detection*

Traditional color histogram can be understood as statistical information about color in image. However, the decrease of retrieval efficiency and larger storage occupation caused by the high color dimension will make it in a dilemma. The proposed algorithm adapts the dynamic color quantization histogram [15] to extract main color information and improve the retrieval accuracy. In other word, we apply the smaller quantization interval to main color space and the larger quantization interval to secondary color space. For every frame in the video, the concrete steps of dynamic color histogram are shown as:

Algorithm 1. The dynamic color quantization histogram:

Step 1. Calculate the existence frequency of every color in image and store them in column vector. The calculation of $\boldsymbol{h}$ is defined as:

$$\boldsymbol{h}(i)=\frac{n_i}{N}, i=1,2,3,\cdots,256, \tag{6}$$

where n_i represents the number of ith color pixel, and N is the total number of frame image pixels in video.

Step 2. Color classification. The color is divided into main color and secondary color. We apply the threshold τ_c to distinguish them. The main color is defined as the color whose frequency $\boldsymbol{h}$ is greater than or equal to τ_c. On the contrary, the secondary color's is less than τ_c. The formula of classification is as shown as:

$$C_i=\begin{cases}\text{main color}, & h(i)\geq\tau_c\\ \text{secondary}, & h(i)<\tau_c\end{cases}, \tag{7}$$

Step 3. Set main color interval. Main color interval can be defined as the most of color that is located in one interval. On the contrary, the remaining color is set to be secondary color interval. When the probability of a continuous three colors are greater than τ_c, the first color in the three colors is considered to be the starting point $\boldsymbol{m}_s$ of the main color interval.

Step 4. The division of quantization interval. Main color is crucial to the performance of similarity. Therefore, smaller quantization interval is applied to main color interval for the purpose of improving the retrieval accuracy. Inversely, larger quantization interval is applied to secondary color interval for the purpose of decreasing the color dimension and improving the retrieval accuracy. We choose 4 and 8 as color interval for main color interval and secondary color interval respectively. Thus, the starting point $\boldsymbol{m}_s$, main color interval and secondary color interval should be multiples of 8. The $\boldsymbol{m}_s$ should be processed to multiply by 8:

$$\boldsymbol{m}_s=ceil(\boldsymbol{m}_s/8)\times 8, \tag{8}$$

where the *ceil* function returns the smallest integer greater than or equal to the specified expression. Furthermore, a fixed main color integer length is set so as to unite a same dimension of color histogram for all frame images. We set the length of main color interval as 104. In this way, the quantization interval is 104/4+(256-104)/8=45 for all frame images.

Step 5. Color histogram statistics. s_i and e_i represent the starting point of ith quantization interval and the ending point of ith quantization interval, respectively. The frequency of the ith color histogram quantization interval is shown as:

$$h(i)=\sum_{n=s_i}^{e_i}h(n), i=1,2,\cdots,45 \tag{9}$$

When the difference of the content of the two frame images is small, the variance of the histogram is smaller in every shot and vice versa. Therefore, we can analyze the variance curve of the video to detect the abrupt shot transition and gradual shot transition. The specific process consists of abrupt shot detection and gradual shot detection. That is as follows:

Algorithm 2. The process of shot boundary detection:

Step 1. Threshold initialization. We set τ as the threshold for mutant shot and τ is set to 0.02 (0.02 is obtained by a large number of experiments). If the histogram difference is larger than τ, the abrupt shot transition occurs.

Step 2. Calculate the mean M and variance V of $h(i)$. The calculations are shown:

$$M = \frac{\sum_{i=1}^{45} h(i)}{45}, \tag{10}$$

$$V = \frac{\sum_{i=1}^{45} (h(i) - M)}{45 - 1}. \tag{11}$$

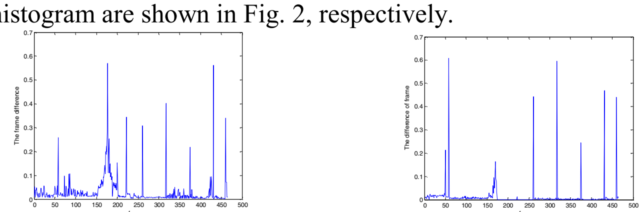
Step 3. Calculate the average of V. In order to reflect the change features of V more obviously, n values from V are taken to calculate the mean and the another curve V' has been gotten.

$$V' = \begin{cases} \dfrac{\sum_{i=t}^{t+n-1} v(i)}{n} & 1 \le t \le F - n \\ \dfrac{\sum_{i=t}^{F} v(i)}{F - t + 1} & F - n \le t \le F \end{cases} \tag{12}$$

where F is the amount of all the frames in the video.

Step 4. Determine gradual shot transition. We calculate the difference of V' and get the curve of $d(t)$. If the length of a positive or negative interval is greater than ξ, the gradual transition interval has been gotten.

The color histogram differences of music video by dynamic quantization color histogram and HSV color histogram are shown in Fig. 2, respectively.



(a) (b)

Fig. 2 Color histogram difference by different methods (a) Dynamic color histogram (b) HSV color histogram.

From the Fig. 2, we can conclude that dynamic color quantization can get better performance on mutant shot detection and perform well on small changes in color details.

C. Fast Wavelet Histogram

For a frame with size of $N \times N$, we obtain M different wavelet histograms after decomposing by m-level wavelet. First, the three-wavelet decomposition is performed on the frame image, and the absolute values of the nine high frequency sub-band coefficients are calculated to obtain the energy of the high frequency sub-band. Next, select a level k as required, the sub-bands of other stages are up-sampled or down-sampled to the same size as the k level band.

We apply the non-uniform quantization algorithm to reduce the dimension of eigenvector. Sub-band $\mathrm{HH}_n, \mathrm{HL}_n, \mathrm{LH}_n$ respectively represent diagonal margin, vertical margin and horizontal margin of frame image. Take vertical margin information as example, only sub-band HL_3 and HL_2 can represent the entire vertical margin information. Furthermore, there is a little effect on retrieval result because the corresponding sub-band coefficients are quantized to zero when level n is lower. Therefore, only sub-band $\mathrm{LH}_3, \mathrm{HL}_3, \mathrm{HH}_3, \mathrm{LH}_2, \mathrm{HL}_2$ and HH_2 are applied to represent texture information of frame image.

The fast wavelet histogram reduces computational complexity to some extent. For example, for a gray-scale frame with size of $N \times N$, the extraction of the eigenvectors need to use N^2 data by the method of wavelet histogram. While only $\frac{N^2}{2^{2k}}$ data is necessary to construct the histogram using fast wavelet histogram.

D. The Mutual Information

A series of key frames are obtained using color features and texture features respectively. However, these key frames are not necessarily ideal, and the key frames of the two algorithms may be similar. Thus, mutual information is applied to further select valid key frames.

As for a gray scale frame image M , the information entropy is shown in (13), and the mutual information between frame image M and frame image N is shown in (14):

$$H(M) = -\sum_i p_M(i) \log p_M(i), \tag{13}$$

$$H(M, N) = -\sum_{i,j} p_{M,N}(i,j) \log p_{M,N}(i,j), \tag{14}$$

where $p_M(i)$ represents probability density for frame image M, and $p_{M,N}(i,j)$ is joint probability density for both frame M and frame N. The probability density $p_M(i)$ and the joint probability density $p_{M,N}(i,j)$ can be obtained by normalizing gray histogram and joint gray histogram, respectively. Furthermore, the mutual information between frame image M and frame image N is shown as:

$$I(M, N) = H(M) + H(N) - H(M, N) \tag{15}$$

Moreover, the mutual information for gray frame images is also defined as:

$$I(M, N) = -\sum_{i \in M, j \in N} p_{M,N}(i,j) \log \frac{p_{M,N}(i,j)}{p_M(i) p_N(j)}. \tag{16}$$

For the color frame image, three mutual information can be obtained by calculating R, G ,and B channel respectively and these mutual information are shown as:

$$I^R_{(M,N)} = -\sum_{i \in M, j \in N} p^R_{M,N}(i,j) \log \frac{p^R_{M,N}(i,j)}{p^R_M(i) p^R_N(j)} \tag{17}$$

$$I^G_{(M,N)} = -\sum_{i \in M, j \in N} p^G_{M,N}(i,j) \log \frac{p^G_{M,N}(i,j)}{p^G_M(i) p^G_N(j)} \tag{18}$$

$$I^B_{(M,N)} = -\sum_{i \in M, j \in N} p^B_{M,N}(i,j) \log \frac{p^B_{M,N}(i,j)}{p^B_M(i) p^B_N(j)} \tag{19}$$

Thus, the mutual information between two color frame images can be written as:

$$I_{(M,N)} = I^R_{(M,N)} + I^G_{(M,N)} + I^B_{(M,N)}. \tag{20}$$

If $I_{(M,N)}$ is greater than threshold value τ_i, both frames serve as final key frames; otherwise, remove one of the two frames.

The application of mutual information is effective to reduce duplication of frames and improve the accuracy of retrieval.

IV. Experimental Results

In this section, we will show performance of proposed method. Because there is no uniform definition of key frame, we define it as the frame which can clearly summarize the overall content of a fragment. As discussed in section III, the threshold τ_c of main color is set to 0.004, and the threshold of final key frames τ_i is set to 0.035. The iteration of k-means is set to 50 for the purpose of getting stable clustering results. Fig. 3 shows key frames of movie which are selected by proposed method, k-means method, and motion analysis. From Fig. 3, the proposed algorithm can detect the most key frames, and the contents of the key frames are not repeated. The k-means method and motion analysis are two kinds of classic key frame extraction algorithm. Compared with the proposed method, they have less effective in terms of the number and the precision of the key frames. We can conclude that the proposed algorithm is more efficient than the other two methods on key frames extraction.

(a) (b) (c)

Fig. 3 The comparison of key frames on movies based on different methods: (a) proposed method, (b) k-means and (c) motion analysis.

TABLE I
THE RECALL RATE OF GRADUAL SHOT TRANSITION

Video database	The recall rate	
	Dynamic color histogram	HSV color histogram[13]
movie	1.00	0.17
sport	1.00	1.00
lecture	0.75	0.50

TABLE II
THE RECALL RATE OF GRADUAL SHOT TRANSITION

Video database	The average accuracy	
	Dynamic color histogram	HSV color histogram[13]
movie	0.71	0.33
sport	0.50	0.50
lecture	0.75	0.67

TABLE III
THE RECALL RATE OF KEY FRAME DETECTION

Video database	The recall rate		
	proposed algorithm	k-means	motion analysis
movie	0.83	0.50	0.50
sport	1.00	0.67	0.50
lecture	1.00	0.33	0.25

TABLE IV
THE AVERAGE ACCURACY OF KEY FRAME DETECTION

Video database	The average accuracy		
	proposed algorithm	k-means	motion analysis
movie	0.50	0.20	0.30
sport	0.80	0.60	0.40
lecture	0.75	0.50	0.20

Table. I and Table. II show the recall rate R and the average accuracy A in gradual shot transition. In addition, R and A range from 0 to 1 and the higher value they are, the better performance of method has. From Table. I and Table. II, the recall rate of proposed algorithm can reach 1 almost. Because main colour can be used to calculated the curve of variance, the better performance on gradual shot transition is obtained. From Table. III, the recall rate of proposed algorithm can reach 1, because mutual information plays a crucial role in removing redundant frames. From Table. IV, the proposed algorithm combines the main colour features and texture features to extract key frames and obtain better performance on key frames extraction.

V. Conclusions

In this paper, a new gradual shot detection method and an efficient key frame extraction algorithm are proposed. From the result, the ideal gradual shot detection is obtained by means of the variance of dynamic color histogram. Hence, we get a better performance of key frame extraction. Due to the fact that main color information and texture detail feature are extracted by descriptors, the optimal k-means algorithm selects key frames automatically. What's more, the mutual information is adopted to remove redundant key frames. For further work, more detailed descriptors will be used to extract key frames.

Acknowledgment

The authors would like to thank Sifan Luo and Anran Yang for their valuable suggestions in revising this paper. The authors also thank the anonymous reviewers and the editor for their constructive comments on this paper.

References

[1] G.L. Guan, Z.Y. Wang, S.Y. Lu, J.D. Deng, and D.D. Feng. "Keypoint-based keyframe selection", IEEE Transactions on Circuits Systems for Video Technology, Vol. 23, no. 4, pp. 729-734, Apr 2013.

[2] D.G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, Vol. 60, no. 2, pp. 91-110, Nov 2004.

[3] R.M. Jiang, A.H. Sadka, and D. Crookes, "Advance in video summarization and skimming," Springer Berlin Heidelberg, Vol. 231, pp. 27-50, 2009. DOI: 10.1007/978-3-642-02900-4_2.

[4] M. Liu, R. Siegwart, "DP-FACT: Towards topological mapping and scene recognition with color for omnidirectional camera," Proceedings of International Conference on Robotics and Automation, USA: Saint Paul, pp. 3503-3508, June 2012.

[5] M. Liu, R. Siegwart, "Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera," IEEE Transactions on Robotics, Vol. 30, no. 2, pp. 310-323, Apr 2014.

[6] V.M. Ricardo, B. Antonio, "Spatio-temporal feature-based keyframe detection from video shots using spectral clustering," Pattern Recognition Letters, Vol. 34, no. 7, pp. 770-779, 2013.

[7] F. Chen, D. Delannay, D. Vleesschouwer, and D. Christophe, "An autonomous framework to produce and distribute personalized team-sport video summaries: a basketball case study," IEEE Transactions on multimedia, Vol. 13, no. 6, pp. 1381-1394, Dec 2011.

[8] I. Bogdan, L. Patrick, "Intensity-driven dissolve detection adapted to synthetic video contents," Journal of Electronic Imaging, Vol. 22, no. 2, pp.1-10, Jun 2013.

[9] S.H. Mei, G.L. Guan, Z.Y. Wang, S. Wan, M.Y. He, and D.D. Feng, "Video summarization via minimum sparse reconstruction," Journal of Pattern Recognition, Vol. 48, no. 2, pp. 522-533, Feb 2015.

[10] M.M. Yeung and B.L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, no. 5, pp. 771-785, Oct 1997.

[11] Y.T. Zhang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frames extraction using unsupervised clustering," Proceedings of International Conference on Image Processing, USA: Chicago, Vol. 1, pp. 866-870, Oct 1998.

[12] N.D. Doulamis, A.D. Doulamis, Y.S. Avarithis, K.S. Ntalianis , and S.D. Kollisa, "Efficient summarization of stereoscopic video sequence," IEEE Transactions on Cisrcuits and System for Video Technology, Vol. 10, no. 4, pp. 501-517, 2000.

[13] B.T. Yang, X.F. Wang. "A gradual shot detection method," Journal of Computer engineering and Application. Vol. 4. pp. 37-40. Feb. 2005. (in Chinese)

[14] W. Wolf, "Key frame selection by motion analysis," Proceedings of the 21st International Conference on Acoustic, Speech, and Signal Processing, USA: Atlanta, Georgia. Vol. 2. pp. 1228-1231. May 1996.

[15] C.Y. Wang, X.Y. Zhang, R Y. Shan, and X. Zhou, "Grading image retrieval based on DCT and DWT compressed domain using low-level features," Journal of Communications. Vol. 10, no. 1, pp. 64-72. Jan 2015.

[16] Q. Wang, C. Wang, Z. Y. Feng, and J. F. Ye, "Review of k-means clustering algorithm", Electronic Design Engineering, vol. 20, no. 7, (2012) Apr., pp. 21-24. (in Chinese).