The comparison and analysis of extracting video key frame

IOP Conf. Series: Materials Science and Engineering **359** (2018) 012010 doi:10.1088/1757-899X/359/1/012010

The comparision and analysis of extracting video key frame

OUYANG ShiZhuang[1],ZHONG Luo[2],LUO RuiQi[3]
1 Master graduate, Wuhan University of Technology, research interests include data mining and image processing,CHN
2 Ph.D., Professor, WuHan University of Technology, research interests include data mining and intelligent system,CHN
3 Ph.D., WuHan University of Technology, research interests include picture processing and computer vision, CHN
Abstract Video key frame extraction is an important part of the large data processing. Based on the previous work in key frame extraction, we summarized four important key frame extraction algorithms, and these methods are largely developed by comparing the differences between each of two frames. If the difference exceeds a threshold value, take the corresponding frame as two different keyframes. After the research, the key frame extraction based on the amount of mutual trust is proposed, the introduction of information entropy, by selecting the appropriate threshold values into the initial class, and finally take a similar mean mutual information as a candidate key frame. On this paper, several algorithms is used to extract the key frame of tunnel traffic videos. Then, with the analysis to the experimental results and comparisons between the pros and cons of these algorithms, the basis of practical applications is well provided.
Key words keyframes; frame difference; threshold; mutual trust

1. Introduction

monitoring video is widely distributed in cities at present. The video monitoring analysis that rely on artificial human eyes is tedious and inefficient considering the large amount of data resources in surveillance video. By extracting the keyframe of the video, the video data will be optimized and compressed, which offer the foundation for the further video semantic analysis.

Rong Pan[1] proposes a keyframe extraction method based on cluster, which finishes the shot boundary detection firstly, and then obtains the child video with smaller data, and finally pick out the nearest frame to the center of the cluster as the keyframe. Zhong adopts the histogram intersection and block weighting, improves the traditional histogram algorithm and solves the defect of shot detection. Thepade[3] comes up with the block encoding idea to extract keyframe, which makes use of the failure vector and similarity degree of two consecutive frames. Huang[4] figures out the mutual information method according to information theory, and confirms that mutual information method contributes to improve the extraction efficiency. Mutual information is a method by calculating the amount of information to determine the threshold and dividing initial class to get keyframe sequence. A. Wang [5] exploits the background difference and light flow method to detect motion object. Wang[6] introduces the adaptive weighted affine propagation algorithm to recognize the keyframe of human body movements among three dimensional video sequence, which has a high precision. K.S.Thakre[7] makes a systematic research on compressing video streaming and presents a way of using self-adaptive threshold method to extract the keyframe of compressed video streaming.

The above words give a summary of domestic and overseas scholars' research on extracting video keyframe within the last five years. The paper presents a mutual information method extracting keyframe and gives an overall analysis combining with other four classical keyframe extraction

algorithm, shot edge analysis, arithmetic average algorithm, histogram averaging algorithm and video content matches.

2. Shot Segmentation

This paper discusses the key frame extraction is based on a single shot key frame extraction. We will be a video for structured analysis，As shown in Figure 1，First, require shot segmentation, and then the keyframe is extracted based on the shot segmentation, and a number of frame images are obtained. The frame image is grouped according to the image content for obtain scene. When shot segmentation, the first need to calculate the frame difference, and the commonly used frame difference method is as follows：Pixel comparison, Histogram comparison, marginal comparison and block matching[8,9]。In this paper, we propose a sub shot segmentation algorithm based on sliding window. The lens segmentation adopts the traditional double threshold method to segment the video for the first time to get the sub shot[10,11].

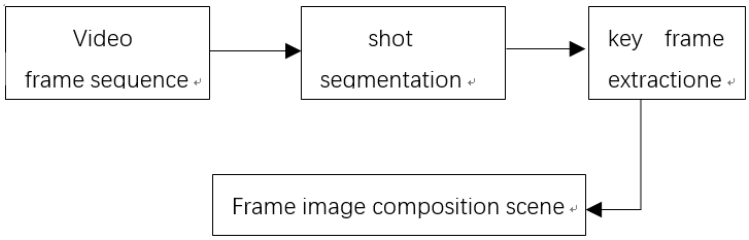

figure1 Video structure analysis flow chart

First, add a sliding window to the video frame sequence and look for possible sub shot boundaries in the small area of the window. The sliding block divides the window into two parts from the middle, and the frame difference between successive frames in the lens is relatively small and can be treated as a class. In this paper, Using fisher linear two kinds of problems. On the two kinds of problems, the idea of Fisher linearity is to project the data samples of multidimensional space into 1-dimensional space, so that the variance between the two classes is as large as possible, and the variance within the class is as small as possible.

Set two classes are C1 and C2，the sample is x, the mean is m1 and m2, the variance is s1 and s2.

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n, (k = 0,1) \quad (1)$$

$$s_k^{\,2} = \sum_{n \in C_k} (x_n - m_k)^2, (k = 0,1) \quad (2)$$

So that different types of class variance as much as possible, that is, so that the ratio of the two $\frac{(m_2-m_1)^2}{s_1^2+s_2^2}$ as large as possible. When the ratio is maximum, the two classes are optimally divided in 1-dim space.

Ideally, the Fisher criterion inside the sub shot is smaller and approaches zero. In practice, the Fisher criterion at the sub shot boundary due to camera or object motion is larger. Since there is a local maximum due to a small fluctuation inside the sub shot, it should be considered as noise. Thus, a local maxima above the Fisher criterion mean is used as the sub shot boundary.

3. Key Frame Extraction

The key frame extraction[12] in this section is based on the sub shot segmentation.

3.*1. Analyze the shot edge to extract keyframes*

Analysis of the edge of the lens to extract the key frame is to select the end of these fragments as the key frame.

Take f_1, $f_{N/2}$, f_N as the pre-selected key frame. f_i and f_j between the interframe difference $D(f_i, f_j)$. The calculations $D(f_1, f_{N2})$, $D(f_1, f_N)$, $D(f_{N2}, f_N)$ are each compared with the threshold T.

If $D(f_1, f_{N2}), D(f_1, f_N)$, $D(f_{N2}, f_N) < T$. it is considered a similar image, Can take any one of them as a key frame; If $D(f_1, f_{N2})$, $D(f_1, f_N)$, $D(f_{N2}, f_N) > T$,the difference between the three frames is relatively large, and they are all as key frames. In addition, determine their distance, the maximum distance corresponding to the key frame[13].

3.2. Frame average method to extract key frames

Keyframe extraction The most straightforward algorithm is based on the frame image algorithm, including the frame averaging method and the histogram averaging method.
The frame averaging method first takes the mean of all frames of the video clip at i:

$$\overline{R_i} = \frac{(\sum_{k=1}^{n} R_{ki})}{n}, \overline{G_i} = \frac{(\sum_{k=1}^{n} G_{ki})}{n}, \overline{B_i} = \frac{(\sum_{k=1}^{n} B_{ki})}{n} \qquad (3)$$

Where n represents the number of frames in the video clip, $\overline{R_i}, \overline{G_i}, \overline{B_i}$ represent the RGB mean, R_{ki}, G_{ki}, B_{ki} representing the RGB value. The difference between each frame and the mean is:

$$D_{ki} = |R_{ki} - \overline{R}| + |G_{ki} - \overline{G}| + |B_{ki} - \overline{B}| \qquad (4)$$

Select the maximum and minimum values of D_{ki} as the key frame.

3.3. The Histogram Averaging Method Extracts Key Frames

The histogram averaging is done by selecting each frame in the video clip, dividing it into a histogram and averaging it. Then the histogram of each frame is compared with the average, and the two extremes are obtained and the corresponding Frame to find out as a key frame [14]。
Histogram comparison method is usually based on the image brightness, color, grayscale these three values, respectively, they are divided into N levels, statistics of each level of the number of pixels, drawn into a histogram. The total pixel of the image frame is M, the gray level is N, the pixel with k-level gray scale has f_k , the statistical histogram of the image color feature is a discrete function, and the frequency of k-level gray scale is：

$$h_k = \frac{f_k}{M}(k = 1,2,\Lambda\ ,N) \qquad (5)$$

The histogram shows the gray value distribution of the image .Figure 2 for the video to take a frame image of the gray histogram. The gray scale histogram of all the frames of the lens is averaged to obtain the key frame.

figure2 gray-level histogram Example

3.4. The Visual Content Matching Extracts Key Frames

Through the visual content matching to extract key frames, the algorithm ideas are as follows: Select a video clip, the first frame as the key frame, the same will also be set to the reference frame, and then after each frame and the reference frame for comparison, when the k-th frame and the selected frame is

greater than the threshold T, then the first k frame to update the new reference frame, will continue to cycle, to the end of all the video clips are detected[12,13]. The flowsheet shown in Figure 3.

A video stream with a frame count of N，

$I = \{ f_1, f_2, \Lambda , f_N \}$；

Select I_1 as the reference frame, define the similarity of the two frames according to the color histogram, and set a threshold T. The correlation coefficients for f_i and f_i are as follows:

$$\rho_{ij} = C_{ij} /(\sigma_i, \sigma_j) \quad (6)$$

In the above formula, $C_{ij} = (f_i - m)(f_j - m), \sigma_i^2 = C_{ii}$, M is the mean vector. Compare the correlation coefficient with the threshold to arrive at a certain number of key frames.

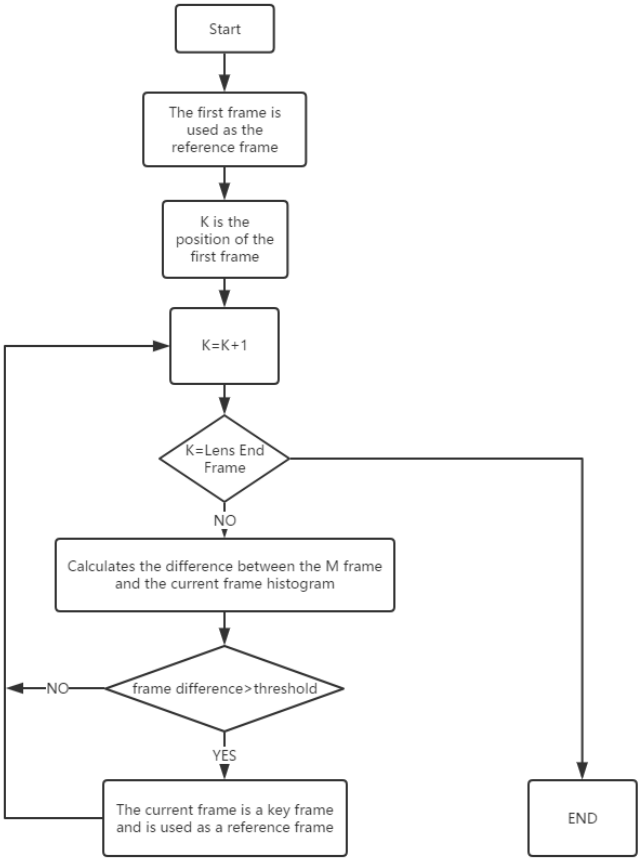

figure3 Extract keyframe flow graphs through visual content matching

4. Key Frame Extraction Base On Mutual Information

The four key frame extraction methods introduced in Section 3 can achieve the extraction of key frames to a certain extent, but there are still some shortcomings. Analysis of the shot edge of the key frame is vulnerable to external influences, when the shot (camera) movement changes, will lead to the shot boundary misjudgment, can not correctly extract the key frame. When the video content is more complex, the frame mean and the histogram method to extract the key frame representation is relatively poor. Through the visual content matching to extract the key frame calculation is moderate, the key frame sequence is reasonable, but does not apply to all the video. Through the above method, this paper proposes a key frame method based on mutual information extraction. The mutual information method introduces the information entropy, can calculate the inter-frame mutual information value, and divide the initial class. At the same time, the method of this paper can calculate the threshold independently [14,15]. In this paper, based on mutual information extraction key frame algorithm steps are as follows.

Step 1: Calculate the higher the ratio of the mutual information I (X, Y) to H (X, Y) for two consecutive frames, the higher the similarity;

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (7)$$

$$H(X,Y) = -\sum_{i,j} P_{XY}(i,j)\log P_{XY}(i,j) \quad (8)$$

Where X, Y represent two random variables, H (X) is the information entropy, and I (X, Y) is called the mutual information of X and Y[16]. P_{XY} is the joint probability distribution of X and Y.

The amount of mutual information for the RGB component is calculated from the mutual information formula given above.

$$I_{X,Y}^{R} = -\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P_{X,Y}^{R}(i,j)\log\frac{P_{X,Y}^{R}(i,j)}{P_{X}^{R}(i)P_{Y}^{R}(j)} \quad (9)$$

the amount of information between the G component and the B component can be obtained.

The mutual information of two consecutive frames is：

$$I_{XY} = I_{X,Y}^{R} + I_{X,Y}^{G} + I_{X,Y}^{B} \quad (10)$$

Step 2: Select the threshold T [17-20].

A video stream with a frame count of N， $I = \{I_{1,2}, I_{2,3}, \Lambda, I_{N-1,N}\}$；

Sort the elements in I to get the new A. $I_{sort} = \{I_1, I_2, \Lambda, I_{n-1}\}$， $I_1 \geq I_2 \geq \Lambda \geq I_{n-1}$

The following five equations（3.11）~（3.15）， Find the value of T.

$$T = \arg\min \delta_w^2 \quad （11）$$

$$\delta_w^2 = q_H\delta_H^2 + q_L\delta_H, q_H = T, q_L = N - T - 1 \quad (12)$$

$$\mu_H = \frac{1}{q_H}\sum_{i=1}^{T} I_i, \mu_L = \frac{1}{q_L}\sum_{T+1}^{N-1} I_i \quad （13）$$

$$\delta_H^2 = \frac{1}{q_H}\sum_{i=1}^{T}[I_i - \mu_H]^2 \quad （14）$$

$$\delta_L^2 = \frac{1}{q_L}\sum_{T+1}^{N-1}[I_i - \mu_L]^2 \quad （15）$$

If the interval between two consecutive frames is $D(k,k+1) \leq I_T$， Then from here to start another division.

Finally, the video stream is divided into k segments, such as $part_1, part_2, \Lambda, part_k$.

Step 3: Keyframe extraction

The above steps mainly complete the division of the initial class, and then complete the core algorithm based on the mutual information extraction key frame.

（1）$part_1$ as the first initial class D_1， the mutual information is equal to the value of $\overline{M_1}$ 。

（2）the calculation of the mutual information

First determine whether A> B is established， if k=k+1， the new initial class is divided；Otherwise, $part_t$ is merged into D_k , Then calculate $\overline{M_k}$.

（3）when t<K ,repeat step(2)。

（4）to determine all the D_k ,take the mutual information similarity to the mean $\overline{M_k}$ of the corresponding frame as a candidate key frame.

Step 4: Fix the results

The method of extracting the key frame there is redundancy, in order to make the shot content is more concise, need to modify the results, the following are two amendments to the principle:

（1）because in the same shot, the image content is similar, so for some of the shorter key frame, can only take one of the frames.

If the amount of mutual information between adjacent frames is less than A, the difference between these two frames is very large, and these two frames are taken as key frames. Otherwise one should be discarded[21].

5. Comparison and Analysis

Experimental environment: Visual Studio2010 and MFC develop kit.

Data resources: the presented data come from the tunnel intelligent monitoring system in Wuhan, Hubei including Fruit Lake, Deposited, Hankou Railway Station and Bayi Road.

The video 1 records 291 frames from 2 to 2:30 p.m. in monitor video. We divide them into three child shot through Sliding window segmentation algorithm and then extract the keyframe using mentioned 5 algorithm. Picture 5 shows the results.

a Frame average method to extract key frames

b The Histogram Averaging Method Extracts Key Frames

c Analyze the shot edge to extract keyframes

d The Visual Content Matching Extracts Key Frames

e Key Frame Extraction Base On Mutual Information

Figure 4 Five key frame extraction algorithms

a threshold T=0.23 Key Frame Extraction Base On Mutual Information

b threshold T=0.6 Key Frame Extraction Base On Mutual Information

Figure 5 Threshold Selection Based on Mutual Information Method

There are a red car and two black cars in Video 1. Picture 4.a and 4.d both fail to extract the important information, picture 4.b and 4.c present one car respectively and picture 4.e shows the three cars. When video has less contents, more keyframes have to be extracted for semantic analysis and the keyframes extracted by the first three algorithms is not enough to meet the requirements. The video content matches method only gets one keyframe and misses some important contents. The mutual information method presents the correct result with three cars captured. The experiment results indicate that the keyframe extracted by mutual information method is reliable and reasonable, and reflects the core contents in the video despite some redundant contents are attached. It is noted that the threshold has a significant impact on the final result as the picture 5 shows. The experiments suggest that roughly 0.23 is relatively reasonable[22].

More experiments are performed and analyzed in order to be more persuasive. Table 1 shows the experiment results.

table1 Monitor the video data sheet

video	Test the number of key frames	Actual number of key frames	Mutual information method to extract the number of key frames	duration /s
Shuiguo lake tunnel video 1	6451	160	206	258

Shuiguo lake tunnel video 2	5751	127	156	230
Shuiguo lake tunnel video 3	2441	19	24	81
Shuiguo lake tunnel video 4	15598	390	489	634
Shuiguo lake tunnel video 5	8931	234	302	357
Shuiguo lake tunnel video 6	3543	74	97	142
Shuiguo lake tunnel video 7	5542	148	182	222

The target video of this paper adopts random sampling method to select the video data set from Wuhan tunnel network, and uses social labeling method to determine the truth value.

$$\text{recall} = \frac{\text{Detect the actual number of key frames}}{\text{The actual number of key frames}} \times 100\%$$

$$\text{precision} = \frac{\text{Detect the actual number of key frames}}{\text{The total number key frames}} \times 100\%$$

The results of several key frame extraction methods for video lake monitoring video 1 are compared, as shown in Table 2 and Figure 6:

Table 2 Monitor video key frame extraction method

Key frame extraction method	recall ratio	precision ratio
Frame averaging method	78.6%	48%
Histogram averaging method	84.1%	68.6%
shot boundary method	85.4%	66.7%
Through video content matching	89.4%	70.3%
Mutual information method	89.7%	69.9%

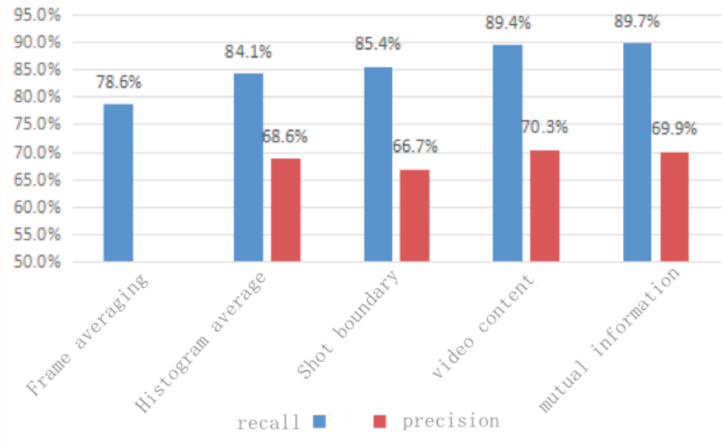

Figure 6 Monitor video histogram

The above analysis proves that the video content matches method and the mutual information method basically fulfill the requirements with a validity and accuracy[23-25]. The two method successfully extract the key contents and remove unnecessary fragments. Meanwhile, the other two method based on frame images give a lower precision ratio with some useless keyframes presented. They are suitable for extracting rough keyframes from a large amount of data since they are easy to understand and operate. The rest shot edge analysis is easily affected by the external factors and lacks of effective features especially with flash video, moved objects or cameras, and so on. Compared with the five method, they should be thoughtfully adopted according to video type and scenario with respective advantages and disadvantages in order to achieve the best efficiency and performance.

6. Conclusion

The paper compares five mainstream keyframe extraction methods and the mutual information method that takes advantage in terms of validity and accuracy is primarily focused. It shows that the keyframe extraction has been mature and the basic semantic analysis approach needs to be enhanced, such as how to recognize the contents of the video. Higher level semantics can be recognized by human like object, behavior, scenarios and events, while artificial intelligence helps to deal with lower level semantics like shape, texture, color. Artificial intelligence has the ability to handle some specific object and scenarios on the basis of pattern recognition and machine learning. However, it still faces challenges in identifying behavior and events, which are the main research fields in the future.

References

[1] Rong Pan， Yumin Tian; Zhong Wang . Key-frame Extraction Based on Clustering [C]. The 2010 IEEE International Conference on Progress in Informatics and Computing, 2010.
[2] Qu Z, Lin L, Gao T, et al. An Improved Keyframe Extraction Method Based on HSV Colour Space[J]. Journal of Software, 2013, 8(7).
[3] Sudeep D. Thepade, Pritam H. Patil. Novel visual content summarization in videos using keyframe extraction with Thepade's Sorted Ternary Block truncation Coding and Assorted similarity measures[C]. Communication, Information & Computing Technology (ICCICT), 2015 International Conference on. IEEE, 2015:1-5.
[4] Huang L, Ye C H. The research and implementation of keyframe extracion methods in content-based teaching video retrieval[C]. Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. IEEE, 2012:2179-2182.
[5] A. Wang. Vehicle Detection and Tracking using the Optical Flow and Background Subtraction [J]. china Railway Science . 2014,35 (6) :131-137
[6] Wang X, Sun S, Ding X. A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition, J. Vis. Commun. Image R. 2015: 193 – 202.
[7] Thakre K S, Rajurkar A M, Manthalkar R R. Video Partitioning and Secured Keyframe Extraction of MPEG Video[J]. Procedia Computer Science, 2016,78:790-798.
[8] Deng-yin ZHANG,Min YANG.Shot-Segmentation-Based Video Watermarking Algorithm Using Motion Vector. Proceedings of 2014 International Conference on Computer,Network Security and Communication Engineering(CNSCE 2014)
[9] A Sengupta,DM Thounaojam,KM Singh,S Roy. Video shot boundary detection: A review. IEEE International Conference on Electrical , 2015 :1-6
[10] Zhong xian. A Dissertation Submitted in Partial Fulfillment of Requirements for the Degree of Doctor Philosophy in Engineering [D]. Huazhong University of Science and Technology, 2013.
[11] Zhong xian, Yang guang, Lu yansheng. Method Of Key Frames Extraction Base on Double-threshold Values Sliding Window Sub-shot Segmentation and Fully Connected Graph [J].Computer S, 2016, 43(6):289-293.
[12] Pal G, Rudrapaul D, Acharjee S, et al. Video Shot Boundary Detection: A Review[M]. Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Springer International Publishing, 2015:119-127.
[13] BH Shekar， KP Uma. Kirsch Directional Derivatives Based Shot Boundary Detection: An Efficient and Accurate Method. Procedia Computer Science ,2015 , 58 :565-571
[14] Dhagdi S T, Deshmukh P R. New Technique for Keyframe Extraction Using Block Based Histogram[J]. International Journal of Advanced Research in Computer Science, 2012,3(3):794-801.
[15] Zhang Q, Zhang S, Zhou D. Keyframe Extraction from Human Motion Capture Data Based on a Multiple Population Genetic Algorithm[J]. Symmetry, 2014, 6(4):926-937.
[16] Sainui J, Sugiyama M. Minimum dependency key frames selection via quadratic mutual information[C]Tenth International Conference on Digital Information Management. 2015.
[17] SheenaC V. Author links open the author workspace.N.K.Narayanan. Key-frame Extraction by

Analysis of Histograms of Video Frames Using Statistical Methods. Procedia Computer Science.Volume 70, 2015, 36-40

[18] JL Lai , Y Yi. Key frame extraction based on visual attention model. Journal of Visual Communication and Image Representation. 2012, 114-125

[19] Chen B W, Bharanitharan K, Wang J C, et al. Novel Mutual Information Analysis of Attentive Motion Entropy Algorithm for Sports Video Summarization[J]. Lecture Notes in Electrical Engineering, 2014, 260(2):1031-1042.

[20] Zhang Jun,Honghua Tan. Key-Frame Extraction Method from Single-Lens Undersea Video Sequences. Applied Mechanics and Materials, 2013, 514-519

[21] M Chatzigiorgaki,AN Skodras, Real-time keyframe extraction towards video content identificationInternational Conference on Digital Signal Processing. 2009 :934-939

[22] G Liu,X Wen,W Zheng,P He. Shot Boundary Detection and Keyframe Extraction Based on Scale Invariant Feature Transform. Eigth IEEE/ACIS International Conference on Computer & Information Science, 2009 :1126-1130

[23] C Sujatha,U Mudenagudi. A Study on Keyframe Extraction Methods for Video Summary. International Conference on Computational Intelligence & Communication Networks. 2011 :73-77.

[24] N Lv , Z Feng , J Peng. Mutual information based video shot boundary detection. International Conference on Image Analysis & Signal Processing, 2013, 20: 1-5

[25] L Krulikovská,M Mardiak,J Pavlovic,J Polec. Video Analysis Based on Mutual Information. Computer Vision & Graphics-international Conference, 201:6375:73-80.