An Improved Keyframe Extraction Method Based on HSV Colour Space

Zhong Qu
College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China
Email: quzhong@cqupt.edu.cn

Lidan Lin, Tengfei Gao and Yongkun Wang
College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China
Email: lldvictor@163.com, gaotengfei1840@163.com, cq789@hotmail.com

***Abstract*—Video segmentation and keyframe extraction are the basis of Content-based Video Retrieval (CBVR), in which keyframe selection is at the very core of CBVR. At shot level, key-frame extraction provides sufficient indexing and browsing of large video databases. In this paper, we proposed two improved approaches of key-frame extraction for video summarization. In our first synthesis method based on histogram-based method and pixel-based method, videos were firstly segmented into shots according to video content, by our improved histogram-based method, with the use of histogram intersection and nonuniform partitioning and weighting. Then, the obtained results are secondly detected to optimize the results. On the other hand, we realized an improved clustering algorithm for video shot segmentation, in consideration of video characteristics. Within each shot, key-frames were determined with the calculation of image entropy of every frame in HSV colour space. Our simulation results in section 4 prove that extracted key frames with our method are compact and faithful to the original video. Moreover, according to the types of test videos, different methods for shot segmentation are highly recommended.**

***Index Terms*—Content-based Video Retrieval, Video Summary, Video Shot Detection, Keyframe Extraction, Histogram-based Method, Clustering-based Method, Image Entropy**

I. INTRODUCTION

With the development of multimedia technology, video data are highly increased, and Content-based Video Retrieval [1,2] is widely used. Key-frame extraction is the foundation of video content analysis and Content-based Video Retrieval. Key-frames of one shot could reflect the main content of the shot, under the circumstances of large video data and limited memory size, storing key-frames would compress redundant video data. Therefore, in the field of video retrieval, the problem of key-frame extraction is of great importance.

Video summarization based on key-frame is an effective method of video indexing and retrieval, presenting users with the selected key-frames as video summary [3], in order to help them understand and perceive video basic content. From the visual point of view, video summarization based on key-frames is more intuitive and could express the content of test video better than traditional video summarization in the form of title. Also, video summarization based on key-frames is the basis of other static video summarization, and is of reference value for dynamic thumbnail video. Therefore, for years it has been permanently focused on from the field of video content analysis [4].

The rest of the paper is organized as follows: Section 2 presents the former related work. Video shot detection and inter-shot key-frame extraction are detailed in section 3. Section 4 presents some opposite experimental results. Section 5 acts as the conclusion of the whole paper.

II. RELATED WORK

Several mechanisms have been proposed to extract key-frames [5]. Yeung et al. (1995) introduces tolerance-band into key-frame extraction. The method regards the first frame as the first key-frame, then for the following frames, the distance between every following frame and the last key-frame is compared with certain threshold value and intended for determining whether it's the new key-frame. One of the possible approaches by Zhang and Hanjalic et al. (1999) for key-frame selection is to divide the consecutive sequence of frames into M clusters (M is pre-determined), and select the frames nearest to the corresponding cluster centers as the key-frames, while the method is of high time complexity [6].

The drawback to most of these approaches is that most of them over-depend on the semantic structure of specific video and parameter settings, such as threshold, it's also difficult to set a suitable interval of time, or frames. Moreover, high time and computation complexity are also concerns.

III. SHOT DETECTION

doi:10.4304/jsw.8.7.1751-1758

Shot is defined as a group of correlative continuous frames shot continuously by a camera shot, which is used for showing a group of motion continuous in time or space. The ideal video shot detection is a process of semantic analysis, but as the current algorithms couldn't analyze video semantics well, most algorithms segment shots based on the low-level features [7] of the position of video shot transition(such as colour, contour, texture, roughness, etc.). Generally, shot transitions will lead to obvious changes of low-level features of video content, abrupt changes [8,9] of colour distribution for example. However, in some special cases, for instance, for video transitions of gradual changes [9] (a generic tem of a variety of shot changes, such as fade in, fade off, dissolve, etc.), low-level features changes are gradual and unobvious, usually lasting several frames. In addition, in one shot, video content changes and noise may cause large changes of low-level features. As a result, under extreme conditions of fast motion in a video and radical ray changes and in the process of gradual change, many algorithms are far from satisfactory.

The most-used method of shot detection is to calculate the frame differences of low-level features between every two continuous frames, then compare the differences with the preset or self-adaptive threshold [10] thus the best combination of frame differences and threshold is the key to video shot detection.

A. Histogram Quantizing based on HSV Colour Space

Generally, it is the three colour values of RGB space that are directly obtained from images, while there are some disadvantages for the RGB space. For instance, different images are possible to have the same colour histogram, while the improvement of histogram partitioning is complicated, and it's also difficult to determine the block size.

Compared to the RGB colour space model, the HSV space model could embody the location of an image better. Moreover, if we computer histogram with properly quantized HSV space, the amount of calculation would be lessened greatly. Thus, with the view of saving memory space and decrease computation complexity, HSV space is nonuniformly quantized [11] according to human colour perception, dividing the hue space into 8 parts, saturation space and value space both 3 parts. And a one-dimensional eigenvector is introduced to simplify the colour feature according to the quantizing levels of H, S and V elements.

$$G = H Q_S Q_V + S Q_V + V \tag{1}$$

Where Q_S and Q_V denote the quantizing levels of element S and V. Here $Q_S = 3$, $Q_V = 3$, then the above formula is indicated as follows:

$$G = 9H + 3S + V \tag{2}$$

The three elements of H, S and V distribute in a one-dimensional eigenvector. In terms of the above formula, $G \in [0,...,71]$, which means obtaining a one-dimensional histogram of 72 bins through the calculation of G.

B. Improved Methods for Video Shot Detection

First, our synthesis method based on histogram-based method and pixel-based method is introduced as below.

Histogram-based methods are the most common-used methods to calculate frame differences. Since color histograms take no account of the location information of given pixels, histogram-based methods are not sensitive to the motion of objects/camera.

There are many applications of histograms. Here we follow the form of frame histogram intersection [8], which could be computed in many different ways.

Since color histograms only record the amount of colour information, without consideration of the location information of given pixels, images with similar color histograms can have dramatically different appearances. Therefore, we improve our histogram-based method with the preprocessing of nonuniform partitioning and weighting, hence the obtained results are more relevant to human visual cognition than the traditional global histogram-based method.

Moreover, in order to improve the recall and precision [12] of shot detection, after detecting shots with our improved histogram-based method, the results are filtered by pixel-based method, which could further decrease the miss detection that have possibly been caused by histogram-based method.

Also, we realized an improved clustering algorithm comparatively. The fundamental idea of traditional clustering algorithm [13] is as below: Start from an initial cluster, and distribute each element of the sample set among some cluster, in order to meet the requirements of the system or user. However, as the traditional method compares one frame with every known shot and takes the frame into the most similar shot, the traditional clustering tends to discontinuous frame numbers and false drop.

Therefore, we improved the traditional clustering algorithm with the following measures: When a new shot appeared, the shots that are already segmented wouldn't be taken into account. With regards to this, we introduced a reference Boolean variance *access* into the traditional clustering algorithm. It means shots are all already segmented when *access* equals to 0, while shots are not yet segmented when *access* equals to 1. Also, we apply the HSV colour space, and compute the similarity between the undetected frame and the last shot by use of the weighting in the three colour channels of H, S and V.

For the video sequence $V = \{f_1, f_2, ..., f_n\}$, we projected it into the HSV colour space, and nonuniformly quantized the H, S and V element, thus the histograms could be established as follows:

$$\begin{cases} H(i) = \dfrac{H\_fol(i)}{M \times N} \\ S(j) = \dfrac{S\_fol(j)}{M \times N} \\ V(k) = \dfrac{V\_fol(k)}{M \times N} \end{cases} \tag{3}$$

Where $H\_fol(i)$, $S\_fol(j)$, $V\_fol(k)$ stand for the number of pixels whose pixel values of H, S and V

element fall into the i, j and k th section, and M, N are the number of pixels in the horizontal and vertical axis. Here, $i \in [1,8]; j \in [1,3]; k \in [1,3]$. Therefore, the histogram of the HSV space $H(i, j, k)$ is a 3-dimensional array, corresponding to the histograms of the three elements H, S and V. Then the similarity between the current frame and shots on the H, S and V elements are defined as follows:

$$\begin{cases} S_H(f, Shot) = \sum_{i=1}^{8} \min(H(i), Shot\_H(i)) \\ S_S(f, Shot) = \sum_{j=1}^{3} \min(S(j), Shot\_S(j)) \\ S_V(f, Shot) = \sum_{k=1}^{3} \min(V(k), Shot\_V(k)) \end{cases} \quad (4)$$

Therefore, our improved clustering algorithm could be described as follows:

(1) Take the first frame f_1 as the first shot, with itself being the inter-class center, and for the shot, $Shot.access \equiv 1$.

(2) Extract the next frame f_{next}, and compute the similarity between f_{next} and the shot based on the above formulas:

$$\begin{aligned} S(f, Shot) = (&\alpha \times S_H(f, Shot) \\ &+ \beta \times S_S(f, Shot) \\ &+ \chi \times S_V(f, Shot))/3 \end{aligned} \quad (5)$$

Where α, β, r are the weighting coefficients of H, S and V elements. Generally, human vision is most sensitive to the H element, $\alpha \geq \beta, \alpha \geq r$. Here they are given the values of 0.9, 0.3, 0.1, corresponding to the weighting proportion of the three elements in the HSV space. The clustering shot must meet the criteria of $Shot.access \equiv 1$.

(3) if $S(f, shot)$ is greater than the threshold of shot segmentation T, f_{next} is considered to belong to $Shot$. Then f_{next} is classified into $Shot$, and the inter-class center of the shot is recomputed as:

$$Shot = \frac{f_{next} + \sum_{i=1}^{Shot.len} f_i}{Shot.len + 1} \quad (6)$$

$$Shot.len = Shot.len + 1 \quad (7)$$

Where f_i is one known frame of the shot.

Else if $S(f, shot)$ is less than the threshold of shot segmentation T, f_{next} is considered not to belong to $Shot$, hence reconstruct a new shot, only containing f_{next}, and also take f_{next} as the inter-class center, $Shot.access \equiv 1$. Meanwhile, for the last shot, $Shot.access \equiv 0$.

(4) If the video hasn't been finished processing, turn to (2), otherwise the process is over.

From the description of our improved clustering algorithm, we can see that our improved algorithm avoid the problem of discontinuous frame in shots and false drop efficiently. And our following experiments also show that our algorithm could get better results.

Further, for the situation of violent illumination change especially flashlight [10], since it usually only sustain several frames, and in terms of the function of human visual persistence for the visual media such as animation, movie, etc, the shots whose frames are lower than 20 are reclassified into the former shots, which also corresponds to characteristics of human vision.

IV. IMPROVED ENTROPY METHOD FOR KEY-FRAME EXTRACTION

The present algorithms of key-frame extraction are divided into 3 classes: first, extract key-frames according to shot boundaries. In the method, take the first frame (or the first frame and last frame) as the key-frame of the shot, so it is easy to implement, but the result is not reliable; second, extract key-frames according to the distances between every adjacent two frames. If one distance is above the threshold preset by users, then it's considered that there appeared a new key-frame; third, extract key-frames according to motion analysis. Calculate the optical flow by Horn-Schunck algorithm, sum up the modulus of optical flow elements of every pixel, and regard the sum as the motion quantity of the k th frame $M(k)$, that is

$$M(k) = \sum_i \sum_j |O_x(i, j, k)| + |O_y(i, j, k)| \quad (8)$$

Then find out the local minimum of $M(k)$, and the frame with the local minimum is considered as a key-frame.

Analyze the above-mentioned three kinds of methods: since in many cases, the first or last frame is not the key-frame, the main content of video shots couldn't be obtained efficiently with the first method; in the second method, the only criteria is the distances between every adjacent two frames, so in many cases, some real key-frames are missed, especially for gradual changes; while in the third method, the computation complexity is high, and the computed local minimum is not always accurate.

After video shot segmentation, the next step is to extract key-frames from segmented shots, which is used for describing main contents of shots. Key-frame (also known as representative frame) is the key image frame used for describing a shot, which always reflect the main content of a shot. Key-frame reduces the data size of video index greatly, and provides an organization structure for video retrieval and browsing. Due to its characteristics, we should "put quantity before quality" for key-frame extraction.

While common-used key-frame extraction methods are referred to in section 2, here we introduce the image entropy into key-frame extraction [14]. Information entropy is first proposed by Shannon in 1948, which is a probability density function of random variable. For an image, its histogram is always considered as the probability density function. Suppose h_k denotes the

proportion of pixels with the value k in the whole image. In addition, a constraint is plus: if $h_k=0$, then $\log h_k=0$. Thus, the image entropy could be formulated as

$$E(f_i)=-\sum_{k=1}^{n} h_k \log h_k \tag{9}$$

In our case with HSV colour space, as the hue space is divided into 8 parts, saturation space and value space both 3 parts, then for the three separate elements, according to the above formula, the histogram feature drops from n dimension to one dimension. The whole image entropy is always computed by the following formula:

$$\begin{aligned} E(f_i) = {} & \alpha \times E(f_i)\_H \\ & + \beta \times E(f_i)\_S \\ & + \gamma \times E(f_i)\_V \end{aligned} \tag{10}$$

Here, as human are most sensitive to the H element, we set $\alpha \geq \beta$, $\alpha \geq \gamma$. In our case, corresponding to the quantizing scales of HSV colour space as is referred to in section 3.1, we suppose α, β, γ at the rate of 9:3:1.

Finally, regard the frame with the largest image entropy in one shot as the key-frame. In this case, the extracted key-frames could stand for the substantive information of shots more efficiently.

V. SIMULATION EXPERIMENTS AND ANALYSIS

Considering the generality of video materials, we select 4 types of 5 videos as test videos, which are advertisement, news, music and preview videos. Generally, the two parameters of recall and precision[12] are used to measure algorithms of video shot detection.

Here, in order to assess the improvement of our methods compared with traditional methods such as histogram-based method and pixel-based method, we also do the tests based on those methods. The contrast is as follows:

TABLE I.
THE RESULTS OF VIDEO SHOT DETECTION

Video name	Video shot detection methods	Shot	Detected number	False drop	Recall	Precision
[CM] Beelzebub ED	Synthesis method	13	13	1	92.9%	92.9%
	Histogram-based method	13	18	5	100%	78.3%
	Pixel-based method	13	31	19	96.9%	62.0%
	Improved clustering method	13	18	6	94.7%	75.0%
[CM]Innisfree cm	Synthesis method	13	13	0	100%	100%
	Histogram-based method	13	13	0	100%	100%
	Pixel-based method	13	13	0	100%	100%
	Improved clustering method	13	8	0	61.5%	100.0%
[News]Cctv_news	Synthesis method	13	15	2	100%	86.7%
	Histogram-based method	13	15	2	100%	86.7%
	Pixel-based method	13	18	5	100%	72.2%
	Improved clustering method	13	16	3	100%	81.3%
[Preview]Anime 10th anniversary	Synthesis method	4	4	0	100%	100%
	Histogram-based method	4	4	0	100%	100%
	Pixel-based method	4	5	1	100%	83.3%
	Improved clustering method	4	4	0	100%	100%
[MV]Taiyou no Uta_clip	Synthesis method	39	33	0	84.6%	100%
	Histogram-based method	39	40	4	93.0%	90.9%
	Pixel-based method	39	39	8	83.0%	83.0%
	Improved clustering method	39	36	4	83.7%	90.0%

From the statistics of the table 1, we can see that the detection precision with our synthesis method (based on histogram-based method and pixel-based method) is higher than the two traditional methods. However, the detection recall is subject to the results of the two methods respectively. While our improved clustering method, compared to the traditional clustering method, also improved the precision to a certain degree. Take the last music video “Taiyou no Uta_clip” for instruction, as there are plenty of abrupt scene changes, some inter-shot illumination changes, abrupt shot moves and gradual changes (suppose the frames before, in and after one gradual change belong to different shots), there exist miss detections to different degrees with all the methods.

However, our two improved method is of simple principle and not difficult to realize, moreover it is of low computational complexity, and enhance the detection precision to a certain degree without increasing the time and computational complexity greatly.

In addition, we also do tests for the situation of continuous shooting, which is supposed to be one type of gradual change that there is only one shot in these videos. Such a video will be exemplified by our two improved methods as follows:

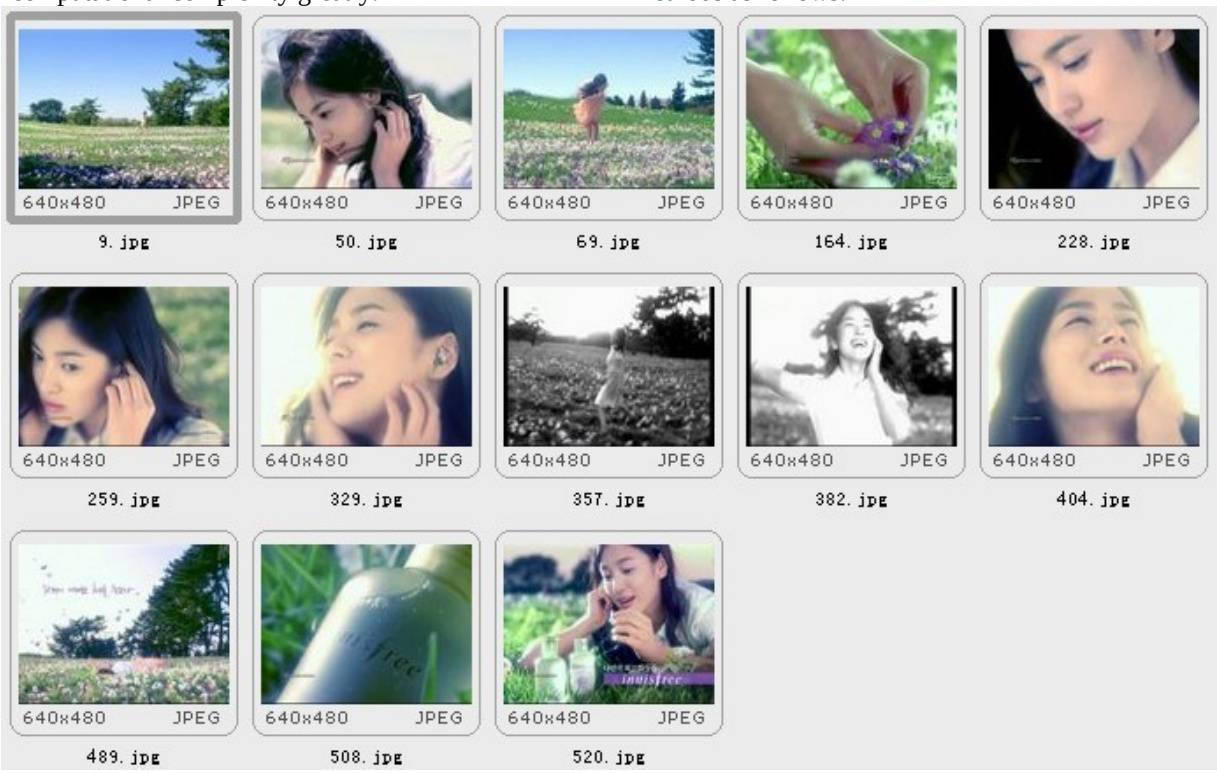

Figure 1. Extracted frames with our method

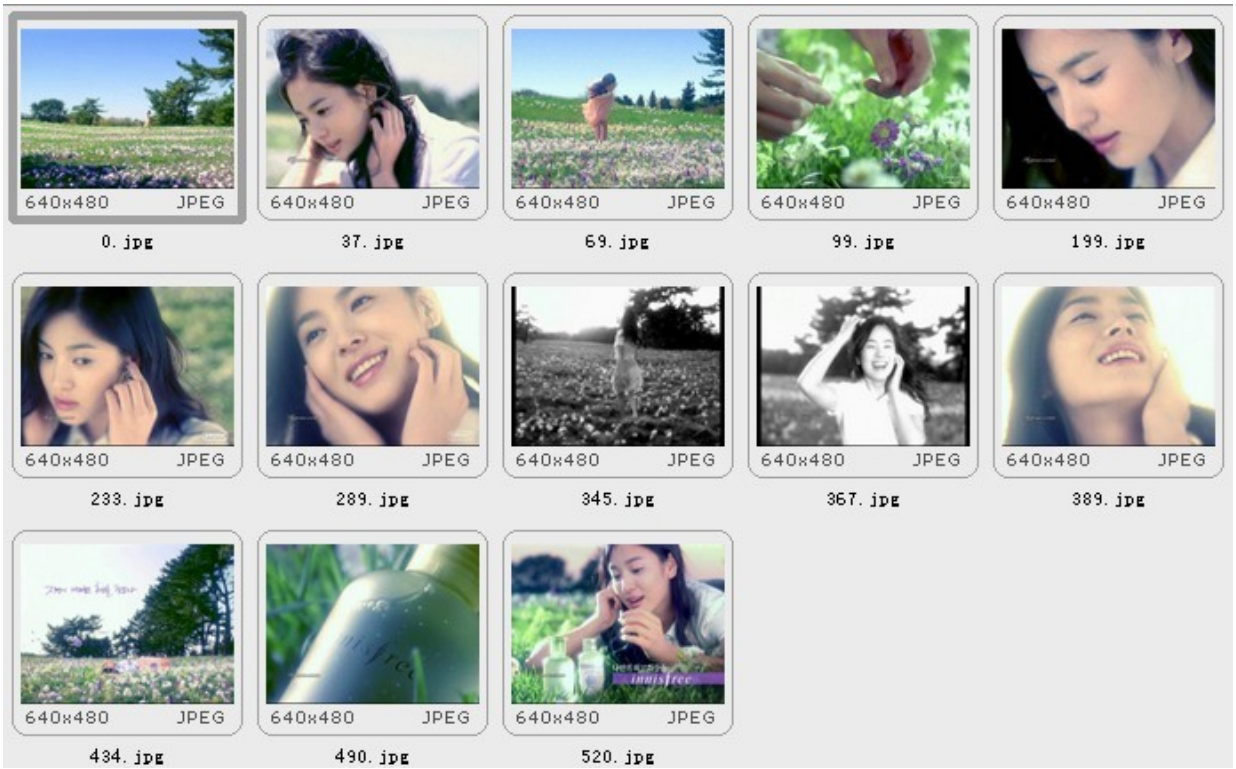

Figure 2. Extracted frames of the first frame

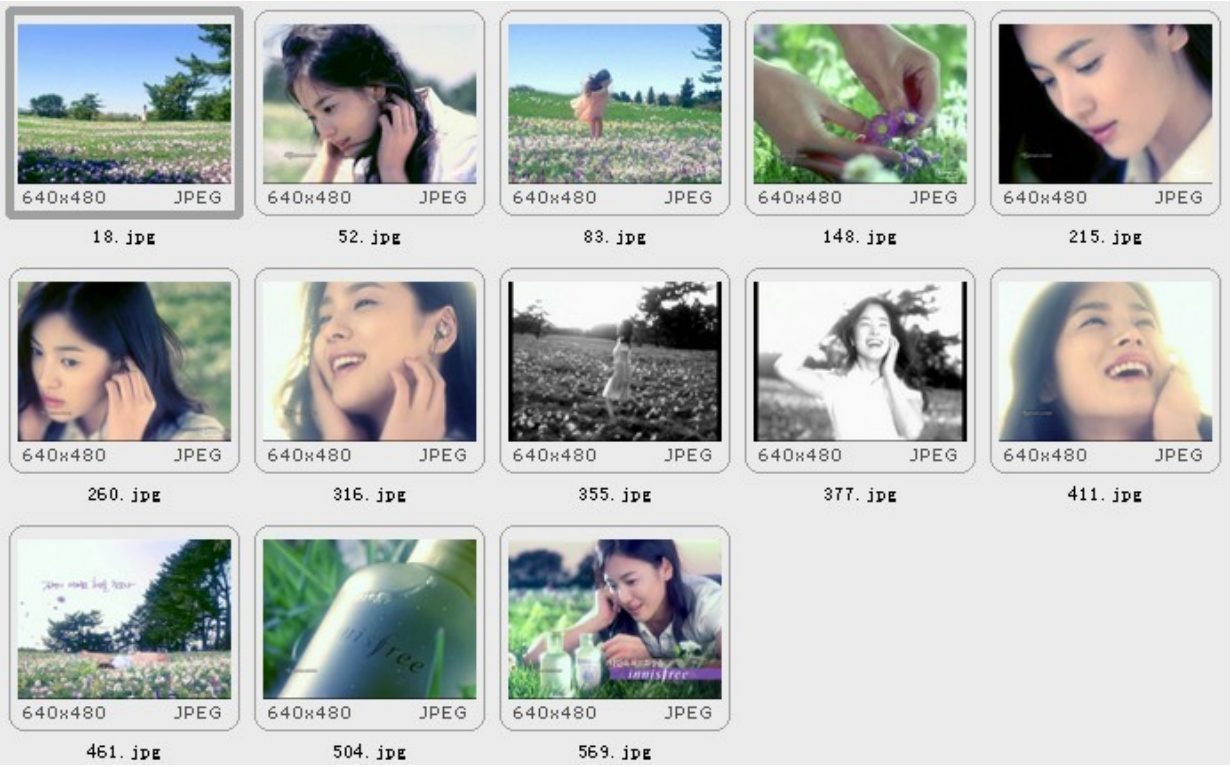

Figure 3. Extracted frames of the middle frame

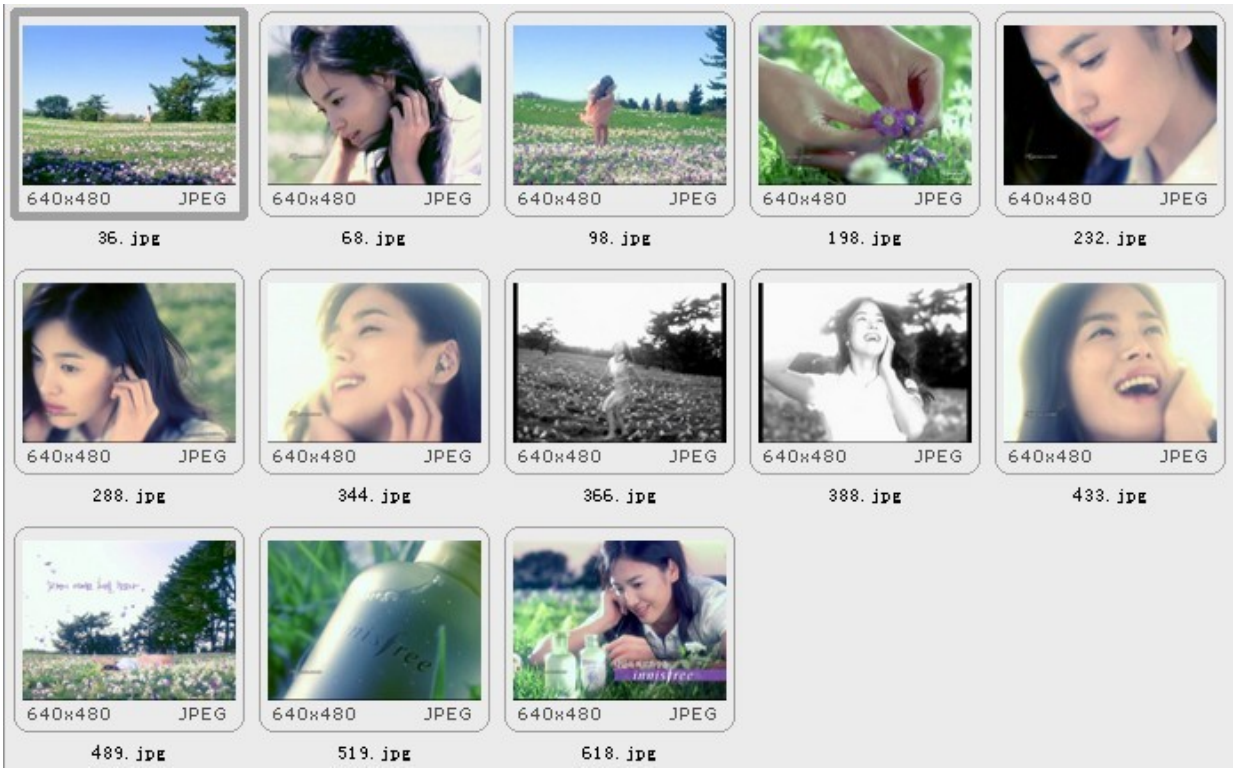

Figure 4. Extracted frames of the last frame

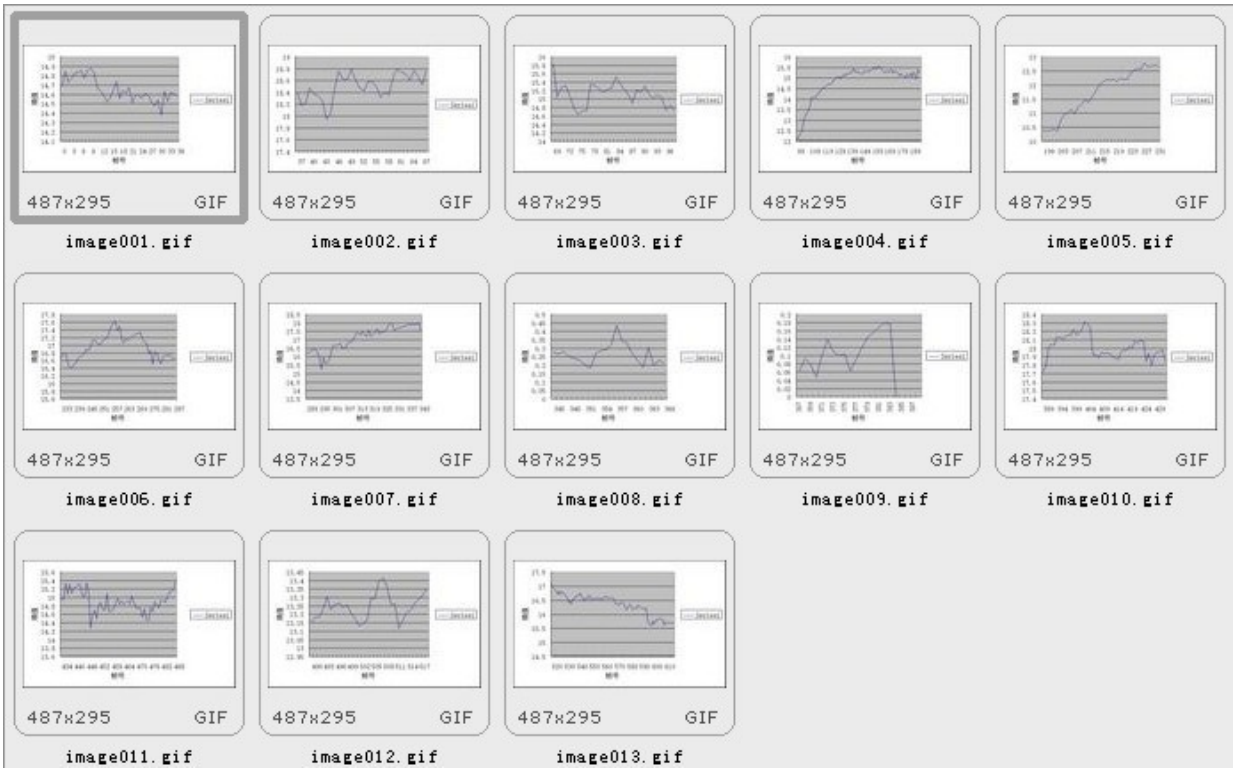

Figure 5. The entropy-change curves of frames within every segmented shot

Table II.
The results of shot detection for a section of continuous shooting video

Video name	Video shot detection methods	Shot	Detected number	False drop	Recall	Precision
[continuous shooting] DOLLY_clip	Synthesis method	1	1	0	100.0%	100.0%
	Improved clustering method	1	1	0	100.0%	100.0%

From the above experiment results, we can see that the key-frames extracted with our entropy method could represent the shot content better, and the key-frame distribution is the most uniform of all the methods of key-frame extraction.

Therefore, on the whole, our set of improved key-frame extraction method could meet the demands of video shot detection and key-frame extraction better. As our synthesis method of shot detection is combined with blocked weighting histogram intersection and a second detection of pixel-based methods, also considering the effects of flashlight, the recall and precision is improved greatly though the time computation complexity is still low, and thus promote the accuracy of key-frame extraction. For our improved clustering algorithm, due to improvements to the defects when traditional clustering is applied to shot detection, it could get better results for some kinds of videos.

VI. Conclusions

In this paper, we present improved algorithms of key-frame extraction based on video segmentation. As the pretreatment of key-frame extraction, we improve the traditional histogram algorithm for video shot detection, combined with frame histograms intersection, blocked weighting, etc. Moreover, pixel-based methods are made use of to do a post processing. Experiment results show that compared to the traditional histogram-based method and pixel-based method, our improved algorithm improves the recall and precision of shot detection to a certain degree, and further, extracted key-frames with the entropy method could represent the contents of video shots better.

Acknowledgments

This work is supported by Chongqing Natural Science Foundation under the Grant No. CSTC 2010BB2399 and Science and Technology Project of Chongqing Municipal Education Commission under the Grant No. KJ110528. The authors wish to thank the associate editors and anonymous reviewers for their valuable comments and suggestions on this paper.

References

[1] Jianwu Long, et al., “A Robust Thresholding Algorithm Framework based on Reconstruction and Dimensionality Reduction of the Three Dimensional Histogram”, Journal of Computers, vol. 8, no. 3, pp.645-652, 2013.

[2] Xinming Zhang, et al., "A Fast Image Thresholding Method Based on Chaos Optimization and Recursive Algorithm for Two-Dimensional Tsallis Entropy", *Journal of Computers*, vol. 5, no. 7, pp.1054-1061, 2010.
[3] Xiang Fu, et al., "Local Features Based Image Sequence Retrieval", *Journal of Computers*, vol. 5, no. 7, pp. 987-994, 2010.
[4] Carlos A.B. et al., "Thresholding Images of Historical Documents Using a Tsallis-Entropy Based Algorithm", *Journal of Software*, vol. 3, no. 6, pp.29-36, 2008.
[5] Kelm, P. et al., "Feature-based video key frame extraction for low quality video sequences", *10th Workshop on Image Analysis for Multimedia Interactive Services*, 2009, pp.25-28.
[6] Bo-Wei Chen et al., "A Novel Video Summarization Based on Mining the Story-Structure and Semantic Relations Among Concept Entities", *IEEE Transactions on Multimedia*, vol.11, no.2, pp.295-312, 2009.
[7] Zhao Yaqin et al., "News Video Clip Retrieval Based on Topic Caption Text and Audio Information", *WRI Global Congress on Intelligent Systems*, no.4, pp.477-481, 2009.
[8] Gunal, E.S. et al., "Gradual shot change detection in soccer videos via fractals", *International Conference on Electrical and Electronics Engineering*, 2009, pp.88-92.
[9] Padalkar, Milind G. and Zaveri, Mukesh A., "Dissolve Detection Based Shot Identification Using Singular Value Decomposition", *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, 2009, pp.312-316.
[10] Feng Hong-cai et al., "A Shot Boundary Detection Method Based on Colour Space", *2010 International Conference on E-Business and E-Government*, 2010, pp.1647-1650.
[11] Spahiu, C.S., "A multimedia database server for information storage and querying", *International Multiconference on Computer Science and Information Technology*, 2009, pp.517-523.
[12] Tapu, R. et al., "A scale-space filtering-based shot detection algorithm", *2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel*, 2010, pp.000919-000923.
[13] Wenzhu Xu and Lihong Xu, "A novel shot detection algorithm based on clustering", *2010 2nd International Conference on Education Technology and Computer*, vol.1, 2010, pp.570-572.
[14] Chasanis, V.T. et al., "Scene Detection in Videos Using Shot Clustering and Sequence Alignment", *IEEE Transactions on Multimedia*, vol.11, no.1, pp.89-100, 2009.

Zhong Qu received the Ph.D degree in computer application technology from Chongqing University in 2009, the M.S degree in computer architecture from Chongqing University in 2003. He is currently an associate professor in Chongqing University of Posts and Telecommunications. His research interests are in the areas of digital image processing, digital media technology and cloud computing.

Lidan Lin is Master Degree Candidates of Chongqing University of Posts and Telecommunications in computer application technology. Her main research interest is digital image processing.

Tengfei Gao received the M.S degree in computer application technology from Chongqing University of Posts and Telecommunications. His main research interest is digital image processing.

Yongkun Wang is Master Degree Candidates of Chongqing University of Posts and Telecommunications in computer application technology. His main research interest is digital image processing.