Video Summarization by Employing Visual Saliency in a Sufficient Content Change Method

Naveed Ejaz, Irfan Mehmood, Muhammad Sajjad, and Sung Wook Baik

***Abstract*—The growing number of videos on the internet requires effective management strategies. Video summarization is a method for managing video data which provides succinct versions of the videos for efficient browsing and retrieval. The inter-frame disparity based key frame extraction is a popular scheme for summarizing videos. However, the performance of such schemes is limited by the fact that most of these techniques select the first frame in a shot as key frame which may not be representative of the shot. In this paper, we propose a saliency inspired inter-frame difference based summarization scheme which selects the most important frame of the shot based on color contrast saliency. The preliminary experimental results prove the efficacy of the proposed method.**

***Index Terms*—Video summarization, key frame extraction, image saliency, visual attention model, frame difference measure.**

I. Introduction

Video summarization is an emerging area of research which deals with the generation of small and succinct version of a given full length video. The field of video summarization can facilitate management of ever-increasing online video data [1]. The prime applications of video summaries are effective browsing, indexing, retrieval, and editing of the online videos. Particular for browsing, the summaries can be used to quickly navigate to the desired contents. Moreover, the summaries can assist the users in video searching by quickly indicating the relevance of the searched videos with the desired ones. In the literature of video summarization, the video summaries are either generated in the form of key frames or skims. The summary in the form of key frames is provided as a set of salient frames that can convey the semantics of the video to the user. The skims provide summary as a shorter video compared to the original video. Naturally, the skims are more expressive and powerful as compared to key frames. However, genration and representation of video skims are generally harder in video skims as compared to key frames.

There are quite a few techniques for summarizing videos in the form of key frames. A popular set of techniques called "Sufficient Content Change based methods" uses the difference among frames for extraction of key frames [2]. The difference is computed among two adjacent frames based on index features and if the distance is increased by a certain threshold then a new key frame is declared. The existence of a high difference is usually affiliated with the advent of a new shot representing a new semantic entity. In the literature, the used features include color histogram [3], edge histogram [3], accumulated energy function [4], compressed domain features [5] and MPEG-7 descriptor [6]. A problem with these type of techniques is that when a significant change among frames is detected, then usually the first frame is selected as the key frames. Practically, the new frame may not be a true representative of the shot and there may be a better frame to represent the semantics. There is a need that the slected frame must be the best representative of the videos.

In this paper, we propose an inter-frame difference based key frame extraction scheme which selects the most appropriate frame from a shot instead of selecting the first frame. For selecting appropriate frame, the visual saliency based frame ranking system is used. After a significant difference is established and the shot boundaries are defined, the frames having the highest saliency rankings are selected as key frame. The experimental results indicate the the proposed modification results in the selection of effective key frames.

Manuscript received July 5, 2013; revised September 4, 2013.

The authors are with the Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea (e-mail: naveed@sju.ac.kr,irfanmehmood@sju.ac.kr, Sajjad@sju.ac.kr, sbaik@sejong.ac.kr).

II. Methodology

The proposed methodology determines the frame difference based on color and texture features. For the representation of color features, color histograms in HSV space are used. For capturing texture features, gray-level co-occurence matrix based textures features are employed. The flow charts of the proposed system are shown in Fig. 1. Each step of the proposed methodology is discussed in detail as under.

A. Computation of Saliency Value

The saliency of an image is a practical realization of the concept of visual attention [7]. The human visual attention is a neourobiological concept that covers the ability of human mind to concentrate on 'salient' areas while looking on a scene. The saliency map of an image is those a grayscale image where a high gray scale value at a pixel represents a high salient area and vice versa. For the purpose of finding saliency value, we used color contrast based saliency measure. The contrast value for a particular pixel '*p*' of an image '*I*', for a color channel '*c*' is given as:

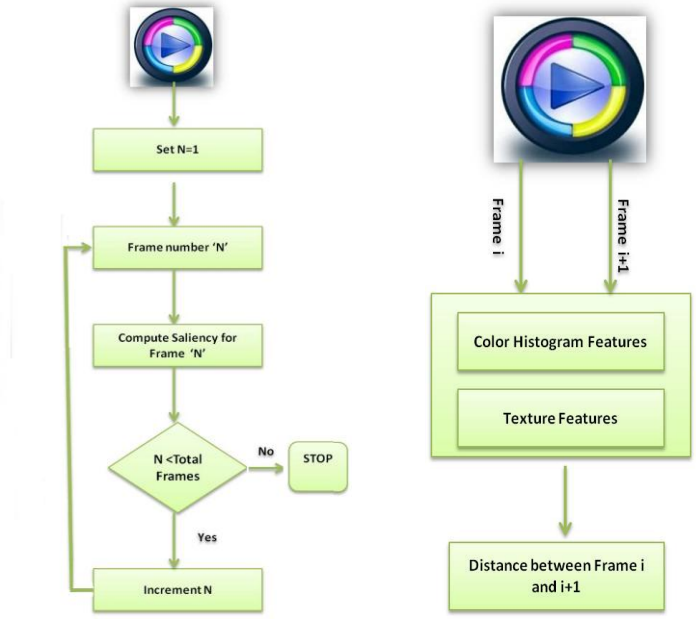
$$C_c(I,p) = \sum_{i \in N(p)} \left\| I_c(p) - I_c(i) \right\| \qquad (1)$$

$$c = red, green, blue$$

where $N(p)$ is a 5×5 neighborhood around pixel 'p'. Next, the final value of contrast at pixel 'p' is achieved by adding contrast values of each individual color channels.

$$C(I,p) = C_{red}(I,p) + C_{green}(I,p) + C_{blue}(I,p) \quad (2)$$

The contrast saliency value can then be computed for each individual pixel of an image. The contrast values are then mapped to the range (0, 1). Finally, the average of all non-zero contrast values is taken which represents the 'Frame Importance Value' (FIV). This concept can be extended to video by calculating FIV for all frames in the video. A frame with high value of FIV indicates high contrast value and thus is assumed to be more significant than others.



(a) (b)

Fig. 1. (a). Flow chart for computing saliency value of each frame, (b) Computing distance among two adjacent frames.

B. Feature Extraction

1) Texture feature extraction

For extraction of texture features, we used gray-level co-occurence matrix (GLCM) [8]. The main theme of GLCM is to calculate the spatial relationship of pixels by enumerating the frequencies of occurrence of pixels in a specified spatial neighborhood. Using the co-occurrence matrix, the four texture parameters contrast, correlation, energy and homogeneity can be computed. For the details of GLCM and computation of texture parameters from it, we refer readers to [8]-[12]. The four texture values combined with color feature vector of size 32 yields a final feature vector of size 36. The feature vector is calculated for each frame in the video and ultimately used for finding difference among frames.

2) Color histogram based features

The color histogram has been developed in HSV (Hue, Saturation, Value) color space because of its relative closer represntation of human perception. Moreover, HSV space separates the color perception from shading and brightness. The quantization step is performed after making histogram and thus assigning 16 bins for Hue component, and 8 bins for each of the saturation and intensity. The quantization step reduces the amount of data to be processed. This makes a feature vector of size 32 for color histogram. All values are mapped in the range 0 to 1.

3) Measuring content change

In the next step, difference value is computed between each individual adjacent frames based on feature vector of dimension 36. For computation of difference, we simply used Euclidean distance between feature vectors. The Euclidean distance returns one value showing difference between neighboring frames. It is assumend that a high distance value indicates a large difference between two frames and vice versa.

4) Extraction of key frames

As a basic ingredient of inter-frame difference based schemes, if the difference between two frames exceed beyond a certain threshold, a new key frame is declared. However while doing so, generally the first frame after the "big difference" is selected as key frame. This is because the emergence of 'big difference' is generally assumed to be the advent of a new shot and the selected frame is exected to reprent a new shot. However, the first frame may not be a representative of a shot. For this reason, we used saliency based frame importance to determine the most important frame in the shot. Each new "big differences" is taken as boundary of the shot. Among each shot, the frame with the highest "FIV" is selected as the key frame.

III. Experiments and Results

In order to evaluate the performance of the proposed scheme, 5 videos form Open Video Project [13] are used. Table I lists some of the properties of these videos.

For evaluation of the proposed method, we utilized a famous subjective evaluation strategy. We asked 7

independent users to rate the generated summaries on criteria of Informativeness and Enjoyability. Informativeness determines the capability of the summary in retaining the maximum content coverage with minimum redundancy. Enjoyability determines the extent to which the users found the experience of watching summaries as perceptually enjoyable. The users were asked to rate the summaries on a scale of 0 to 1.

Table II shows the results of Informativeness and Enjoyability for the 5 videos of the data set. For comparison, we also provided results of the summarization for the same method except that the first key frame among shot is detected. As it can be observed that the proposed change of selecting frames based on saliency has resulted in better results as compared to that of first frame selection in the shot.

TABLE I: DETAILS OF TEST DATA

No.	Video Name	No. of Frames	Genre
1	The Great Web of Water, segment 01	3279	Documentary
2	Drift Ice as a Geologic Agent, segment 05	2187	Documentary
3	Technology at Home: A Digital Personal Scale	3346	Educational
4	The Future of Energy Gases, segment 09	1884	Documentary
5	Introduction to HCIL 2000 reports	2454	Educational

TABLE II: EVALUATION OF THE PROPOSED TECHNIQUE

No.	Selecting First Frame from each Shot		Proposed Scheme	
	Informativeness	Enjoyability	Informativeness	Enjoyability
1	0.73	0.79	0.78	0.84
2	0.81	0.78	0.85	0.86
3	0.75	0.81	0.79	0.89
4	0.69	0.73	0.77	0.82
5	0.83	0.76	0.88	0.90

IV. CONCLUSION

This paper proposes a modification in inter-frame distance based key frame extraction scheme. The selection of first frame after detection of a big change is problematic and may not lead to semantically relevant key frames. The usage of visual attention models and image saliency help in selecting appropriate key frames from the videos. In future, we want to extend this framework by utilizing more sophisticated visual attention models and index features.

ACKNOWLEDGMENT

This work was supported by the Industrial Strategic technology development program, 10041772, The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the Ministry of Science, ICT & Future Planning(MSIP).

REFERENCES

[1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 3, no. 1, pp. 3, 2007.

[2] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *Journal of Visual Communication and Image Representation*, vol. 23, no. 7, pp. 1031-1040, October 2012.

[3] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern recognition*, vol. 30, no. 4, pp. 643-658, 1997.

[4] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng, "Dynamic selection and effective compression of key frames for video abstraction," *Pattern recognition letters*, vol. 24, no. 9, pp. 1523-1532, 2003.

[5] J. Ren, J. Jiang, and Y. Feng, "Activity-driven content adaptation for effective video summarization," *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 930-938, 2010.

[6] J.-H. Lee, G.-G. Lee, and W.-Y. Kim, "Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder," *IEEE Trans. on Consumer Electronics*, vol. 49, no. 3, pp. 742-749, 2003.

[7] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 34-44, January 2013.

[8] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley, ISBN 0-201-50803-6, 1992, pp. 508-510.

[9] N. Ejaz and S. W. Baik, "Video summarization using a network of radial basis functions," *Multimedia Systems*, vol. 18, no. 6, pp. 483-497, 2012.

[10] I. Mehmood, N. Ejaz, M. Sajjad, and S. W. Baik, "Prioritization of brain MRI Volumes using medical image perception model and tumor region segmentation," *Computers in Biology and Medicine*, vol. 43, no. 10, pp. 1471-1483, October 2013.

[11] M. Sajjad, N. Ejaz, and S. W. Baik, "Multi-kernel based adaptive interpolation for image super-resolution," *Multimedia Tools and Applications*, December 2012.

[12] M. Sajjad, N. Ejaz, I. Mehmood, and S. W. Baik, "Digital image super-resolution using adaptive interpolation based on Gaussian function," *Multimedia Tools and Applications*, July 2013.

[13] G. Marchionini and G. Geisler, "The open video digital library," *D-Lib Magazine*, vol. 8, no. 12, pp. 1082-9873, 2002.

Naveed Ejaz received his MS degree in Computer Software Engineering from National University of Sciences and Technology, Pakistan. He is currently pursuing Ph.D. course in Sejong University, Seoul, Korea. His research interests include digital image and video retrieval, video summarization, and video analytics.

Irfan Mehmood received his BS degree in Computer Science from National University of Computer and Emerging Sciences from Pakistan. He is currently pursuing his MS degree at Sejong University, Seoul, Korea. His research interests include medical video summarization and computer aided diagnostics systems.

Muhammad Sajjad received his MS degree in Computer Software Engineering from National University of Sciences and Technology, Pakistan. He is currently pursuing Ph.D. course in Sejong University, Seoul, Korea. His research interests include image super-resolution and reconstruction, sparse coding, video quality assessment and virtual reality.

Sung Wook Baik is a professor in the College of Electronics and Information Engineering at Sejong University. His research interests include Computer vision, Pattern recognition, Computer game and AI. He has a Ph.D. in Information Technology and Engineering from George Mason University.