A STUDY ON KEYFRAME EXTRACTION METHODS FOR VIDEO SUMMARY

Sujatha C
Department of Computer Science and Engineering
BVBCET, Hubli-580031,Karnataka
Email: sujata_c@bvb.edu

Uma Mudenagudi
Department of Electronics and Communication
BVBCET, Hubli-580031, Karnataka
Email:uma@bvb.edu

***Abstract*—In this paper we carry out a survey on key frame extraction methods for Video Summary. We also discuss the summary evaluation criteria and compare the approaches based on the method, data set and the results. Video Summary is a process of presenting an abstract of entire video within a short period of time. It aims to provide a compact video representation, while preserving the essential activities of the original video. It is an essential task in video analysis and indexing applications. Most of the video summaries are based on selection of key frames within the shots of a video. Many of them use motion features and few use visual features for extracting the keyframes. The video summary quality assessment methods are based more on subjective and less on objective measures. Tongwei Ren et al has provided a framework to assess the quality of the video against a given reference summary using both subjective and objective measures. Ciocca et al used the objective measures for evaluation of summary and most of them evaluate by taking the subjective opinion of experts. A framework for automatic evalution is needed based on both subjective and objective measures without the reference summary.**

Key words: Video Summary, Shot Detection, Key frames, quality measures.

I. INTRODUCTION

In this paper we carry out a study on the recent key frame methods for video summary. We also discuss on the evaluation criteria, since the assessment is the major challenge in video summary. The rapid development of digital video capture and editing technology has led to enormous increase in video data, creating the need for effective techniques for video retrieval and analysis. Video summary is a temporally condensed representation of a video. The purpose of the video summary evolves mainly due to viewing time constraints. It help us to assess the relevance or value of information within a shorter period of time while decision making. It also plays prime role where the resources like storage, communication bandwidth and power are limited. It finds main applications in security, military, and entertainment.

Most of the work in video summarization extracts the key frames within each shot. A Shot is defined as a sequence of frames captured from a single camera operation. Shot Detection in a video sequence is a process of identifying visual discontinuities along the time domain. The basic first step is to segment the video into shots. The first and last frames of each shot are chosen as key frames. And key frames within each shot are extracted based on various models with fixed or variable number of key frames.

One of the most challenging task in video summarization is to evaluate the summary produced by the algorithms. The most common method is to take the subjective opinion from panel of experts/users where they compare the summary with the original sequence of video. And only few approaches use qualitative measures to compare and quantify their results.

In Section 2, we present some recent approaches of key frame extraction for video summary. We describe the various ways to evaluate the summaries in Section 3. The results and observations of different key frame extraction methods are discussed in Section 4. We conclude in Section 5.

II. KEY FRAME EXTRACTION METHODS

Here we discuss the different methods used to extract key frames which are based on the motion patterns, frame descriptors and frame visual features. Also we discuss the data set and the evaluation criteria used in different approaches.

A. Perceived Motion Energy Model (PME)

Tianming Liu et al. [1] proposed a triangle model of perceived motion energy to model motion patterns in video and a scheme to extract key frames based on this model. The PME is a combined metric of motion intensity and the kind of motion with more emphasis on dominant video motion. The motion data is extracted from MPEG video streams. The average magnitude Mag(t) of motion vectors in the entire frame is calculated as:

$$Mag(t) = \frac{\left(\frac{\sum(MixFEN_{i,j}(t)}{N} + \frac{\sum(MixBEn_{i,j}(t)}{N}\right)}{2} \quad (1)$$

where $MixFEN_{i,j}(t)$ represents forward motion vectors and $MixBEn_{i,j}(t)$ represents backward motion vectors. N is number of macro blocks in the frame.

The percentage of dominant motion direction $\alpha(t)$ is defined as:

$$\alpha(t) = \frac{max(AH(t,k), k \in [1,n]}{\sum_n^{k=1} AH(t,k)} \quad (2)$$

$AH(t,k)$ represents the angle histogram with n bins. The PME of a B-frame is computed as $PME(t) = Mag(t).\alpha(t)$. These PME values of the frames are plotted which represent the

978-0-7695-4587-5/11 $26.00 © 2011 IEEE
DOI 10.1109/CICN.2011.15

sequence of motion triangles. The frames at the turning point of the motion acceleration and motion deceleration are selected as key frames. The key frame selection process is threshold free and fast. Here first the video sequence is segmented into shots using twin comparison method [2]. The key frames are selected based on the motion patterns within the shots. For shots having motion pattern the triangle model is used to select the key frame, whereas for shots with no motion pattern, the first frame is chosen as a key frame. The algorithm is tested on 3 hours video sequences of various categories such as home, sports, news and entertainment video. Ten testers from their research lab rated based on their satisfaction to how well the key frames capture the salient content of a shot. The satisfactory rate for sports and entertainment video found to be good as more actions exist when compared to home and news video.

B. Visual frame Descriptors

G. Ciocca et al [3] introduced an algorithm that uses three visual features: color histogram [4], wavelet statistics [5] and edge direction histogram [6] for selection of key frames. Similarity measures are computed for each descriptor and combined to form a frame difference measure. The distance between two color histograms d_H using the intersection measure is given as:

$$d_H(H_t, H_{t+1}) = 1 - \sum_{j=0}^{63} min(H_t(j), H_{t+1}(j) \quad (3)$$

The difference between two edge direction histograms d_D is computed using Euclidean distance as such in the case of two wavelet statistics d_W :

$$d_D(D_t, D_{t+1}) = \sqrt{\sum_{j=0}^{71}(D_t(j) - D_{t+1}(j))^2} \quad (4)$$

$$d_W(W_t, W_{t+1}) = \sqrt{\sum_{j=0}^{19}(W_t(j) - W_{t+1}(j))^2} \quad (5)$$

These differences are combined to form the final frame difference measure d_{HWD} as:

$$d_{HWD} = d_H.d_W + d_W.d_D + d_D.d_H \quad (6)$$

These difference values are used to construct a curve of the cumulative frame differences which describes how visual content of the frames changes over the entire shot. The high curvature points are determined and key frames are extracted by taking the midpoint of two consecutive points. The authors have used three quality measures: Fidelity [7], Shot Reconstruction Degree (SRD)[8] and Compression Ratio (CR) to evaluate the video summary. The Fidelity measure is defined as a semi Hausdorff distance. SRD measure is that using a suitable frame interpolation algorithm, we should be able to reconstruct the whole sequence from the set of key frames. CR is defined as ratio of number of key frames and total number of frames in the video sequence. Five categories of video are tested and compared with other six key frame extraction methods such as adaptive temporal sampling (ATS), flexible rectangles (FR), shot reconstruction degree interpolation (SRDI), midpoint (MP) and perceived motion energy (PME). Overall the algorithm outperforms the other methods. But only for news video FR outperforms than this method.

C. Motion Attention Model

I Cheng Chang et al. [9] detect the shots using color distribution and edge covering ratio [10] that increase the accuracy of shot detection. Key frames are extracted from each shot by using the motion attention model [11]. Here the first and last frame of every shots are considered as key frame and the others are extracted by adopting motion attention model. These key frames are then clustered and a priority value is computed by estimating motion energy and color variation of shots. The motion energy TMA is defined as:

$$TMA(S_i = \sum_{f_j \in S_i} MA(f_j) \times \log(\sum_{f_j \in S_i} MA(f_j)) \quad (7)$$

where $TMA(S_i)$ denotes the sum of motion attention value of shot i. And the energy motion change (EMC) is defined as:

$$EMC(S_i) = TMA(S_i) \times CF(S_i) \quad (8)$$

where $CF(S_i)$ denotes the total number of frames that have significant intensity variation in shot i. The priority value of shot is defined as the following:

$$PV = e^{-(\frac{EMC(S_i)}{\sum_{S_i \in C_j} EMC(S_i)})} \quad (9)$$

The higher motion energy and color variation are the larger priority value of shot will have. A higher PV value means that this shot is more important of this cluster and the shot will be the highlight of cluster. The algorithm is tested on home video which consists of two scenes one playing table-tennis in an indoor space and other playing tennis at outdoor court captured randomly. The method summarized the video by clustering the two scenes. Here the chronological order of frames is not maintained.

D. Multiple Visual Descriptor Features

Chitra A.D et al. [12] used same visual features as Ciocca[13] along with one additional feature, weighted standard deviation. The grayscale image is focused to L-level discrete wavelet decomposition. At each ith level (i=1..L) there are LH,HL,HH detail images and an approximation image at level L. The standard deviation is for all these images are calculated and the weighted standard deviation feature vector is defined as:

$$f = \Big\{\sigma_1^{LH}, \sigma_1^{HL}, \sigma_1^{HH}, \frac{1}{2}\sigma_2^{LH}, \sigma_1^{HL}, ... \frac{1}{2^{L-1}}\sigma_L^{LH}, \frac{1}{2^{L-1}}\sigma_L^{HL}, \frac{1}{2^{L-1}}\sigma_L^{HH}, \frac{1}{2^{L-1}}\sigma^A, \mu^A \quad (10)$$

The key frames are selected by constructing the cumulative graph for the frame difference values. The frames at the sharp slope indicate the significant visual change, hence they are selected and included in the final summary. And the key frames corresponding to the mid points between each pair of consecutive curvature point are considered as representative frames. The algorithm is tested on educational video sequence and compared with the I-frames obtained by CueVideo and found that the method gives better result.

E. Motion focusing

Congcong et al. [14] proposed motion focusing method to extract key frames and generated summary for lane surveillance videos. This method focuses on one constant-speed motion and aligns the video frames by fixing focused motion into a static situation. A summary is generated containing all moving objects and embedded with spatial and motion information. The method begins with background subtraction to extract the moving foreground for each frame. In this method background subtraction is combined with min cut to get a smooth segmentation of foreground objects. A labeling function f labels each pixel i as foreground f_i = 1 or background $f_i = 0$. The labeling problem is solved minimizing the Gibbs energy [15], defined as:

$$E(f) = \sum_{i \in V} E_1(f_i) + \lambda \sum_{(i,j) \in N} E_2(f_i, f_j) \qquad (11)$$

where E_1 and E_2 are defined as:

$$E_1(1) = \begin{cases} 0 & d_i > k_i^1 \\ k_i^1 - d_i & k_i^2 < d_i < k_i^1 \\ \inf & d_i < k_i^2 \end{cases}$$

$$E_1(0) = \begin{cases} 0 & d_i > k_i^3 \\ d_i - k_i^3 & k_i^3 < d_i < k_i^1 \\ \inf & d_i < k_i^1 \end{cases}$$

$E_2 = \delta(f_i - f_j), d_i$ is difference between the current frame and the Gaussian mean for the i^{th} pixel and $k_i^t | t = 1, 2, 3$ are the thresholds for the i^{th} pixel. The key frame extraction and summary image generation is done through two steps of mosaicing. The initial mosaicing is done with the foreground segmentation results. A greedy search method is used to find out the key frames which increase the foreground coverage on the mosaic foreground image most. Then a second time mosaicing is carried on by mosacing the key frames to generate the summarization image. The proposed method is tested on lane surveillance video where the moving speed is almost the same for different objects. The summary not only represents all objects in the focused motion but also provide temporal and spatial relation and a comparison is made with a similar method described by Y. Pritch, et al [16] which cannot provide temporal and spatial relation of input video.

F. Camera Motion and Object Motion

Jiebo Luo et al [17] have proposed a method to extract semantically meaningful key frames from personal video clips. Authors have attempted to select the key frames from personal or consumer video space where the content is unconstrained and lack of pre-imposed structures. Firstly in a psycho visual study, a ground truth collection of key frames from video clips taken by digital cameras were obtained using first and third party judges and a reference database were created. The proposed key frame extraction framework is based on camera motion and object motion. The video is segmented using camera motion-based classes: pan, zoom in, zoom out and fixed. The key frames are selected from each of these segments. For zoom in class the focus is on the end of the motion when the object is closest. In case of pan the selection is based on local motion descriptor and global translation parameters. For a fixed segment the mid frame of the segment or the frame where the object motion is maximum is chosen. Final key frame selections from each of these segments are extracted based on confidence value formulated for the zoom, pan and steady segments. The global confidence function d_{pan} is given as: $d_{pan} = \alpha_1 d_{spat} + \alpha_2 d_{know}$ with $\alpha_1 + \alpha_2 = 1, d_{spat}$ is probability function of the cumulative camera displacements and $d_{know} = G(\mu + \sigma)$ is a Guassian function, with being the location of local minimum and σ the standard deviation computed from the translation curve. The algorithm is evaluated on 18 videos data set with 8 indoor and 10 outdoor videos having variety of scene content and camera motions. The proposed algorithm motion based key-frame extraction (MKFE) was compared with evenly spaced (ESKF) and color histogram based method [18]. The authors observed that all these methods gave similar accuracy for key frames representative of pan motions. But for action and zoom related key frames are significantly better detected by the proposed method (MKFE). The key frames representative of a zoom event are detected with a success rate of 87.5% using MKFE, 37.5% for histogram based and 62.5% by ESKF.

G. Visual Attention Clues

Jiang Penag et al.[19] made an attempt to bridge the semantic gap between low level descriptors used by computer systems and high level concepts perceived by human users. The method represents a new visual attention index (VAI) descriptor based on a visual attention model to bridge the above mentioned gap. With VAI, they extracted key frames that are most aligned with a human's perception both at the shot and clip levels. The attention detection model is comprised into two parts: dynamic and static. In the dynamic attention detection, the optical flow of each frame is calculated to detect the temporal conspicuous areas of video sequence. Motion attention for sample blocks are computed and normalized to yield a dynamic range of [0,255]. Static attention detection module use two contrast-based features: intensity feature (I) and color feature (H) and calculate the static attention value of block. Combining these both attention values, the final visual attention index (VAI) is obtained as: $VAI = w_S.A_S + w_T.A_T$ where w_S is motion priority static and w_T motion weight. A_S is static attention model and A_T is motion attention. For the video summary the frames having the highest VAI from each shots are selected that comprise the most attractive key frames.

The method was tested on documentary, indoor, outdoor, news and sports videos. The highest VAI was obtained for documentary/landscape video. The results were compared with K-Means single value decomposition [20] and information theory [21]. Authors conducted a user study and evaluated the quality of key frames on the basis of subjective rating for three evaluation criterion: informativeness,enjoyability and rank. Assessors assigned scores ranging from 0 to 100 for each of these parameters. The results were compared with time sampled, K-Means with single value decomposition and information theory methods. The proposed method (VAI) significantly outperforms these methods on different types of videos with average score of 87.6,86 and 3.6% for the above mentioned three evaluation parameters.

III. VIDEO SUMMARY EVALUATION CRITERION

In this section we discuss about the evaluation criteria used for video summary. Most of them give the success rate as the ratio of number of key frames extracted to the actual number of key frames obtained. And another most common evaluation of summary relies on the subjective opinion of panel of users where they compare the summary with original sequence. It is time consuming and gives only the perception of the individuals. This leads a need of an automatic evalution technique that can measure the quality of a good summary. Tongwei Ren et al [22] proposed a framework based on 4C criteria: coverage, consiseness, coherence and context [23] for a given reference summary. It provides a complete objective assessment framework that maps the subjective evlaution by human being. G Ciocca et al. used three objective measures: fidelity measure proposed by Chang et al. [7], the shot reconstruction degree proposed by Liu et al. [8] and the compression ratio for evaluating the quality of summary. Jiebo Peng, et al used three subjective measures: in formativeness, enjoy ability and rank where the assessors assigned scores ranging from 0 to 100 for each of these parameters and again this also depends on the individual perceptions.

IV. COMPARISON OF THE METHODS

The key frames extraction methods for video summaries which were discussed above are compared as shown in Table I. A comparison is made with respect to video data set, evalution criteria and results.

PME model uses motion pattern to select the keyframes, hence this method is more suitable for videos were actions are more. In curvature point method three visual feature descriptors: color histogram, wavelet statistics and edge direction are used. This combined metric has led to analyze complex events and select appropriate keyframes. For lane surveillance system the motion focusing method is useful as the vehicles move with similar constant speed. The camera motion and object motion features are more suitable for personal videos as the contents are unconstrained and unstructured. The visual attention clue method uses both the low level descriptors used by computers and high level concepts perceived by human users, hence more semantic keyframes are selected.

TABLE I
COMPARISON OF KEYFRAME EXTRACTION METHODS FOR VIDEO SUMMARY

Keyframe Extraction method.	Video Data Set Tested on	Type of measure	Results
Perceived Motion Energy Model	home, sports, news and entertainment videos with duration of 3 hrs	subjective. 10 testers rated satisfactory	Good for sports and entertainment were action is more.
Visual frame Descriptors	Six videos of categories such as TV series, news, sports, cartoon with video length of 1-8 mins.	Three objective measures Fidelity, Shot reconstruction degree and compression ratio	Compared with five other algorithms, except for cartoon video this method gave good results than others
Motion Attention model	Indoor and outdoor games with 7512 total frames.	subjective	Gives more concise summary for similar color and motion type shots.
Multiple Visual Descriptor Features	Video sequences like news programs, sports, academic with total frames ranging from 173 to 180.	one objective measure: Number of key frames.	Compared with the I frames obtained by Cue Video data set.
Motion focusing	Lane surveillance video with 50 shots	subjective	provides the spatial and temporal relation.
Camera Motion and Object Motion	Eight indoor and ten outdoor games with total number of frames ranging from 194 to 3000.	one objective measure: Number of key frames detected	Compared with three algorithms, for action and zoom related key frames this method is better than others.
Visual Attention Clues	Six videos of documentary, outdoor scene news and sports with 14 to 80 number of shots.	3 subjective measures: in formativeness, enjoy ability and rank	Compared with 3 other methods. 11 assessors assessed by rating the parameters with 0 to 100 scores. This method gave better scores 86-88.

V. CONCLUSION

In this paper we discussed on various key frame extraction methods for video summary and found that most of the methods use motion as one of the important feature and few use visual features. The quality assessment for summary is another important factor that describe the good summary. Both the subjective and objective measures need to be used

to evaluate the summary. Such a framework is proposed by Tongwei Ren *et al* with a given reference summary. Ciocca et al used the objective measures to quantify the results. Most of the video summaries are assessed by the invitee assessors who rate according to their perception and this may lead to bias. Hence, we conclude that a framework is needed to use objective measures that maps the subjective measures without a given reference summary. Many other key frame extraction methods such as inter-shot information, entropy and discrete contour evolution also exists, which may be used for summarizing the video to give a better summary.

REFERENCES

[1] T. Liu, H. J. Zhang, and F. Qi, "A novel video key frame extraction algorithm based on percived motion energy model," *IEEE transactions on circuits and systems for video technology*, vol. 13, no. 10, Oct 2003.

[2] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion videol," *ACM Multimedia Syst*, vol. 1, no. 1, pp. 10–28, 1993.

[3] G. Ciocca and R. Schettini, "An innovative algorithm for keyframe extraction in video summarization," *Journal of Real-Time Image Processing (Springer)*, vol. 1, no. 1, pp. 69–88, 2006.

[4] P, Agraim, H, Zhang, D, and Petkovic, "Content-based representation and retrieval of visual media: A state of the art review," *Multimedia Tools and Applications*, vol. 3, no. 1, pp. 179–202, 1996.

[5] M. Swain and D. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[6] J. R. Smith and S. F. Chang, "Tools and techniques for color image retrieval," *IST/SPIE Storage and Retrieval for Image and Video Databases IV*, vol. 26, no. 70, pp. 426–437, 1996.

[7] H. S. Chang, S. Sull, and U. L. Sang, "Efficient video indexing scheme for content-based retrieval," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1269–1279, 1999.

[8] L. Tieyan, X. Zhang, J. Feng, and K. T. Lo, "Shot reconstruction degree: a novel criterion for keyframe selectionl," *Pattern Recognition Letters*, vol. 25, pp. 1451–1457, 2004.

[9] I. C. Chang and K. Y. Cheng, "Content-selection based video summarization," in *IEEE International Conference On Consumer Electronics*, Las Vegas Convention Center, USA, Jan 2007, pp. 11–14.

[10] C. Huang and B. Liao, "A robust scene-change detection method for video segmentation," *IEEE Transactions on Circuits and System for Video Technology*, vol. 11, no. 12, Dec 2001.

[11] C. Ngo, Y. Ma, and H. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Transactions on Circuits and System for Video Technology*, vol. 15, no. 2, Feb 2005.

[12] A. Chitra, Dhawale, and S. Jain, "A novel approach towards key frame selection for video summarization," *Asian Journal of Information Technology*, vol. 7, no. 4, pp. 133–137, 2008.

[13] D. Chetverikov and Z. S. Szabo, "A simple and efficient algorithm for detection of high curvature points in planar curves," in *Proc. 23rd Workshop of the Austrian Pattern Recognition Group*, 1999, pp. 175–184.

[14] L. Congcong, Y. T. Wu, Y. Shiaw-Shian, and T. Chen, "Motion-focusing key frame extraction and video summarization for lane surveillance system," in *ICIP 2009*, 2009, pp. 4329–4332.

[15] Y. Boykov and V. Kolmogotov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. On PAMI*, vol. 26, no. 9, pp. 1124–1137, Sept 2004.

[16] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Trans. on PAMI*, vol. 30, no. 11, pp. 1971–1984, Nov 2008.

[17] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips:from humans to computers," *IEEE Transactions On Circuits And Systems For Video Technology*, vol. 19, no. 2, February 2009.

[18] Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1097–1105, Dec 2005.

[19] J. Peng and Q. Xiaolin, "Keyframe based video summary using visual attention clues," *MultMedMag*, vol. 17, no. 2, pp. 64–73, Apr-Jun 2010.

[20] S. Lee and M. Hayes, "Properties of the singular value decomposition for efficient data clustering," *IEEE Signal Processing Letters*, vol. 11, no. 11, pp. 862–866, 2004.

[21] Z. Cernekova. and et al, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits and Systems for video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[22] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *ACM International Conf. Multimedia*, 1999.

[23] T. Ren, Y. Liu, and G. Wu, "Full-reference quality assessment for video summary," in *IEEE International Conference on Data Mining Workshops*, 2008.