ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

A Study on Video Summarization Techniques

RaviKansagara[1], DarshakThakore[2], MahaswetaJoshi[3]

P.G.Student, Department of Computer Engineering, B.V.M. Engineering College, V. V. Nagar, India[1]

Associate Professor, Department of Computer Engineering, B.V.M. Engineering College, V. V. Nagar, India[2]

Assistant Professor, Department of Information Technology, B.V.M. Engineering College, V. V. Nagar, India[3]

ABSTRACT: The paper covers study of various techniques for key frames based video summarization available in the literature. There have been tremendous needs of video processing applications to deal with abundantly available & accessible videos. One of the research areas of interest is Video Summarization that aims creating summary of video to enable a quick browsing of a collection of large video database. It is also useful for allied video processing applications like video indexing, retrieval etc. Video Summarization is a process of creating & presenting a meaningful abstract view of entire video within a short period of time. Mainly two types of video summarization techniques are available in the literature, viz. key frame based and video skimming. For key frame based video summarization, selection of key frames plays important role for effective, meaningful and efficient summarizing process.

KEYWORDS: Video Summarization, Key Frame, Video Skim, Euclidian Distance, Depth Factor

I. INTRODUCTION

The rapid development of digital video capture and editing technology led to increase in video data, creating the need for effective techniques for video retrieval and analysis [2].

Advances in digital content distribution and digital video recorders, has caused digital content recording easy. However, the user may not have enough time to watch the entire video. In such cases, the user may just want to view the abstract of the video instead of watching the whole video which provides more information about the occurrence of various incidents in the video. [2]

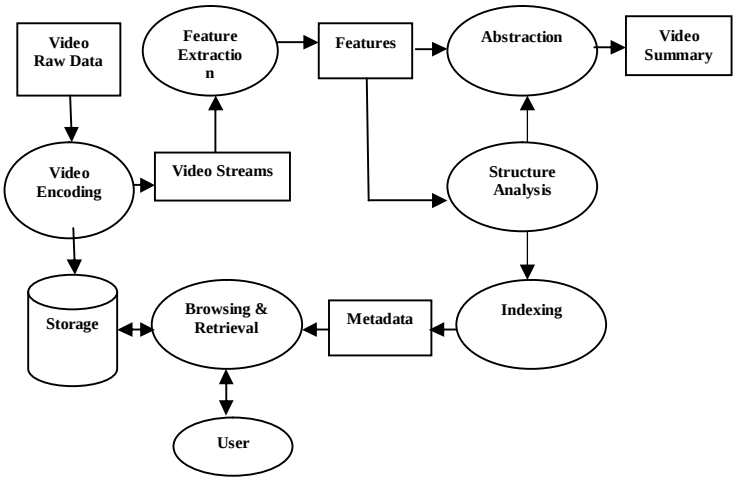

Fig 1: General application of the video analysis and indexing tasks [5]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

As the name implies, video summarization is a mechanism for generating a short summary of a video, which can either be a sequence of stationary images (key frames) or moving images (video skims) [2]. Video can be summarized by two different ways which are as follows.

1) Key Frame Based Video Summarization

These are also called representative frames, R-frames, still-image abstracts or static storyboard, and a set consists of a collection of salient images extracted from the underlying video source [2]. Following are some of the challenges that should be taken care while implementing Key frame based algorithm

1. Redundancy: frames with minor difference are selected as key frame.
2. When there are various changes in content it is difficult to make clustering.

2) Video Skim Based Video Summarization

This is also called a moving-image abstract, moving story board, or summary sequence [2]. The original video is segmented into various parts which is a video clip with shorter duration. Each segment is joined by either a cut or a gradual effect. The trailer of movie is the best example for video skimming.

The paper is organized as follows. Section II shows related work, Section III gives the overview and classification of key frame based video summarization. Section IV describes methods for video summarization. And section V concludes the paper.

II. RELATED WORK

A video summarization is a summary which represents abstract view of original video sequence and can be used as video browsing and retrieval systems. It can be a highlight of original sequence which is the concatenation of a user defined number of selected video segments or can be a collection of key frames. Different methods can be used to select key frames.

By using triangle model of perceived motion energy (PME) [4] motion patterns are modeled in video. 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 and the extracted key frames are representative.

In Visual frame Descriptors algorithm [5] three visual features: color histogram, wavelet statistics and edge direction histogram are used for selection of key frames. Similarity measures are computed for each descriptor and combined to form a frame difference measure. Fidelity, Shot Reconstruction Degree, Compression Ratio qualities are used to evaluate the video summarization [5].

In Motion Attention Model [6] shots are detected using color distribution and edge covering ratio that increase the accuracy of shot detection. Key frames are extracted from each shot by using the motion attention model. Here the first and last frame of every shots are considered as key frame and the others are extracted by adopting motion attention model [3][6]. These key frames are then clustered and a priority value is computed by estimating motion energy and color variation of shots.

In Multiple Visual Descriptor Features algorithm [7], 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.

Motion focusing method [8] 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. Background subtraction and min cut are mainly used in motion focusing.

In Camera Motion and Object Motion [9], the video is segmented using camera motion-based classes: pan, zoom in, zoom out and fixed. Final key frame selections from each of these segments are extracted based on confidence value formulated for the zoom, pan and steady segments.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

III. KEY FRAMES BASED VIDEO SUMMARIZATION

As explained in paper [1], key frames based video summarization works on frames so first step is to extract frames from original video frame sequence. In step two extracted video frames are cluster that have redundant content obviating the need for shot detection. Selection of key frames is proceeding in step three. The entire procedure is shown in fig 2.

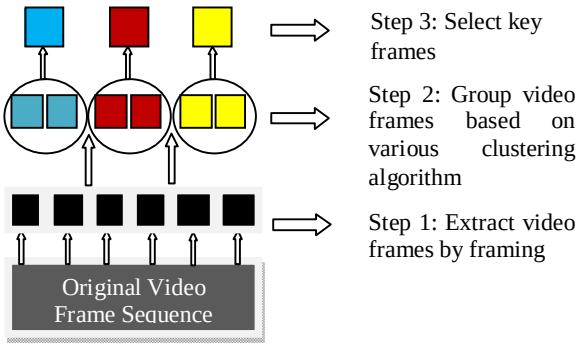

Fig.2: Key frames based video summarization [1]

As summarized in paper [11], Key frames based video summarization can be classified in three different ways. These are as follows.

1) Classification based on sampling

This method [11] chooses key frames uniformly or randomly under-sampling, without considering the video content. The summary produced by these methods does not represent all the video parts and may cause some redundancy of key frames with similar contents.

2) Classification based on scene segmentation

This method [11] extracts key frames using scenes detection, the scene includes all parts with a semantic link in the video or in the same space or in the same time. The disadvantage of these techniques is producing a summary, which does not take into account the temporal position of frames.

3) Classification based on shot segmentation

This method [11] extracts adapted key frames to video content. They extract the first image as shot key frames or the first and the last frames of the shot. These methods are effective for stationary shot and small content variation, but they don't provide an adequate representation of shot with strong movements.

IV. VIDEO SUMMARIZATION METHOD

Following are the various key frame extraction methods described by Sujatha C and Mudenagudi U in [3] along with other methods.

1) Video Summarization By Clustering Using Euclidean Distance [1]

This method is based on removing the redundant video frames which has almost similar content. Like many other approaches, the entire video material is first clustered into nodes, each containing frames of similar visual

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

content [1][10]. By representing each cluster with its most representative frame, a set of key frames is obtained which then summarizes the given sequence [1][10]. Procedure for this method is shown in fig 3 [1].

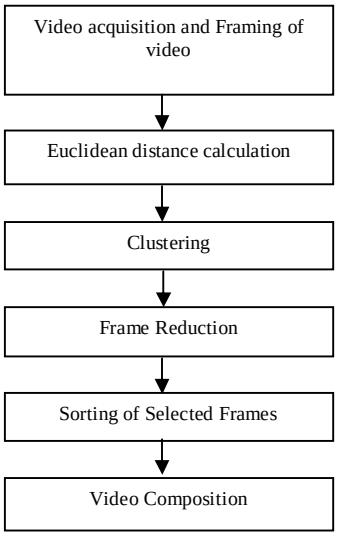

Fig 3: Video Summarization by Clustering Using Euclidean Distance [1]

2) Perceived Motion Energy Model (PME)

As described by A. Liu et al. [4], Motion is the more salient feature in presenting actions or events in video and, thus, should be the feature to determine key frames [4]. 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 [3]. The average magnitude Mag(t) of motion vectors in the entire frame is calculated as described in [3][4] as,

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

Where $MixFENi,(t)$ represents forward motion vectors and $MixBEni,j(t)$ represents backward motion vectors. N is number of macro blocks in the frame.

The percentage of dominant motion direction (t) is defined in [3][4] as,

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

(t, k) represents the angle histogram with n bins. The PME of a B-frame is computed in [4] as $P(t) = Mag(t).\alpha(t)$. These PME values of the frames are plotted which represent the 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 [3][4]. Here first the video sequence is segmented into shots using twin comparison method. The key frames are selected based on the motion patterns within the shots. For shots

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

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 [4]. The satisfactory rate for sports and entertainment video is found to be good as more actions exist when compared to home and news video [3].

3) Visual frame Descriptors

G. Ciocca and R. Schettini [5] introduced an algorithm with three visual features: color histogram, wavelet statistics and edge direction histogram are used 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 dx_H using the intersection measure is given in [5] as,

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

As defined in [5], 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,

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

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

These differences are combined to form the final frame difference measure d_{HWD} defined in [5] as,

$$d_{HWD} = d_H . d_W + d_W . d_D + d_D . d_H$$

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 [5]. The high curvature points are determined and by using two consecutive points key frames are extracted. Following qualities are used to evaluate the video summary [5].

1. Fidelity: The Fidelity measure is defined as a semi Hausdorff distance.
2. Shot Reconstruction Degree (SRD): It uses a suitable frame interpolation algorithm; we should be able to reconstruct the whole sequence from the set of key frames.
3. Compression Ratio (CR): CR is defined as ratio of number of key frames and total number of frames in the video sequence

4) Motion Attention Model

I. C. Chang et al. [6] used this model to detect shots. In this model, shots are detected using color distribution and edge covering ratio that increase the accuracy of shot detection. Key frames are extracted from each shot by using the motion attention model. Here the first and last frame of every shots are considered as key frame and the others are extracted by adopting motion attention model [3][6]. 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 in [6] as,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

$$TMA\left(S_i = \sum_{f_j \in S_i} MA(f_i) \times \log\left(\sum_{f_j \in S_i} MA(f_j)\right)\right)$$

Where $T(Si)$ denotes the sum of motion attention [6] value of shot i. And the energy motion change (EMC) is defined in [6] as,

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

Where $CF(Si)$ denotes the total number of frames that have significant intensity variation in shot i. The priority value of shot is defined in [6] as,

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

A higher PV value means that this shot is more important of this cluster and the shot will be the highlight of cluster [3].

5) Multiple Visual Descriptor Features

Chitra A.D et al. [7] used same visual features as Ciocca [5] 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 in [7] as,

$$f = \left\{ \begin{matrix} \sigma_1^{LH}, \sigma_1^{HL}, \sigma_1^{HH}, \frac{1}{2}\sigma_2^{LH}, \sigma_1^{HL}, \ldots . \\ \frac{1}{2^{L-1}}\sigma_L^{LH}, \frac{1}{2^{L-1}}\sigma_L^{HL}, \frac{1}{2^{L-1}}\sigma_L^{HH}, \frac{1}{2^{L-1}}\sigma^A, \mu^A \end{matrix} \right\}$$

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 [7]. The algorithm is tested on educational video sequence and compared with the I-frames obtained by Cue Video and found that the method gives better result [3].

6) Motion focusing

Congcong et al. [8] proposed motion focusing method. This method extracts key frames and generate 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 [3][8]. 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 $fi = 1$ or background $f_i = 0$. The labeling problem is solved minimizing the Gibbs energy, defined in [8] as,

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

Where $E1$ and $E2$ are defined in [8] as,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

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

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

$E2 = (fi - ff)$, di is difference between the current frame and the Gaussian mean for the i_{th} pixel and $k^t t$ = 1, 2, 3 are the thresholds for the i_{th} pixel [3][8]. The key frame extraction and summary image generation is done through two steps of mosacing. The initial mosacing 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 [8]. Then a second time mosacing is carried on by mosacing the key frames to generate the summarization image. The summary not only represents all objects in the focused motion but also provide temporal and spatial relation.

7) Camera Motion and Object Motion

Jiebo Luo et al [9] have proposed a method to extract key frames from personal video clips. The key frames are extracted from consumer video space where the content is unconstrained and lack of pre-imposed structures [3]. The 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 [3][9]. 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 [9]. 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 $dpan$ is given in [9] as: $dpan = \alpha 1\, dspat + \alpha 2\, dknow$ with $\alpha 1 + \alpha 2 = 1$, $dspat$ is probability function of the cumulative camera displacements and $dknow = G(\mu + \sigma)$ is a Gaussian function, with being the location of local minimum and σ the standard deviation computed from the translation curve [9].

V. CONCLUSION

Video summarization plays important role in many video applications. A survey on various methods for key frame based video summarization has been carried out. But there is no any universally accepted method available for video summarization that gives better output in all kinds of videos. The summarization viewpoint and perspective are often application-dependent. The semantic understanding and its representation are the biggest issues to be addressed for incorporating diversities in video and human perception. Depending upon the changes in contents of the video, the key frames are extracted. As the key frames need to be processed for summarization purpose, the important contents must not be missed.

REFERENCES.

1. Sony, A.; Ajith, K.; Thomas, K.; Thomas, T.; Deepa, P.L., "Video summarization by clustering using euclidean distance," *Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on* , vol., no., pp.642,646, 21-22 July 2011

2. Truong, B. T. and Venkatesh, S. 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. 3, 1, Article 3, Feb. 2007

3. Sujatha, C.; Mudenagudi, U., "A Study on Keyframe Extraction Methods for Video Summary," *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on* , vol., no., pp.73,77, 7-9 Oct. 2011

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

4. T. Liu, H. J. Zhang, and F. Qi, "A novel video key frame extraction algorithm based on perceived motion energy model," *IEEE transactions on circuits and systems for video technology*, vol. 13, no. 10, Oct 2003, pp 1006-1013.

5. 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.

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

7. 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.

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

9. 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.

10. Nalini Vasudevan, Arjun Jain and Himanshu Agrawal, "Iterative Image Based Video Summarization by Node Segmentation".

11. Sabbar, W.; Chergui, A.; Bekkhoucha, A., "Video summarization using shot segmentation and local motion estimation," *Innovative Computing Technology (INTECH), 2012 Second International Conference on*, vol., no., pp.190, 193, 18-20 Sept. 2012