International Journal of the Physical Sciences Vol. 6(14), pp. 3377–3388, 18 July, 2011
Available online at http://www.academicjournals.org/IJPS
DOI: 10.5897/IJPS11.662
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

Weighting low level frame difference features for key frame extraction using Fuzzy comprehensive evaluation and indirect feedback relevance mechanism

Naveed Ejaz and Sung Wook Baik*

College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea.

Accepted 8 June, 2011

Video summarization is a method to generate succinct version of a video by eliminating the redundant frames. The representation of video summaries using key frames is a simple and effective way to generate video summaries. However, eliciting the frames that effectively characterize a video is a daunting task. A popular way to extract key frames is to compute the frame difference between the consecutive frames and then labeling a frame as key frame if a significant difference is located. In this paper, we propose a novel framework in which multiple index features, obtained from video frames, are combined to describe the frame difference between consecutive frames. It is observed that certain frame difference features have more influence in generating a representative frame difference measure. Moreover, some features are more relevant than others in different video genres. Therefore, for each video genre, the weights of different features are pre-determined at training phase by indirectly utilizing the Relevance Feedback Mechanism. Fuzzy Comprehensive Evaluation has been used to evaluate the efficiency of a particular frame difference measure based on the users' feedback about summaries and thus generating the weights of each measure. The framework is evaluated based on three popular frame difference measures including color histogram, correlation and edge orientation histogram. The experimental results, based on an objective evaluation criteria, show that our technique gives better results as compared to some of the other techniques in the literature.

Key words: Image processing, key frame extraction, video summarization, fuzzy comprehensive evaluation, relevance feedback mechanism.

INTRODUCTION

Recent years have witnessed an enormous increase in video data on the internet. This rapid increase demands efficient techniques for management and storage of video data. Video summarization is one of the commonly used mechanisms to build an efficient video archiving system. The video summarization methods generate summaries of the videos which are the sequences of stationary or moving images (Money and Agius, 2008). Key frame extraction is a widely used method for video summarization. The key frames are the characteristic frames of the video which render limited, but meaningful information about the contents of the video (Li et al., 2001). The extracted key frames from the video can be arranged chronologically to generate a storyboard. In video archiving systems, the key frames can be used for indexing in such a way that the content based indexing and retrieval techniques developed for image retrieval can be applied for video retrieval (Ciocca and Schettini, 2006).

In general, a key frame extraction technique must be fully automated in nature and must use the contents of the video to generate summary. Theoretically, key frames must be extracted using high level features such as objects, actions and events. However, the key frame extraction based on high level features is mostly specific

*Corresponding author. E-mail: sbaik@sejong.ac.kr. Tel: +82-02-3408-3797. Fax: +82-02-3408-4339.

to certain applications and usually low level features have been employed (Ren et al., 2010). Some of the examples of low features that are commonly used are color histogram, correlation, moments, edges and motion features. These low level features can then be employed to derive high level features to generate domain specific applications (Antani et al., 2002; Fredembach et al., 2004; Schettini et al., 2004).

The low level features, extracted from individual frames of the video, have been employed in a variety of ways by researchers. A common methodology is to compare consecutive frames based on some low level Frame Difference Measures (FDMs) and extract a key frame if this difference exceeds a certain threshold (Li et al., 2001). However, a single feature for computing frame difference is usually not sufficient to capture all the visual details of the image. Therefore certain low level features can be combined to get an effective representation of a frame (Yu and Seah, 2011). For instance, color histograms have been a very popular feature for image representation and computation of key frames. However, key frame methods that use color histograms as FDM, tends to fail in scenes with illumination changes. Edge orientation feature, on the other hand, is expected to behave well under this situation. The choice of a particular low level feature depends on the video genre to some extent. Also, some features are visually more important then others in certain video genres. For instance, in a video of a soccer game, where the camera is mostly focused on the field, edge orientation is an appropriate feature to capture the camera motion. This means that for a particular genre of videos, different visual features must be combined with varying weights, giving more weight to the visual feature (or FDM) which provides more detail about the visual content of the video.

There can be different ways of assigning weights to the features. One of the most widely used schemes to assign weights to individual features in content based image retrieval (CBIR) systems is “Relevance Feedback Mechanism” (Zhou and Huang, 2003). In CBIR systems, the Relevance Feedback Mechanism incorporates the human users’ feedback in the search mechanism and adjusts the feature weights based on this feedback. This weight adjustment process is usually performed at the time of system design. In the context of key frame extraction, the Relevance Feedback can be employed by asking human users to appraise the summaries and adjusting weights accordingly. However, fine-tuning weights in this way is a time consuming and tedious job for the problem of key frame extraction.

In this context, we propose a Fuzzy Comprehensive Evaluation (FCE) based mechanism to compute the weights of each feature for specific genres of video. Our proposed technique can be used to combine any number and type of FDMs. For experimental purposes we use color histogram, correlation and edge orientation measures. Our technique indirectly incorporates the Relevance Feedback Mechanism in generating weights of different FDMs. For this purpose we have exploited the evaluation of key frames strategy suggested by Avila et al. (2011). This strategy compares the automated video summary with human users created summaries, and then represents the closeness of two summaries using two metrics: Accuracy Rate and Error Rate. In training phase, we extract key frames separately for each frame difference measure and compute Accuracy and Error Rates. The Accuracy and Error Rates are fuzzified and then FCE is applied to determine the weights of each FDM. The fuzzy set theory is applicable to the problem as the evaluation of efficiency of the key frame extraction techniques is intrinsically subjective in nature. In the end, we compared our technique with some of the well known techniques for key frame extraction, based on the evaluation scheme of Avila et al. (2011). The results are promising and show the effectiveness of assigning different weights to different FDMs.

The main contributions of this paper include:

1. Design of a completely automated and unsupervised method for determination of weights of FDMs for key frame extraction
2. Incorporation of Indirect Relevance Feedback Mechanism without direct involvement of the user at training time for adjusting weights.
3. Use of fuzzy methods and Fuzzy Comprehensive Evaluation to determine the weights of individual features.

LITERATURE REVIEW

As narrated earlier in introduction, the researchers have attempted to exploit various features for the extraction of key frames in videos. These features have been utilized in a variety of different ways. Some of the low level features which are commonly used includes color histogram, frame correlation, motion information and edge histogram etc. (Jiang et al., 2009).

Hanjalic et al. (1996) extracted key frames by computing the percentage of the accumulative histogram differences of each shot to the total frame difference accumulation of the sequence. Each shot is then assigned a part of the given key frames based on this percentage. Zhang et al. (1997) used the color histogram difference between the current frame and the last extracted key frame to draw out key frames from the video. Gunsel and Tekalp (1998) compared the histogram of current frame with the average color histograms of the previous frames to compute the discontinuity value. The Open Video project (www.open-video.org) utilizes the algorithm that was originally proposed by DeMenthon et al. (1998). In this technique,

the video sequence is represented as a trajectory curve and the discontinuities on this curve represent key frames. A key frame extraction scheme based on the cumulative frame differences of color histogram, histogram of edge orientation and wavelets was presented by Ciocca and Schettini (2006). In this scheme, the cumulative frame differences are used to build a curve. The sharp changes on this curve are identified and midpoint of two sharp changes is selected as key frame. In the Adaptive Sampling algorithm (Hoon et al., 2000), key frames are extracted by uniformly sampling the y-axis of the curve of cumulative frame differences. The key frames are obtained by sampling x-axis. Another low feature that has been utilized by some researchers for key frame extraction is the motion of objects and cameras in the video. For instance, Tianming et al. (2003) used motion vectors by modeling them using a triangle model of perceived motion energy to determine the frame where quick change starts. The technique presented by Yanzhuo et al. (2003) monitors the variation in magnitude and angle to determine the rapid change in the frames. Zhu et al. (2005) computes the motion intensity using motion vectors and selects those frames as key frames which have relatively high intensity of motion. An interesting technique, that is only based on camera motion, was presented by Guironnet et al.(2007). In this scheme, the rules defined on sequence and magnitudes of camera motions determine the key frames.

Some of the techniques for extraction of key frames are based on clustering. Such techniques aim to cluster video frames based on some similarity measure to select key clusters. Then one of the frames from each key cluster is selected as key frame. An unsupervised clustering approach based on color histogram features was presented by Zhuang et al. (1998). In this approach, the similarity measure of each frame is computed and compared with a threshold δ. If the similarity measure is less than δ, the node is not added to the cluster. The key clusters are those whose size is larger then the average cluster size. Ferman and Tekalp (2003) introduced a two stage method to extract hierarchal summaries in MPEG-7 videos. The first stage is carried at the time of contents production which uses fuzzy clustering and data pruning methods to obtain key frames. In the second stage, the number of key frames is reduced based on the browsing preferences of the user. The approach of Mundur et al. (2006) uses clustering based on Delaunay triangulation to cluster the color contents of the frame which are represented as multi-dimensional point data. The Delaunay diagram is then built and clusters are obtained by removing the separating edges. Furini et al. (2010) used HSV color descriptors to cluster frames. The technique is interactive where users are given choice to specify the desired number of key frames and the computational time limit. Avila et al. (2011) extracted key frames by using a slightly modified version of k means clustering to cluster the color features in HSV color space. The clustering is preceded by pre-sampling and removal of useless frames. The number of clusters 'k' used in k-means clustering is guessed by computing the pair-wise distance of consecutive frames.

A thorough survey of existing techniques reveals that the researchers have used many different visual features for the problem of key frame extraction. Some researchers have also tried to combine multiple features. However, the user feedback is generally not incorporated in deciding the weight of a particular factor. Our technique extends the work of key frame extraction by adjusting system parameters at training time.

METHODOLOGY

This section describes the main steps of our framework for the extraction of key frames. Figure 1 shows the main steps involved in training phase using one specific feature for as FDM. This training phase is repeated for every FDM that is used to build the feature vector. Our framework is independent of the number and type of FDMs. However, for the evaluation of framework, we used a combination of three comparison measures; color histogram, correlation and edge orientation histogram.

There are 10 videos in the training database for each genre. Each video is first pre-sampled to pick a frame after every 30 s to reduce the overall computational cost. Munder et al. (2006) assert that this pre-sampling does not affect the quality of key frame extraction. After pre-sampling the video, the particular frame difference feature is computed from the video frames. This feature is then used to extract key frames for the training video. To start the extraction process, the first frame is declared as a key frame. Then the frame difference is computed between the current frame and the last extracted key frame. If the frame difference is greater than a certain threshold, the current frame is selected as key frame. The threshold is automatically computed from the video, whereby the average value of the frame difference of the sampled frames is used for threshold. In computation of threshold, the useless frames are excluded. The useless frames are the frames which are not meaningful; for example a frame with dominantly black or white color. To detect a useless frame, we use a simple strategy suggested by Furini et al. (2010). The standard deviation of pixels in the frame is computed, and if the standard deviation is below a certain level then the frame is discarded. After extraction of key frames, we use the evaluation strategy of Avila et al. (2011) to compare the summaries generated by that technique with the human users' created summaries. This strategy compares the auto generated summaries with the summary view of human users and based on this comparison generates two quality metrics; Accuracy and Error Rates.

Before proceeding further, we briefly describe the comparison strategy of Avila et al. (2011). This scheme is originally designed to evaluate the quality of the summaries generated by various techniques. The auto generated summaries are compared with the summaries created by human users. This is done by comparing every frame in automatic summary with every frame of the users' summary. If the Manhattan distance between the color histograms of two frames is less than a certain threshold, the frames are said to be matched. In this way, the number of matching and non-matching key frames of automatic summaries and user summaries are determined. These measures are then used to compute Accuracy Rate (CUS_A) and Error Rate (CUS_E) using:

$$CUS_A = \frac{n_{MAS}}{n_{US}} \quad (1)$$

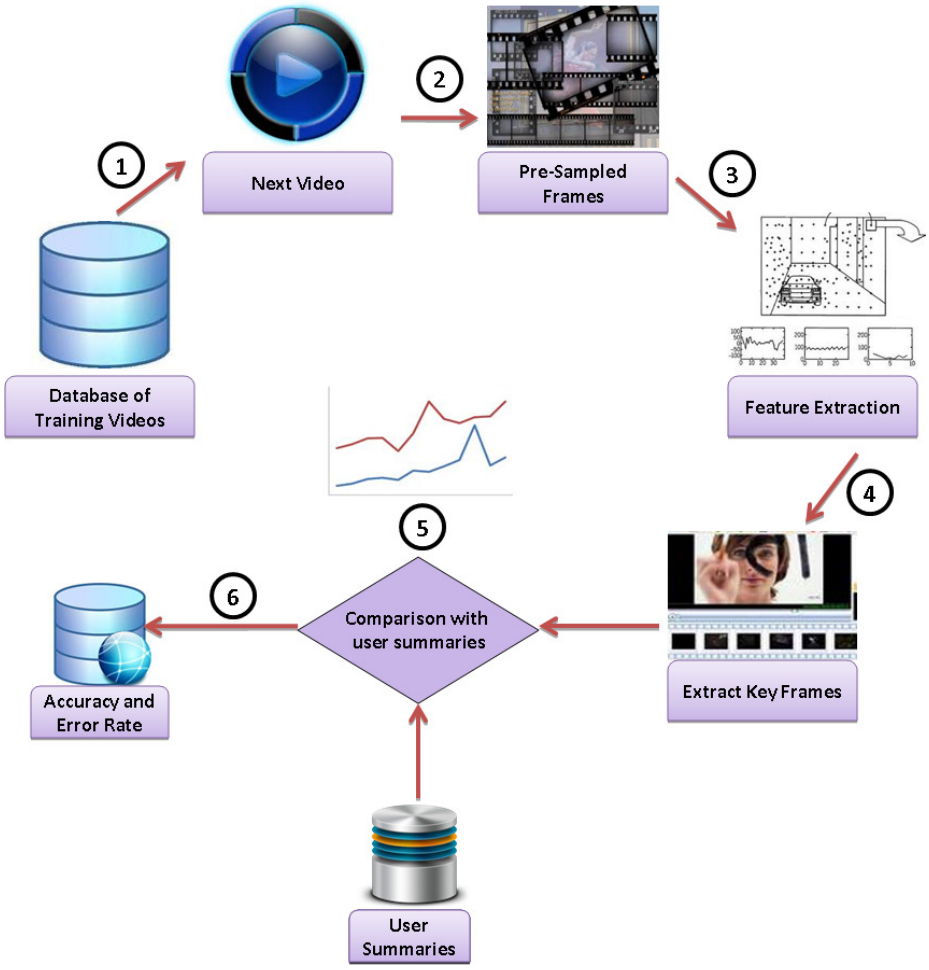

Figure 1. Computation of accuracy and error rates for training videos.

$$CUS_E = \frac{n_{MAS}}{n_{US}} \quad (2)$$

where

n_{MAS} = Number of Matching Key frames from Automatic Summary

n_{MAS}= Number of Non-Matching Key frames from Automatic Summary

n_{US}= Number of Key Frames from User Summary

The value of Accuracy Rate varies from 0 to 1, 1 being the best value where all frames of automated summary matches with all frames of user summaries. The value of Error Rate ranges from 0 to n_{AS} /n_{US} where 0 is the best value (n_{AS} is the number of frames in automatic summary). The quality of a summary is superior if it has high Accuracy Rate and low Error Rate.

Using this strategy, the Accuracy and Error Rates of the summaries for each FDM are computed for every training video. The evaluation of key frame extraction mechanism is inherently subjective in nature and the question of 'what is a key frame' is generally vague. Moreover, the factors effecting the human user's decision to declare a frame as key frame are not determined. This makes the problem of evaluation of summaries to be well suited for fuzzy analysis. The purpose of fuzzy set and fuzzy logic is to deal with problems involving knowledge expressed in vague linguistic terms. In a fuzzy set, each element of universe of disclosure is awarded a degree of membership value in a fuzzy set using a membership function. The membership function is used to associate a grade to each linguistic variable. We use four linguistic terms to represent the quality of the summary of a video. These fuzzy sets for quality include Very Good (VG), Good (G), Average (A) and Poor (P). The Accuracy and Error Rates of training videos are fuzzified based on the triangular like member functions to associate a level of degree of goodness and badness to the user summaries. Figure 2 shows the fuzzy membership functions for Accuracy and Error Rates respectively.

Next the FCE is separately applied to compute the weight for each FDM. FCE is a well known method which comprehensively judges the membership grade status of the items to be evaluated based on some factors. In our problem, we evaluate each FDM based on the factors of Accuracy and Error Rates by applying FCE. We first discuss the requirements of FCE and then explain its applicability to our method.

In general, FCE has following requirements:

1. A Factor Set $U = \{u_1, u_2, \ldots, u_m\}$ which is composed of m different factors. These factors influence the evaluation of objects.
2. An Evaluation Set $V = \{v_1, v_2, \ldots, v_n\}$ which is composed of n types of remarks.

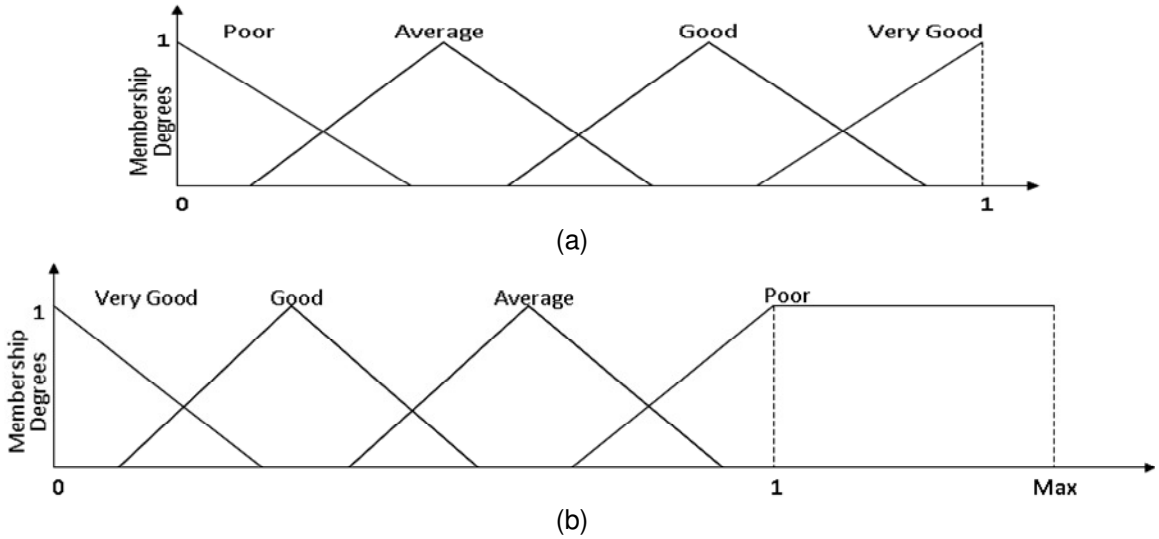

Figure 2. Membership functions for accuracy and error rates: (a) Membership function for accuracy rate; (b) Membership function for Error Rate (Max = nAS /nUS).

3. A Weight Set $W = \{w_1, w_2 \ldots w_m\}$ *where* $\sum_{i=1}^{m} w_i = 1$ *and* $w_i \geq 0$, where each member of this set represents the weight coefficients of each factor in the factor set U.
4. A Fuzzy transformation Γ_f that transforms Factor set U into the evaluation set V.
5. A Fuzzy relation R on UxV defined as:

$$R = (r_{ij})_{mxn} \in R_{mxn} \quad (3)$$

The membership degree of the subject to remark v_j from the viewpoint of factor u_i is given by:

$$r_{ij} = \mu_R(u_i, v_j) \; where \; r_{ij} \in [0,1], \; i = 1 \ldots m, \; j = 1 \ldots n \quad (4)$$

Based on sets U, V, W and Fuzzy relation R, the composition operation can be used to find the relation b that maps elements of U to V.

$$b = W \circ R = max_{v \in V}\{\min(\mu_W(u,v), \mu_R(v,w))\} \quad (5)$$

Each element b_i (I = 1, 2…, n) of set b represents the possibility of remark v_j in the evaluation V. The entries of b are then normalized to make the sum of b equal to 1. The credibility of a single factor is determined by multiplying b with the transformation function Γ_f.
We use FCE to determine the weight of each FDM used for key frame extraction. We modeled our problem into the scenario of FCE as under:

1. The accuracy rate (CUS_A) and error rate (CUS_E) are the criteria used to evaluate the efficacy of a specific measure thus they made up the factor set U.
2. The Evaluation set V includes the scale factor for CUS_A and CUS_E, such as Very Good (VG), Good (G), Average (A) and Poor (P). The values of these scale factors are determined by fuzzification of accuracy and error rates of the training videos and then averaging the degree of membership of each set.
3. For better quality of video summary, the value of CUS_A must be high and value of CUS_E must be low. Therefore, both the factors are equally important and are assigned equal weight to determine weight set W = (0.5, 0.5).
4. The Fuzzy transformation Γ_f is defined as:

$$\Gamma_f = \begin{bmatrix} 1 \\ 0.7 \\ 0.4 \\ 0.1 \end{bmatrix}$$

5. The single factor evaluation matrix R is then determined and b is computed using b = W◦R and the weight/credibility of a technique is found by multiplying b with fuzzy transformation function Γ_f.
6. The process is repeated to determined the weight of each frame difference measure

To make our point clear, we feel it appropriate to include a numerical example of the determination of weights. For this example, the training has been shown only on 3 videos using a single FDM. The Accuracy and Error Rates, the degree of membership after fuzzification and the average values of each scale factor for a particular FDM are given in Table 1.

Using Table 1, the matrix R is then given as:

$$R = \begin{bmatrix} 0.17 & 0.43 & 0 & 0 \\ 0.05 & 0.11 & 0.41 & 0 \end{bmatrix} \; and \; W = [0.5 \quad 0.5]$$

Therefore b = W◦R is determined as $b = [0.17 \quad 0.43 \quad 0.41 \quad 0]$ which is normalized to get $b = [0.16 \quad 0.42 \quad 0.4 \quad 0]$

$$Weight = b \,.\, \Gamma_f = [0.16 \quad 0.42 \quad 0.4 \quad 0] \begin{bmatrix} 1 \\ 0.7 \\ 0.4 \\ 0.1 \end{bmatrix} = 0.61$$

Table 1. A numerical example for fuzzification of accuracy and error rates.

Video	CUS_A	CUS_E	Degree of membership (CUS_A)				Degree of membership (CUS_E)			
			VG	G	A	P	VG	G	A	P
1	0.73	0.41	0.45	0.1	0	0	0	0	0.3	0
2	0.5	0.21	0	0.75	0	0	0.16	0.34	0	0
3	0.66	0.56	0.05	0.45	0	0	0	0	0.95	0
Average			0.17	0.43	0	0	0.05	0.11	0.41	0

For the evaluation of our framework, we have used three simple inter-frame difference measures: color histogram, correlation and edge orientation histogram. Next we briefly describe these FDMs.

The color histograms have been commonly used for key frame extraction in both frame difference based techniques and clustering techniques (Zhang et al., 1997; Gunsel and Tekalp, 1998; Ciocca and Schettinni). This is because the color is one of the most important visual features to describe an image. Color histograms are easy to compute and are robust in case of small camera motions. For computing FDM, color histogram has been built in HSV color space by performing a quantization step to reduce the number of distinct colors to 64. Instead of computing one histogram for the entire image, we divided image in a total of 'T_s' sections, each of size mxm. This is to effectively measure the level of difference between two frames. Each corresponding section of one frame is compared with the corresponding section of other frame using the histogram intersection mechanism. The histogram difference $HD_{i,j,s}$ between two corresponding sections 's' of histogram H_{is} of frame i and histogram H_{js} of frame j is defined as:

$$HD_{i,j,s} = 1 - \sum_{k=0}^{64} \min\left(H_{is}(k), H_{js}(k)\right) \quad (6)$$

The histogram difference "HD" between two frames i and j is then calculated by taking the average of the difference measure between each section.

$$HD_{i,j} = \frac{1}{T_s}\sum_{k=1}^{T_s} HD_{i,j,k} \quad (7)$$

The correlation coefficients have been very popular scheme to find similarity between two data sets. The correlation coefficients are invariant to brightness and changes in the contrast. Again, for computing correlation measure, we divide frames into T_s sections of size mxm. The correlation values of each section are then averaged. The correlation is measured for three color channel values red, green and blue. The correlation difference $CD_{p,q,s,c}$ of a color channel 'c' between two corresponding sections 's' of frame p and q is defined as:

$$CD_{p,q,s,c} = 1 - \frac{\sum_{p=1}^{m}\sum_{q=1}^{m}\left(F_{i,p,q,c} - \bar{F}_{i,c}\right)\left(F_{j,p,q,c} - \bar{F}_{j,c}\right)}{\sqrt{\sum_{p=1}^{m}\sum_{q=1}^{m}\left(F_{i,p,q,c} - \bar{F}_{i,c}\right)^2 \sum_{p=1}^{m}\sum_{q=1}^{m}\left(F_{j,p,q,c} - \bar{F}_{j,c}\right)^2}} \quad (8)$$

$$where s = 1....T_s, \quad c = red, green, blue, \bar{F}_{i,c} = mean\ value\ of\ channel\ c\ of\ the\ frame\ i$$

The correlations of all sections of frame i and j are averaged to obtain the overall correlation $CD_{i,j,c}$ for a color channel.

$$CD_{i,j,c} = \frac{1}{T_s}\sum_{k=1}^{T_s} CD_{i,j,k,c} \quad (9)$$

Then, the overall correlation difference measure $CD_{i,j}$ between frames i and j is obtained by averaging the value of each color channel.

$$CD_{i,j} = \frac{CD_{i,j,Red} + CD_{i,j,Green} + CD_{i,j,Blue}}{3} \quad (10)$$

The third measure used for computing is the histogram of edge orientation. The edges are good under illumination changes. The edges are first computed using horizontal and vertical Sobel operators which are then used to find gradient and angle of edges. The angles are then used to build a histogram of edge orientation. For simplicity, we defined only 72 bins for the angles. Moreover as suggested by Ciocca and Schettini (2006), the angles are computed for only those pixels where value of gradient is above a certain threshold. As in the case of histograms, we compare histograms of corresponding sections of the two frames. The edge histogram difference "ED" between two frames i and j is calculated by taking the average of the difference measure between each section.

$$ED_{i,j} = \frac{1}{T_s}\sum_{i=1}^{T_s}\left|ED_{ik} - ED_{jk}\right| \quad (11)$$

Using Fuzzy Comprehensive Evaluation method as discussed, the three weights W_H, W_C and W_E are generated for histogram, correlation and edge orientation histogram frame differences respectively. These weights are separately calculated for each genre of the video. Figure 3 shows the process of extraction of key frames. Each frame from the sampled video is compared with the key frame of the last step. Again, the first frame is selected as key frame to start the process. In each comparison, three values HD, CD and ED are computed corresponding to three FDMs. Each FDM is then multiplied by the corresponding weight depending upon the genre of the video. The combined value 'W' is then obtained by adding the three values.

$$W = W_H.HD + W_C.CD + W_E.ED \quad (12)$$

The value of W ranges from 0 to 1. If W is greater than a threshold "T", the frame is selected as key frame. The mean value of average color histogram difference, correlation comparison difference and edge orientation histogram difference is taken as threshold.

After extracting the key frames, we used two simple post-processing steps to fine tune the results. The first of these steps eliminates the useless frames which is done using the similar step as discussed in threshold computation. Such useless frames usually have a relatively higher difference with rest of the video frames so they are likely to get selected as key frames. The second step is applied to eliminate those frames from the set of key frames, which are very similar to some other key frames. This is accomplished by comparing the selected key frames among themselves using color histogram. If the difference is less than a

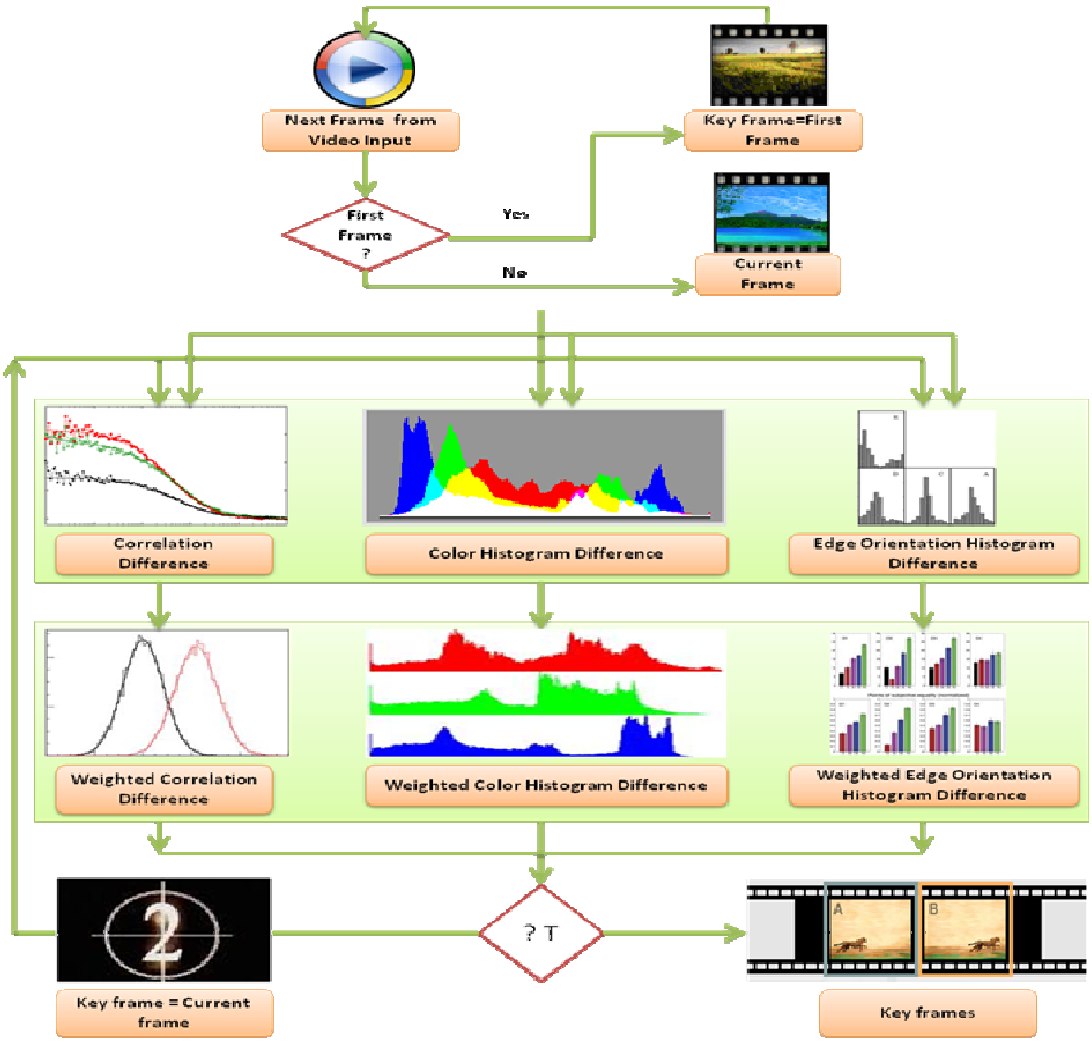

Figure 3. Key frame extraction process.

certain threshold then one of those frames is discarded. The remaining key frames are ordered sequentially.

RESULTS AND DISCUSSION

This section starts with a discussion, based on an example, to demonstrate that a single frame descriptor is usually not enough to capture all visual features of the image. Then, to evaluate the performance of our proposed scheme, we performed two different comparisons. In the first comparison, we compared our scheme with the key frame extraction techniques based on individual FDMs and the technique of combining FDMs using equal weights. In the second comparison, we compared our technique with some of the other techniques in the literature.

Figure 4 shows the change of frame difference for color histogram, correlation and edge orientation histogram respectively for a sample video. The video used is a documentary “Ocean floor Legacy, segment 04” from Open Video Project. It is quite evident from the graph that there is an obvious difference between the differences generated by three techniques. In other words, not all FDMs are capturing the visual details of the video in the same fashion.

Figure 5 shows two sample frames from the video. It can be effortlessly observed that there is not much dissimilarity is the contents of these two frames. However because of illumination changes, the second frame is somewhat darker than the first one. Because of this color change, the histogram difference measure yields a relatively high difference of 0.65 between these two frames. The correlation difference value is 0.16. Edge difference proves to be the most effective in case of illumination changes by producing a difference value 0f 0.11 only.

For the evaluation of the quality of video summary,

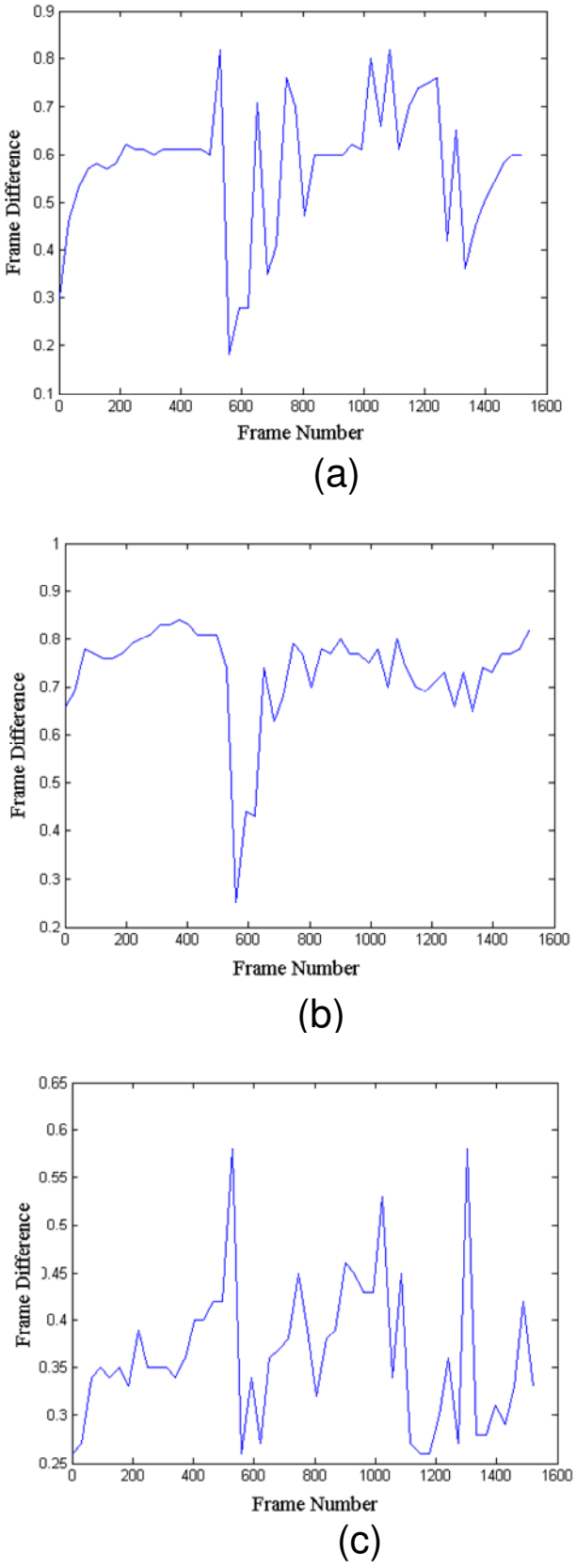

Figure 4. Change in frame difference for three difference measures: (a) Color histogram (b) Correlation difference (c) Edge orientation histogram.

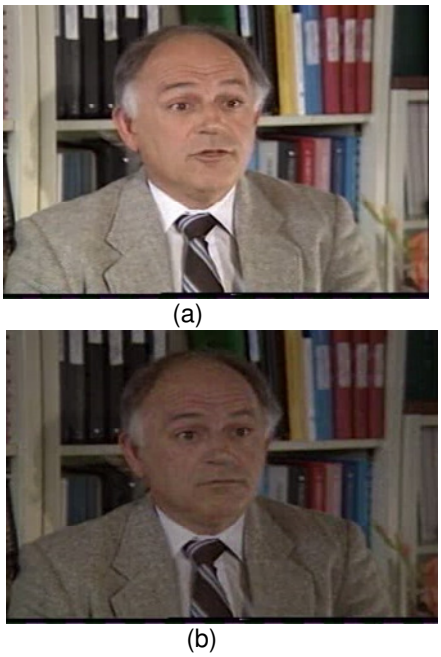

Figure 5. Two sample frame of video: (a) Frame 1321. (b) Frame 1351.

many techniques have been proposed in literature. However, a consistent framework for the evaluation of video summarization does not exist. Some of the evaluation strategies suggested in literature involve human users to judge the quality of summaries based on certain parameters (Yahiaoui et al., 2001; Li et al., 2003; Wang et al., 2007; Furini et al., 2010). An obvious drawback of these methods is that it is difficult to compare a new technique with the existing ones because of change of human subjects in the evaluation. Some researchers have also used some objective measures like Fidelity and Shot Reconstruction Degree to evaluate the summaries (Chang et al., 1999; Tieyan et al., 2004; Ciocca and Schettini, 2006). The problem with such techniques is that the human user feedback is missing and it is not clear that whether the metric is according to the human user judgment. We used the technique presented by Avila et al. (2011) for the evaluation of summaries. This technique has already been described in framework section where the same technique has been implied in the training phase. This technique not only provides objective measures in the form of Accuracy Rate (CUS_A) and Error Rate (CUS_E) but also incorporate user feedback in this comparison. Moreover, the data set and user summaries are publically available thus making comparison of techniques easier. As described earlier, a good technique must have a high Accuracy Rate (maximum 1) and low Error Rate (minimum zero).

Our experimental data contains videos of three genres: cartoons, documentaries and sports (soccer) videos. In our experimental set up, we used 10 videos for training

Table 2. Weights for various FDMs for different genres of videos.

FDMs	Color histogram	Correlation	Edge orientation histogram
Cartoons	0.6	0.2	0.2
Documentaries	0.45	0.25	0.3
Soccer	0.5	0.1	0.4

Table 3. Mean accuracy and error rates for individual and combined FDMs.

	CHDM	CDM	EODM	CEW	Our Technique
CUS_A	0.63	0.61	0.53	0.67	0.78
CUS_E	0.44	0.42	0.51	0.46	0.33

Table 4. Mean accuracy and error rates for various techniques.

	OV	DT	STIMO	Our technique
CUS_A	0.70	0.53	0.72	0.78
CUS_E	0.57	0.29	0.58	0.33

and 15 videos for testing for each genre. The data set and the user summaries used for evaluation are taken from the public data set of Avila et al. Table 2 shows the weights assigned to each FDM scheme using our methodology. It can be observed that the most important feature for the cartoons is color histogram owing to the use of bright colors. For documentaries, no single FDM generates overwhelming response. Finally for the soccer videos, histogram and edge difference measures proved to be more effective because of relatively fast camera and objects' motion.

Table 3 shows the mean Accuracy and Error Rates of the summaries generated separately by color histogram difference measures (CHDM), correlation difference measures (CDM), edge orientation difference measures (EODM), the three measures combined using equal weights (CEW) and the three difference measures combined using our method. It can be observed that combining difference measures by our methodology yields high Accuracy and low Error rates.

Table 4 compares the average results of our technique with OV (DeMenthon et al., 1998), DT (Mundur et al., 2006) and STIMO (Furini et al., 2010) based on Accuracy and Error Rates. As our techniques incorporate the users' opinion in generation of the summaries, therefore we get the highest Accuracy Rate. The Error Rate of our technique is lower than the values of all techniques except DT. This is because the DT approach produces summaries with very few frames, thus resulting in less number of mismatched frames. However this comparatively low Error Rate of DT is achieved at the cost of Accuracy Rate as number of matching frames is also less.

The summaries for the video "Ocean floor Legacy, segment 04" by these four techniques are shown in Figure 6. The user summaries by five different users are shown in Figure 7. It can be observed by visual comparison of summaries that our technique has generated summaries that are more close to the users' perception of summaries.

Figures 8 and 9 show one sample user summary and summary generated by our technique for videos of cartoon and sports genre respectively. The visual results show that our technique works well in videos with different visual content.

Conclusion

In this paper, we demonstrated the use of fuzzy principles in the assignment of weights to low level features for the task of key frame extraction from videos. It is generally concluded that one frame difference measure is usually not enough to capture all the visual contents of the image. The evaluation of summaries is a subjective task therefore fuzzy methods are best suited to assign weights to various frame difference measures. Since summaries are mostly used by human users, the weights must be assigned keeping in view the users' notion of the summary. The different measures for frame differences can be combined using our Fuzzy comprehensive Evaluation based framework. This framework of assigning variable weights based on indirect Relevance Feedback Mechanism and Fuzzy Comprehensive Evaluation has shown better results as compared to some of the other techniques to which it is compared. In future, we intend to

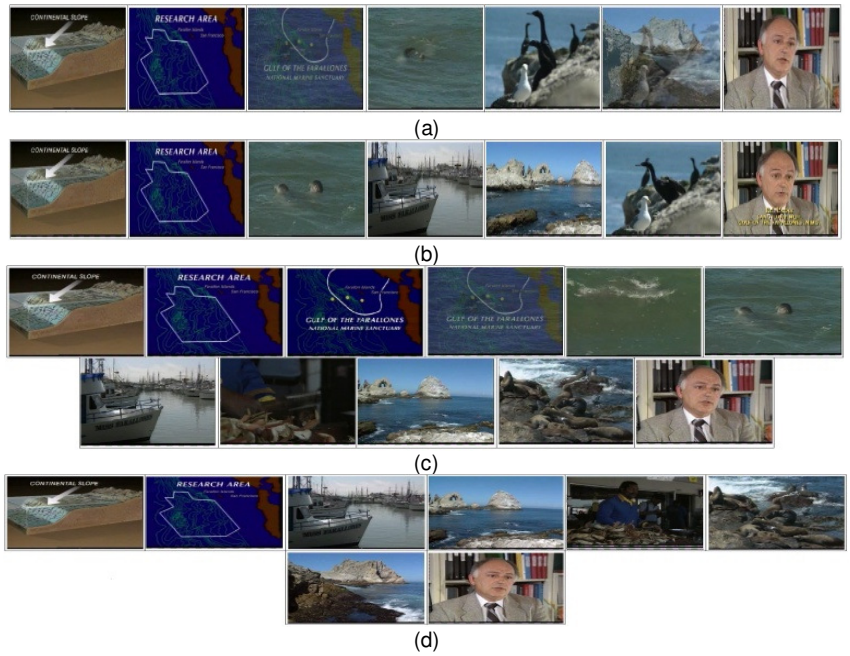

Figure 6. Summaries generated by various techniques for “Ocean floor Legacy, seg. 04”: (a) DT summary (b) OV summary (c) STIMO summary (d) Summary generated by our technique.

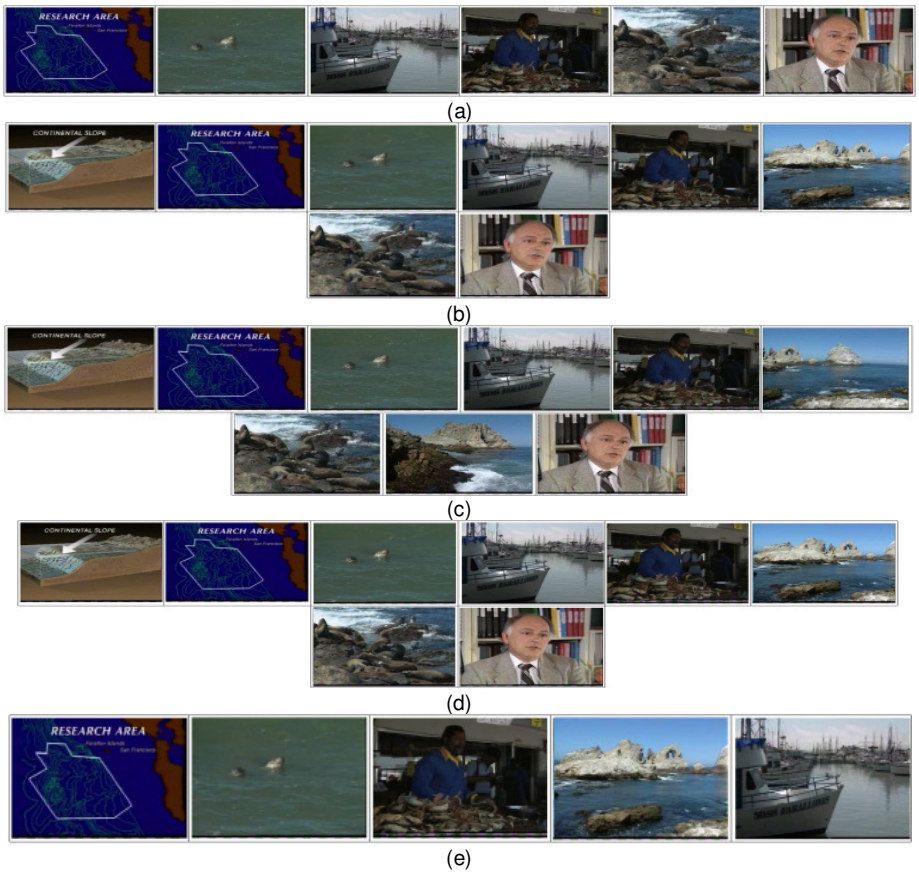

Figure 7. Summaries by five users for “Ocean floor Legacy, seg. 04”: (a) User 1 summary (b) User 2 summary (c) User 3 summary (d) User 4 summary (e) User 5 summary.

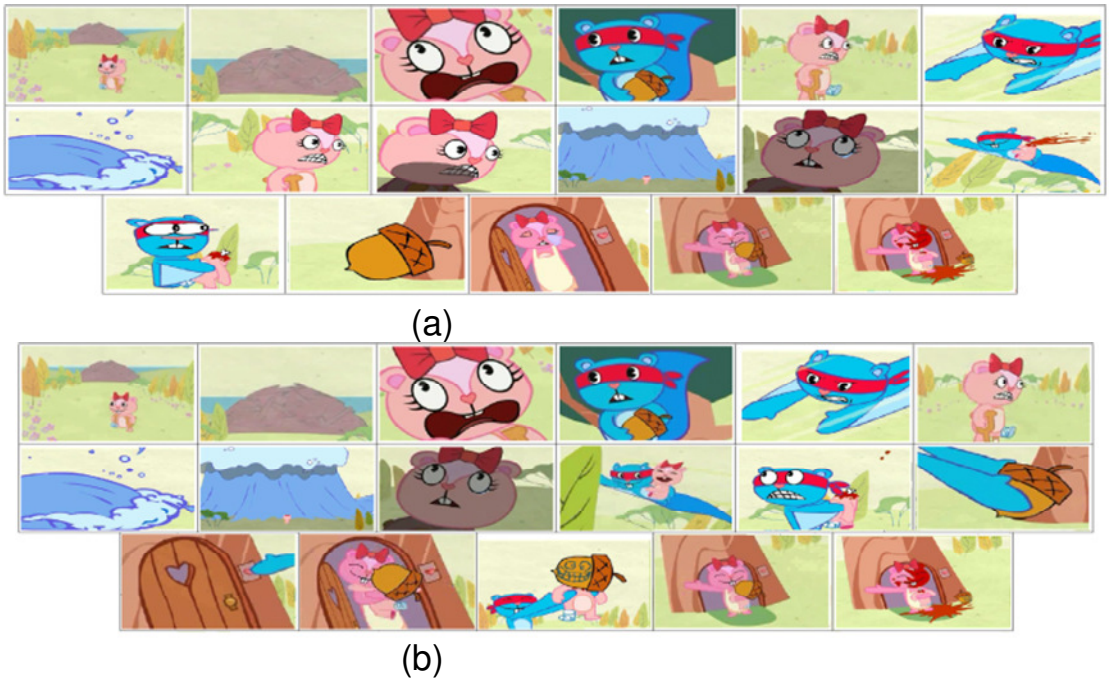

Figure 8. Summary by a user and results of our technique on a cartoon video. (a) A sample user summary (b) Summary generated by our technique.

Figure 9. Summary by a user and results of our technique on a sports video. (a) A sample user summary. (b) Summary generated by our technique.

test our framework by adding more visual features apart from the already used three features. We also intend to check the efficacy of our framework on various other genres of the video.

ACKNOWLEDGEMENT

This work is supported by the Seoul R & BD Program (JP090972M0214831).

REFERENCES

Antani S, Kasturi R, Jain R (2002). A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognit., 35(4): 945–965.

Avila SED, Lopes ABP, Antonio LJ, Araújo AdA (2011). VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognit. Lett., 32(1): 56-68.

Chang HS, Sull S, Lee SU (1999). Efficient Video Indexing Scheme for Content-Based Retrieval. IEEE Trans. Circuits Syst. Video Technol., 9(8): 1269-1279.

Ciocca G, Schettini R (2006). Innovative Algorithm for Key Frame Extraction in Video Summarization. J. Real Time Image Process, 1(1): 69-88.

DeMenthon D, Kobla V, Doermann D (1998). Video summarization by curve simplification. Proceedings of ACM International Conference on Multimedia. NY, USA, pp. 211–218.

Ferman AM, Tekalp AM (2003). Two-stage hierarchical video summary extraction to match low-level user browsing preferences. IEEE Trans. Multimedia, 5(2): 244–256.

Fredembach C, Schröder M, Süsstrunk S (2004). Eigen regions for Image Classification. IEEE Trans. Pattern Anal. Mach. Intell., 26(12): 1645-1649.

Furini M, Geraci F, Montangero M, Pellegrini M (2010). STIMO: STIll and moving video storyboard for the web scenario. Multimedia Tools Appl., 46(1): 47–69.

Guironnet M, Pellerin D, Guyader N, Ladret P (2007). Video Summarization Based on Camera Motion and a Subjective Evaluation Method. EURASIP J. Image Video Process. 2007(2007):1-12.

Gunsel B, Tekalp AM (1998). Content-based video abstraction. Proceedings of IEEE International Conference of Image Processing, Chicago, USA, 1998, pp. 128–132.

Hanjalic A, Langendijk RL, Biemond J (1996). A new key-frame allocation method for representing stored video-streams. 1st International Workshop on Image Databases and Multimedia Search, pp. 67-74.

Hoon SH, Yoon K, Kweon I (2000). A new Technique for Shot Detection and Key Frames Selection in Histogram Space. Proceedings of the 12th Workshop on Image Processing and Image Understanding, pp. 217-220.

Jiang RM, Sadka AH, Crooks D (2009). Advances in Video Summarization and Skimming. In: Grgic M et al. (eds.) Recent Advances in Multimedia Signal Processing and Communications, Springer, Berlin, 231: 27-50.

Li Y, Zhang T, Tretter D (2001). An overview of video abstraction techniques. Tech. Rep., HP-2001-191, HP Laboratory.

Li Z, Katsaggelos K, Gandhi B (2003). Temporal rate-distortion based optimal video summary generation. Proceedings of IEEE International Conference on Multimedia and Expo, Washington DC, USA, pp. 693–696.

Money AG, Agius H (2008). Video summarisation: A conceptual framework and survey of the state of the art. J. Visual Commun. Image Represent., 19(2): 121-143.

Mundur P, Rao Y, Yesha Y (2006). Keyframe-based video summarization using Delaunay clustering. Int. J. Digital Lib., 6(2): 219-232.

Ren J, Jianga J, Feng Y (2010). Activity-driven content adaptation for effective video summarization. J. Visual Commun. Image Represent. 21(8): 930-938

Schettini R, Brambilla C, Cusano C, Ciocca G (2004). Automatic classification of digital photographs based on decision forests. Int. J. Pattern Recognit Artif Intell., 18(5): 819-846.

Tianming L, Zhang HJ, Qi FH (2003). A novel video key-frame-extraction algorithm based on perceived motion energy model. IEEE Trans. Circuits Syst. Video Technol. 13(10): 1006-1013.

Tieyan LM, Zhang X, Feng J, Lo KT (2004). Shot reconstruction degree: a novel criterion for key frame selection. Pattern Recognit. Lett. 25(12): 1451–1457.

Wang T, Mei T, Hua XS, Liu XL, Zhou HQ (2007). Video collage: A novel presentation of video sequence. Proceedings of IEEE International Conference on Multimedia and Expo, pp. 1479–1482.

Yahiaoui I, Mérialdo B, Huet B (2001). Automatic video summarization. Proceedings of Multimedia Content-based Indexing and Retrieval (MCBIR).

Yanzhuo M, Chang Y, Yuan H (2003). Key-frame extraction based on motion acceleration. Opt. Eng., 47(9): 090501.

Yu J,Seah HS (2011). Fuzzy Diffusion Distance Learning for Cartoon Similarity Estimation. J. Comput. Sci Technol., 26(2): 203-216

Zhang HJ,Wu J, Zhong D, Smoliar SW (1997). An integrated system for content-based video retrieval and browsing. Pattern Recognit., 30(4): 643–658.

Zhu X, Elmagarmid AK, Xue X, Wu L, Catlin AC (2005). Insight video: toward hierarchical video content organization for efficient browsing, summarization and retrieval. IEEE Trans. Multimedia, 7(4): 648– 666.

Zhou X, Huang T (2003). Relevance feedback for image retrieval: a comprehensive review. Multimedia Syst., 8(6): 536–544.

Zhuang Y, Rui Y, Huang TS, Mehrotra S (1998). Adaptive Key Frame Extraction Using Unsupervised Clustering. Proceedings of International Conference on Image Processing, pp. 866-870.