Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Research Paper

# An Approach on Keyframe-Based Video Summarization Using Color Histogram and Evaluation Method

Engineering

**KEYWORDS :** Video Summarization, HSV Histogram, Euclidean Distance, Clusters, Correlation, Evaluation method

**Himani Parekh** CGPIT, Bardoli

**Pratik Nayak** SRIMCA,Bardoli

## ABSTRACT

*There has been tremendous needs of video processing applica-tions 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, video retrieval etc. This paper describes the key frame based video summarization, which is a process of creating & presenting a meaningful ab-stract view of entire video within a short period. Mainly two types of key frame based video summarization techniques are availa-ble, viz. Scene segmentation based and shot segmentation based video summarization. For key frame based video summarization, clustering and selection of key frames play important role for effective, meaningful and efficient summarizing process. HSV color histogram is used to find the distance between frames. Ac-cording to that distance frames are clustered into shots. The comparative analysis shows that the implemented technique gen-erates summaries that are closer to the summaries created by humans.*

## I. Introduction

Due to the recent advance in the computing and network infrastructure, together with the widespread use of digital video technology, demand for various multimedia applications is rapidly increase. There is a strong demand for a mechanism that allows the user to gain certain perspectives of a video document without watching the video in its entirety [2]. This mechanism is termed video summarization. A video sequence normally contains a large number of frames. 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). Video can be summarized by two different ways which are as follows.

### *A. Key Frame Based Video Summarization*

It is also known as representative frames or still-image abstracts. The set consists of a collection of salient images extracted from the underlying video source [2].

### *B. Video Skim Based Video Summarization*

This type of summarization is also called a moving-image abstract or summary sequence [2]. This type of abstract consists of a collection of video segments (and corresponding audio) extracted from the original video. These segments are joined by either a cut or a gradual effect. It is itself a video clip, but of significantly shorter duration. One popular kind of video skim in practice is the trailer of movies.

Key frames based video summarization can be classified in three different ways [11]. (1) Classification based on sampling, (2) classification based on scene segmentation, (3) classification based on shot segmentation. Most of the work in video summarization extracts the key frames within each shot [3]. A Shot means sequence of frames captured from a single camera operation. Shot Detection is a process of identifying visual discontinuities along the time domain in video sequence. The disadvantage of shot segmentation is that it increases redundancy in summarized video. While in scene segmentation, it loses temporal order of frames in summarized video. One of the most challenging tasks 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 where they generate the summary with the original sequence of video [3].

The paper is organized as follows. Section II gives the overview of existing approach for key frame based video summarization. Section III and IV describes proposed methods for video summarization using Euclidean distance. Section V shows the results for proposed method and section VI 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 keyframe selection process is threshold free and fast and the extracted key frames are representative.

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

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. 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. *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. And the key frames corresponding to the mid points between each pair of consecutive curvature point are considered as representative frames.

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.

## III. Implemented Method

Our implemented summarization technique depends on removing the visual-content redundancy among video frames and selecting number of user defined unique frames to form summarized video. Like many other approaches, first step is to extract frames from original video frames sequence. And entire video material is clustered into shots; each shot contains frames of similar visual content. The most representative frame from each cluster is selected as a key frame. The combination of these key frames forms a summarized video sequence. As shown in Fig. 1, Framing, clustering, and key frame selection are three basic and major steps for any key frame based video summarization. First of all, frames are extracted from the original video and then these frames are further processed. In steps 2 and 3, we perform clustering or grouping of similar frames and selection of key frames.

## IV. Algorithm

### Shot segmentation using Color Histogram

**A. Video Acquisition**: Video acquisition is the process of converting an analog video signal to digital form. The first step is to acquire the video file to be summarized from the destination file. The resulting digital data are referred to as a digital video stream [1].

**B. Video Framing:** Pre-sampling: The frames are retrieved from the video with constant interval of time. At this step amount of data are reduced to improve the performance. The pre-sampling step is beneficial in videos having long shots. However, in videos having short shots, important parts of the content may be wasted. For this reason, the sampling rate must be selected carefully to prevent any loss of information [12]. We do not consider the audio of the video. Importance is given only to the visual content [10].

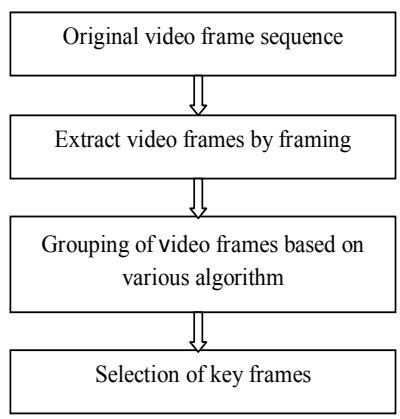



**Fig.1: Key frames based video summarization [1]**

**C. Calculate distance between frames using Color Histogram:** Here we used different approach for finding a difference of two frames because here we do not convert image into block but we just take whole image intensity for three colors. Some key issues of histogram-based techniques are the selection of an appropriate color space and the quantization of that color space. In implemented method, the color histogram algorithm is applied to the HSV color space, which is a popular choice for manipulating color. The HSV color space was developed to provide an intuitive representation of color and to be near to the way in which humans perceive and manipulate color. For that first of all, frames are converted into three different color i.e. H(hue), S(Saturation), V(Value).

The steps for histogram methods are as follows:

1) Compute the Histogram of $i^{th}$ and $(i+1)^{th}$ frames for different three colors $H_H$, $H_S$ and $H_V$, where $H_H$, $H_S$, and $H_V$ are histogram of hue, saturation and value respectively. Now calculate the difference between two frames using Euclidean distance [1]:

$$ED_C = \sqrt{\sum \left(H_{iC} - H_{(i+1)C}\right)^2}$$

$$where\ C = H, S\ and\ V$$

$$ED = \frac{(ED_H + ED_S + ED_V)}{3}$$

Where $H_{iC}$ and $H_{(i+1)C}$ are histograms for consecutive frames in different colors C. $ED_H$ , $ED_S$ and $ED_V$ are the distance of consecutive frames in Hue, Saturation and Value.

**D. Shot Boundary Detection:** The Euclidean distance is the primary parameter to detect boundaries. The following steps are used to detect boundaries.

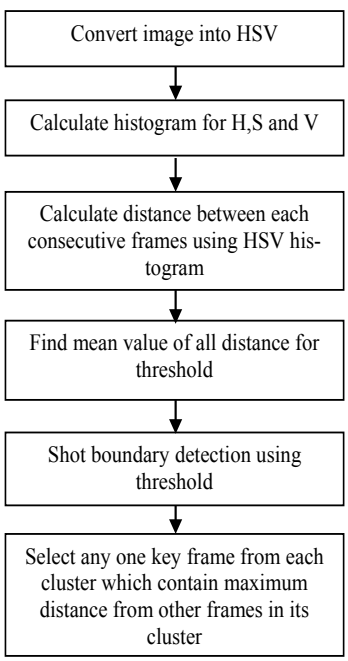



**Fig 2: Flow chart of the whole process**

1) Calculate the threshold

$$T = \sum_{i=1}^{N} \frac{ED_i}{N-1}$$

Distance between two consecutive frames is greater than threshold indicates boundary points of shot. For defining boundaries, following steps have to be performed:

For all value of ED, check

If ED(i) > Threshold then

$i^{th}$ and $(i+1)^{th}$ frames are dissimilar.

Hence, both frames should be in different cluster.

End

**E. Key Frame Extraction:**

For extracting key frames, following steps have to be performed:

For each cluster

If any cluster contains single frame,

that frame is considered as unused frame and that cluster is discarded.

If cluster contains more than one frames then

The frame that has maximum Distance from other frames in its cluster is considered as a key frame.

**F. Video Composition:** The selected frames from each node are combined to form the summarized video and saved as a new .avi file[1].

## Method for Evaluation of Automatic Summary

Evaluation of automatic video summaries is a challenging problem. This method is based on comparing automatic video summaries generated by video summarization techniques with ground-truth user summaries [13]. The objective of this evaluation method is to quantify the quality of video summaries, and allow comparing different video summarization techniques.

### • Correlation frame difference measure

The inter-frame correlation measures the similarity between two frames based on the color contents [12]. The correlation value have been widely used to capture the similarity between two frames. The correlation value is calculated for each color channel red, green and blue. Let F(t) and F(t + 1) be the two consecutive frames for which the correlation is calculated in [12] as:

$$r\left(F(t),F(t+1)\right)_c = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(F(t)_{c,i,j} - F_c(t)\right)\left(F(t+1)_{c,i,j} - F_c(t+1)\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(F(t)_{c,i,j} - \overline{F_c(t)}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n}\left(F(t+1)_{c,i,j} - \overline{F_c(t+1}\right.}$$

where $F(t)_{c,i,j}$ is the pixel value of 'c' color channel of F(t) at row 'i' and column 'j', and $\overline{F_c(t)}$ and $\overline{F_c(t+1)}$ are the mean values of the pixel values of color channel 'c' of frames F(t) and F(t + 1).

Mean of all color channel is taken to obtain the result of correlation comparison measure in [12] as.

$$\rho\left(F(t),F(t+1)\right) = \frac{r\left(F(t),F(t+1)\right)_{red} + r\left(F(t),F(t+1)\right)_{green} + r\left(F(t),F(t+1)\right.}{3}$$

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

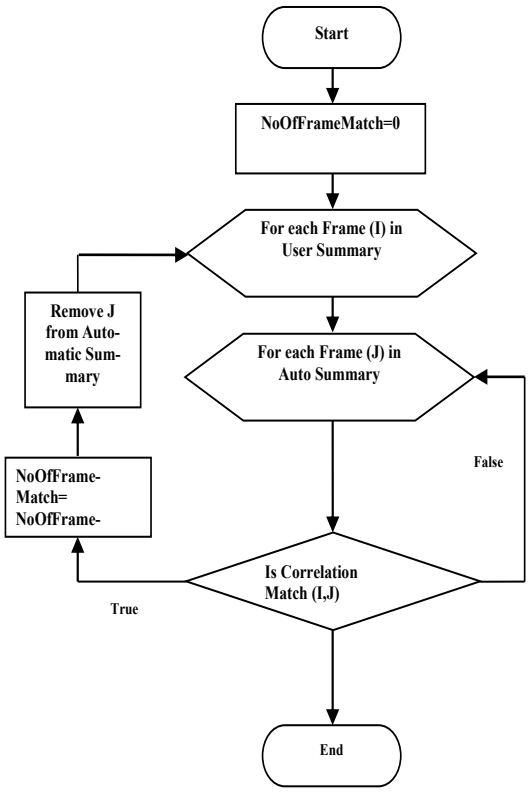



**Fig 3: Flowchart of the Implemented Evaluation Method**

- **Comparison with other techniques**

The summaries created manually by different users are taken as a reference for subjective comparison. Users have selected some important frames after watching the video. The frames selected by automatic summary are compared with these user summaries. If correlation between automatic summary frame and user summary frame is greater than some predefined threshold then both frames are considered as a similar frame. As per this mechanism, the quality of an automatically generated summary is determined by two metrics called Accuracy Rate ($CUS_A$) and Error Rate ($CUS_E$) which are defined in [13] as follows:

$$CUS_A = \frac{n_{mAS}}{n_{US}}$$

$$CUS_E = \frac{n_{m'AS}}{n_{US}}$$

Where $n_{mAS}$ = number of matching key frames from automatic summary (AS) $n_{m'AS}$ = number of non-matching key frames from automatic summary $n_{US}$ = number of key frames from user summary. The value of $CUS_A$ ranges from 0 to 1. The value 0 is the worst case where none of the key frames from automated summary matches with any of the user key frames, and the value 1 is the best case, which means that all the key frames of automatic summary matches with the user key frames [12].

The value of CUSE ranges from 0 to $n_{AS}$ /$n_{US}$ ($n_A$ number of frames in automatic summary). The val the best value for $CUS_E$ where no mismatch occ tween AS and user key frames, whereas $n_{AS}$ /$n_{U}$ worst value means none of the key frames are m The highest summary quality is achieved when CU and $CUS_E$ = 0. The summary having high $CUS_A$ d always mean a high quality summary until its $CUS_I$ ficiently low [12]. For instance, if a technique sel many key frames from the video, then it is likely high $CUS_A$ but low $CUS_E$. The experiments were ca data sets that are made publically available by Avi [13]. The user summaries, for each video in these da are also available. The evaluation on a common helps in proper evaluation of the framework and possible a comparison with other techniques. The consisted of 31 videos chosen from the Open Video (www.open-video.org). Based on these data sets, w pared our technique with OV, DT, and STIMO. A sa rate of 1 frame/s was selected as was used by Avi [13] for the same data set.

## V. Experiment Results And Performance An

Table 1 contains the various videos with its time du and total number of frames. This dataset [14][15] i for the experiments.

| No. | Name of Video[14] | Duration | No. of Frame |
|---|---|---|---|
| 1 | The Great Web of Water, segment 02 | 1:11 | 2118 |
| 2 | The Great Web of Water, segment 07 | 0:59 | 1745 |
| 3 | A New Horizon, segment 01 | 1:01 | 1806 |
| 4 | A New Horizon, segment 02 | 1:00 | 1797 |
| 5 | A New Horizon, segment 03 | 3:29 | 6249 |
| 6 | A New Horizon, segment 04 | 1:47 | 3192 |
| 7 | A New Horizon, segment 05 | 1:59 | 3561 |
| 8 | A New Horizon, segment 06 | 1:05 | 1944 |
| 9 | A New Horizon, segment 08 | 1:01 | 1815 |
| 10 | Take Pride in America, segment 01 | 1:30 | 2691 |
| 11 | HCIL Symposium 2002 - Introduction, segment 01 | 1:18 | 2336 |
| 12 | Exotic Terrane, segment 03 | 1:29 | 2676 |
| 13 | Exotic Terrane, segment 06 | 1:21 | 2425 |
| 14 | Exotic Terrane, segment 08 | 1:21 | 2428 |
| 15 | America's New Frontier, segment 01 | 1:59 | 3591 |
| 16 | America's New Frontier, segment 03 | 1:12 | 2166 |
| 17 | America's New Frontier, segment 07 | 2:00 | 3615 |
| 18 | The Future of Energy Gases, segment 03 | 1:37 | 2934 |
| 19 | The Future of Energy Gases, segment 05 | 2:00 | 3615 |
| 20 | The Future of Energy Gases, segment 09 | 1:02 | 1884 |
| 21 | The Future of Energy Gases, segment 12 | 1:36 | 2886 |
| 22 | Oceanfloor Legacy, segment 01 | 0:58 | 1740 |
| 23 | Oceanfloor Legacy, segment 02 | 1:17 | 2325 |
| 24 | Oceanfloor Legacy, segment 08 | 1:46 | 3186 |
| 25 | Oceanfloor Legacy, segment 09 | 1:10 | 2106 |
| 26 | The Voyage of the Lee, segment 05 | 1:09 | 2094 |
| 27 | The Voyage of the Lee, segment 15 | 1:15 | 2277 |
| 28 | The Voyage of the Lee, segment 16 | 1:27 | 2619 |
| 29 | Hurricane Force - A Coastal Perspective, segment 03 | 1:17 | 2310 |
| 30 | Drift Ice as a Geologic Agent, segment 03 | 1:31 | 2742 |
| 31 | Drift Ice as a Geologic Agent, segment 05 | 1:12 | 2187 |

**Table 1 : Video Dataset [14]**

methods (OV, DT and VISTO) with implemented approach with their $CUS_A$ and $CUS_E$.

**(a) User Summary 1**

**(b) User Summary 2**

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48


(c) User Summary 3

(d) User Summary 4


(e) User Summary 5

**Fig 4: User summaries for the video ‘‘A New Horizon, segment 02” [15]**

(a) OV Summary [15], $CUS_A$=0.33 $CUS_E$=0.09

(b) DT Summary [15], $CUS_A$=0.15 $CUS_E$=0.18


(c) VISTO Summary [15], $CUS_A$=0.38 $CUS_E$=0.12

(d) Implemented Method, $CUS_A$=0.80 $CUS_E$=0.16

**Fig 5: Summaries generated by various techniques for the video ‘‘A New Horizon, segment 02’’**

Fig 6 shows the 5 user summaries for video ‘‘A Oceanfloor Legacy, segment 08’’. Fig 7 shows the comparison of different methods (OV, DT and VISTO) with implemented approach with their $CUS_A$ and $CUS_E$.

(a) User Summary 1

(b) User Summary 2

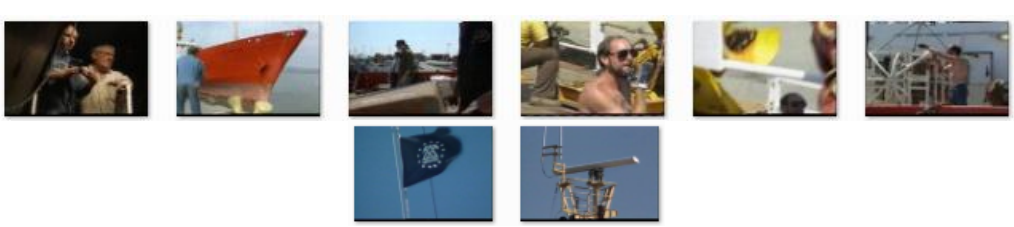
(c) User Summary 3

(d) User Summary 4

(e) User Summary 5

**Fig 6: User summaries for the video ''A Oceanfloor Legacy, segment 08'' [15]**

(a) OV Summary [15], $CUS_A$=0.44 $CUS_E$=0.09



(b) DT Summary [15], $CUS_A$=0.65 $CUS_E$=0.21

Volume : 5 | Issue : 5 | May 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

**(c) VISTO Summary [15], $CUS_A$=0.74 $CUS_E$=0.33**

**(d) Implemented Method, $CUS_A$=0.84 $CUS_E$=0.33**

**Fig 7: Summaries generated by various techniques for the video ''A Oceanfloor Legacy, segment 08''**

Results contain the key frames extracted by different techniques and key frames suggested by five users. We can visually match the key frames of implemented method to that of other techniques and user summaries and conclude that the most of the frames are similar. Compare to other techniques, implemented approach has high $CUS_A$ and low $CUS_E$.

Fig 8 shows the results of OV, DT, VISTO and implemented Method for 31 videos. Accuracy of implemented method is high as compare to other techniques but error is also increase. Main aim of summary is that, important frames must not be left in final output summary and hence to maintain the accuracy high, implemented method selects more key frames so that all key frames can be covered which are suggested by five users. But at the same time if any single user has suggest very less key frames as compare to other users, then it effects the performance and error rate increases due to that single user.

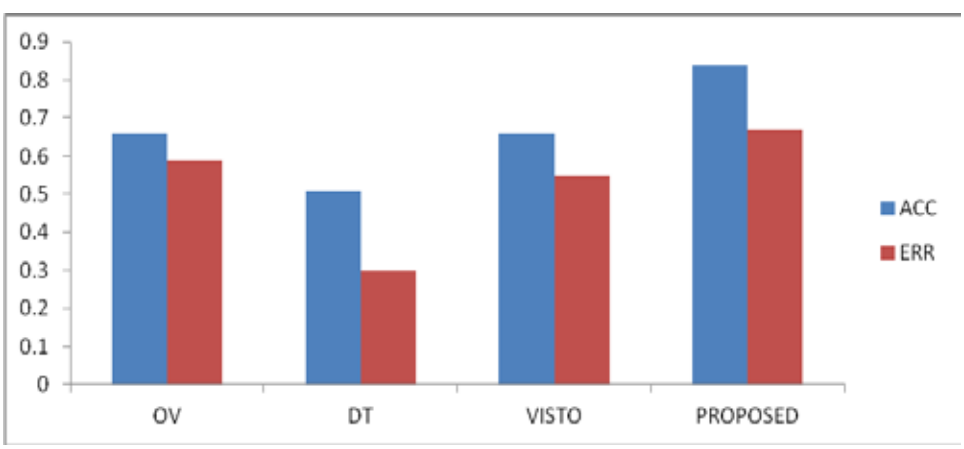



**Fig 8: Comparison Graph for Accuracy and Error**

## VI. CONCLUSION

There is no any universally accepted method available for video summarization giving better output in all kinds of videos. The summarization viewpoint and perspective are often application-dependent. Depending upon the contents and the change in contents of the video, the key frames are extracted.

The implemented technique uses color histogram comparison between frames. It also works well in videos with multiple shots. We evaluated our technique on an objective evaluation criteria developed by Avila et al. [13]. The technique is able to achieve reasonably higher Accuracy at the cost of error. In order to extract important key frames efficiently with high accuracy rate, error rate is compromised.

## VII. 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
[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," in *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," in *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
[12] Naveed Ejaz, Tayyab Bin Tariq and Sung Wook Baik "Adaptive key frame extraction for video summarization using an aggregation mechanism" *Journal of Visual Communication and Image Representation*, Volume 23 Issue 7, October, 2012, pp. 1031-1040
[13] S.E.D. Avila, A.B.P. Lopes, L.J. Antonio, A.d.A. Araújo, VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method, Pattern Recognition Letters 32 (1) (2011) 56–68.
[14] http://www.open-video.org/
[15] https://www.sites.google.com/site/vsummsite/