VISUAL ATTENTION BASED KEYFRAMES EXTRACTION AND VIDEO SUMMARIZATION

P.Geetha[1], S.Thiruchadai Pandeeswari[2], and SonyMohanan[3]

[1]Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai
geethap@annauniv.edu
[2]Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai
eshwarimsp@gmail.com
[3]Department of Electronics and Communication Engineering, Dhaanish Ahamad college of Engineering, Anna University of Technology, Chennai
sonymohanan@gmail.com

ABSTRACT

Recent developments in digital video and drastic increase of internet use have increased the amount of people searching and watching videos online. In order to make the search of the videos easy, Summary of the video may be provided along with each video. The video summary provided thus should be effective so that the user would come to know the content of the video without having to watch it fully. The summary produced should consists of the key frames that effectively express the content and context of the video. This work suggests a method to extract key frames which express most of the information in the video. This is achieved by quantifying Visual attention each frame commands. Visual attention of each frame is quantified using a descriptor called Attention quantifier. This quantification of visual attention is based on the human attention mechanism that indicates color conspicuousness and the motion involved seek more attention. So based on the color conspicuousness and the motion involved each frame is given a Attention parameter. Based on the attention quantifier value the key frames are extracted and are summarized adaptively. This framework suggests a method to produces meaningful video summary.

KEYWORDS

Visual Attention Index, Motion Intensity, Motion Coherence, Center-surround color difference, Adaptive summarization

1. INTRODUCTION

Video summarization is the process of summarizing a video by producing a summary of salient key frames that could convey the overall content of the video to the viewer so that a viewer could understand the content of the video without actually watching the video fully. A video summary conveys the actual content of the video using few most representative frames. Key frames that are used to represent the video in the summary are selected from the pool of thousands of frames of the video using various criteria. Lot of research is going on in this field and many approaches are proposed to extract potential key frames from the video and to summarize them.

Video summarization involves the two basic steps.

David C. Wyld, et al. (Eds): CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 179–190, 2012.
© CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2219

(1)Extracting key frames from the video

(2)Summarizing video by selecting representative keyframes from the extracted potential key frames.

Key frame extraction refers to finding a set of salient images taken from a full length video that represent the visual content in the video efficiently. The key frames extracted should be content based and the process of extraction should be automatic. Previously key frames were extracted randomly from the video to summarize the video. But randomly extracted frames will not represent the content of the video and information in the video efficiently. There are various approaches towards extraction of keyframes. *Sampling-based approach* is very simple, which selects key frames by uniformly sampling the frames from original video at fixed time intervals. The drawback of this approach is that it may cause some important yet short video clips to have no representative frames while less important but stationary clips could have multiple similar frames and lead to information redundancy.

The key frames extracted by *Segment based approach* could represent the video content efficiently, however, the clustering process is complex and it is difficult to determine the number of the segments. *Shot based techniques* are considered to be effective among the three available approaches as they extract representative keyframes from each shot. Most of the shot based approaches use low level spatial features like color and intensity histogram. The systems that use temporal features mostly use a threshold and compare dynamism between two frames. *Object based techniques* are efficient and semantic, but it gives more importance to the foreground and are suitable for certain applications only. Video summarization based on automatic understanding of semantic video content is far from being achieved.

2. RELATED WORKS

Two basic approaches are used to extract frames from video. They are *motion based* (activity based) key frame extraction methods and *color based* key frame extraction methods. In Color based method, color histograms are used to measure the similarity among frames. The distance between the color histogram of each frame is used as similarity measure. But using color histograms alone, the salient frames cannot be found as they may not capture the dynamics of the video. Reference [1] Proposes an approach to extract key frames from the video by considering color distribution with respect to time. It uses an algorithm to find a *temporarily maximum occurrence frame* (TMOF) in each shot. The color histogram of each frame is compared with the Color histogram of the TMOF. By comparing the pixel values at same position through frames within a shot, a reference frame is constructed with the pixel value throughout the frames in the same position with the maximum occurrence. The TMOF frame in each shot is the content descriptive frame of each shot that preserves both the temporal and color information. The distance measure between histograms of each frame and that of TMOF is calculated. The key frames are extracted from the peak of the distance curve in [1].

To capture dynamics of the video, TMOF is constructed instead of comparing color histograms of frames with one another. However, this method succeeds in capturing changes in the visual (color) content over time. The motion or activity involved is not considered. Amount of motion in each frame is not measured.

Reference [2] addresses this problem of lack of temporal consistency in the color based approaches of video summarization by combining both the color and motion based approaches. By doing so, more meaningful summarization can be achieved. The fundamental idea is to compute a single *image map which is representative of the pictorial content of the video sequence* by warping all the frames contained in it into a reference frame and combining their pixel

intensities. The two fundamental video summarization steps are: 1) Fitting a global motion model to the motion between each pair of successive frames. 2) Computing the summarizing image map by accumulating the information from all the frames after they have been aligned according to the motion estimates computed in the previous step. It assumes that there is a dominant motion among the motions of the objects in a scene and the dominant motion alone is taken into account blurring the actions of other objects. Considering only the dominant motion need very accurate implementation techniques, Reference [3] proposes another motion based key frame extraction technique. It quantifies motion involved in each video segment and accordingly decides the number of key frames required to represent the video in the summary. The assumption is that, more the motion in a scene, more key frames is extracted for effective and efficient summarization. In [3], the video is segmented and the intensity of motion in the video segments is directly calculated using MPEG-7 motion descriptors. The shots are divided into parts of equal cumulative motion intensity. *Frames located at the centre of each sub segment are selected to represent the video summary.* It also derives an empirical relationship between the motion activity of the segment and the no. of key frames required to represent them. Reference [4] presents an idea to capture visual content and the dynamism in the video as well. It uses color based key frame extraction technique. However, takes activity in the video into account for determining the no. of key frames to be represented in the summary. Frames are extracted from the video and each frame is converted from RGB color space to HSV color space and color histograms of all the frames are constructed. The first frame is stored as key frame and the threshold Ɛ(which is called dynamic threshold) is chosen. The histograms of consecutive frames are compared with the histogram of key frame. If the difference exceeds threshold value Ɛ, the corresponding frame is taken as key frame and stored in a pool. The upcoming frames are then compared with the histogram of the newly assigned key frame and the same procedure continues on.

Reference [5] extracts key frames based on a three step algorithm. The three major steps are *Preprocessing, Temporal filtering and post processing.* Both the first and last steps are done using queues. In the first step, Entropy based energy minimization method is used to find frames in the video that has gradual transition. Video frames enter the first queue. Frames that are insignificant are removed using an energy-minimization method. The output frames of the first queue are fed into the second queue. Redundant frames are removed in this queue .The frames that passed both the queues are clustered using dynamic clustering techniques. In the second step, content dissimilarity in the video is measured. The dissimilarity measure is found by taking temporal difference between frames. Then the video frames are processed using dynamic clustering techniques.

The first frame is selected as a cluster. For the sequential frames, if its minimum distance to all the existing clusters is larger than 5, the frame as a new cluster. Otherwise, it is merged with the nearest cluster and update the centroid of the cluster. The proposed key frames selection method is different from the clustering-only method in two ways. First, in this algorithm, the clustering is a post processing step that refines key frame results. Since the number of video frames is greatly reduced, this post processing is computationally efficient. Second, the parameter of the clustering process, 6, is automatically determined based on the queuing process and is adaptive to video content.

3. PROBLEMS AND MOTIVATION

The various methods of key frame extraction surveyed in the above section helps us to know that the keyframes are extracted by using either color or motion involved in the video. Some approaches like [1] and [4] tries to combine both dynamism and the visual content to decide the salient frames.

However for automatic keyframes extraction using higher level concepts that are in line with human visual and cognition system, both color and motion should be combined to develop a criteria based on which salient frames from the video can be extracted. Reference [5] appears to suggest one such method. In this paper we have implemented the methods suggested by [5] to extract keyframes from the video and to summarize it.

4. PROPOSED WORK

The proposed system basically extracts key frames and summarizes them, based on a visual attention model. The system helps achieving a meaningful video summary by bridging the semantic gap between the low level descriptors used by computer systems and high level concepts perceived by human based on neurobiological concepts of human perception of vision.

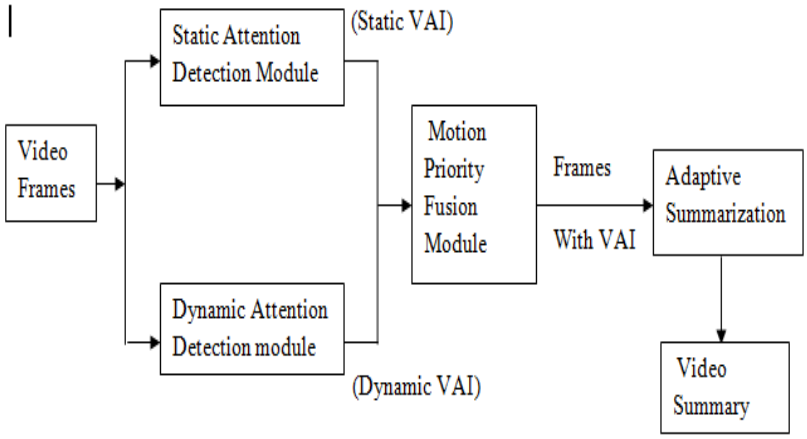

Figure 1. Simplified Architecture of the proposed System

From psychological studies it is known that, human pays more attention to moving objects than a static object. Not just activities, at times interesting static objects at the back ground which are visually appealing get the viewer's attention.So to find the frames that are visually attractive and meaningful, it becomes essential to model both dynamic and static attention each frame commands.

Both the static and dynamic attention are quantified using a descriptor called 'Visual Attention Index'. Motion intensity and orientation in a video are computed to model dynamic attention the video commands.VAI of dynamic attention is calculated based on coherence of orientation that is calculated using Gaussian kernel density estimation. Static attention of each frame is further calculated based on the Red-Green opponency and blue-Yellow opponency in each frame

4.1 Static Attention Detection Model

The static attention detection module quantifies color conspicuousness of each frame. This model is based on color opponent theory and works in LMS color space.

LMS is a color space represented by the response of the three types of cones of the human eye, named after their responsivity (sensitivity) at long, medium and short wavelengths. LMS color space, represents long, medium, and short light wavelengths (red, green, blue respectively). LMS values range from -1 to 1. So, black is (-1,-1,-1) and white is (1,1,1), and other colors lie in the

range from -1 to 1. sum of L+M is a measure of luminance. The cones in the human are classified into three classes. These three classes of cones are the short-wavelength sensitive (S-cones), middle-wavelength sensitive (M-cones) and long-wavelength sensitive (L-cones), and all have different but overlapping spectral sensitivities.

The color opponent process is a color theory that states that the human visual system interprets information about color by processing signals from cones and rods in an antagonistic manner. The opponent color theory suggests that there are three opponent channels: red versus green, blue versus yellow, and black versus white (achromatic and detects light-dark variation or luminance). [6] Responses to one color of an opponent channel are antagonistic to those to the other color. That is, opposite opponent color are never perceived together .There is no "greenish red" or "yellowish blue". We quantify color conspicuousness of each frame based on Red-Green and Blue-yellow opponency.Figure2 shows the operation flow of static attention detection module. The video which is to be summarized is split into individual frames and the individual frames are given as input into the static Attention detection model.

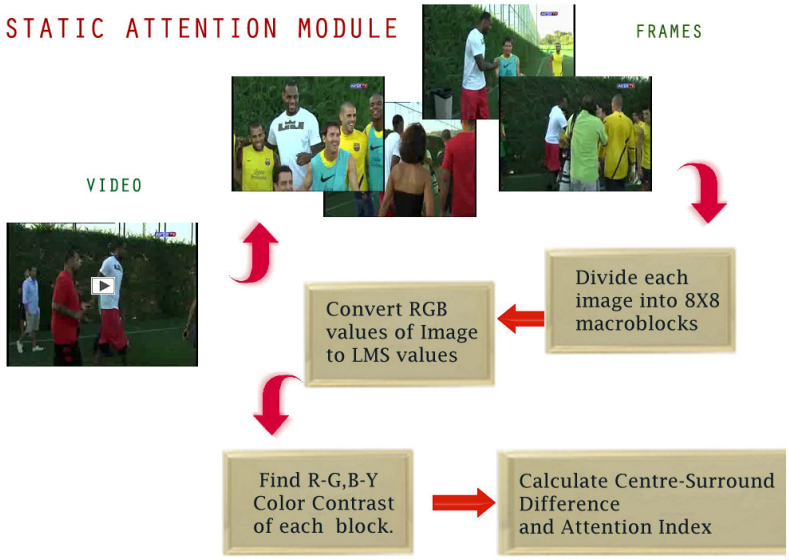

Figure 2. Operation flow of static attention Module

Each frame is split into 64 blocks that are further operated upon by the algorithm given below to get the Visual Attention Index of each frame.

The algorithm is as follows:

1 Divide each frame into smaller macro blocks in such a way that each frame is subdivided into 64 blocks.

2 Take one block, at a time. Convert the color space from RGB to LMS

3 RGB to LMS conversion is achieved in two steps
- 3.1 First RGB to XYZ conversion is carried out.
- 3.2 Followed by XYZ to LMS conversion

4 RGB to XYZ conversion is done using the transformation matrix

$$[x\ y\ z] = [R\ G\ B] \begin{bmatrix} 0.5767309 & 0.2973769 & 0.0270343 \\ 0.1855540 & 0.6273491 & 0.0706872 \\ 0.1881852 & 0.0752741 & 0.9911085 \end{bmatrix}$$

5 Similarly, XYZ to LMS is also achieved through the transformation matrix

$$[L\ M\ S] = [x\ y\ z] \begin{bmatrix} 0.7328 & -0.7036 & 0.0030 \\ 0.4296 & 1.6975 & 0.0136 \\ 0.1624 & 0.0061 & 0.9834 \end{bmatrix}$$

6 Calculate Red-Green, Blue-Yellow opponency using the following formulae,
Red-Green Opponency = (L-M) / (L+M)
Blue- yellow opponency = (s-0.5*(L+M))/(s+0.5*(L+M)

7 Calculate color contrast and intensity of each block.

8 Calculate center-surround difference of each block using the following formula

$$d(p_i,q)=(0.5\ |p_i\ (I)-q(I)|\ )+\ (0.5\ |p_i\ (H)-q(H)|$$

9 Calculate Static Visual attention index as follows:

$$A_s = (1/N)^* (\Sigma\ W_i\ .\ C_i)$$

Where W_i is Gaussian fall off weight.

10 Output the Static Visual Attention Index (VAI) for each frame.

Using the above algorithm, Static Visual Attention index of each frame is calculated. The process involves conversion of each frame from RGB Color space to LMS color model. Using the LMS values Centre-surround difference of each block is calculated and then visual attention index of entire frame is obtained

4.2. Dynamic Attention Detection Module

In the Dynamic attention detection model, the motion associated with each frame of the video is calculated. Motion estimation is done on the basis that the patterns corresponding to objects and background in a frame of video sequence move within the frame to form corresponding objects on the subsequent frame in [7]. The idea behind block matching is to divide the current frame into a matrix of 'macro blocks' that are then compared with corresponding block and its adjacent neighbors in the previous frame to create a vector that quantifies the movement of a macro block from one location to another in the previous frame. This movement calculated for all the macro blocks comprising a frame, constitutes the motion estimated in the current frame. The search area for a good macro block match is constrained up to *p* pixels on all fours sides of the corresponding macro block in previous frame. This '*p*' is called as the search parameter. The matching of one macro block with another is based on the output of a cost function. The macro block that results in the least cost is the one that matches the closest to current block. There are various cost functions mean absolute difference (MAD) and Mean squared error (MSE).

Based on the above idea the Motion attention detection module is built. Figure 3 shows operation flow of the Motion attention detection module .The calculation dynamic attention index by measuring motion associated with each frame is based on the following algorithm:

1. Divide the frame into 64 sub-blocks
2. Compare each block of a frame with the corresponding block in the next frame and calculate the motion vectors dx and dy by Block matching technique.
3. The motion intensity of each block is given by the formula

   Motion intensity $\Upsilon_i = \text{sqrt}\ (dx_i^2 + dy_i^2)$
4. The orientation of each block is calculated

   Motion orientation $\Theta_i = \arctan(dy_i / dx_i)$
5. Then, orientation histogram is built and the motion attention index is calculated from the following formula:

   $A_T(i) = 1 - (v(b(i))/(\Sigma v(j)))$
6. Motion Attention of the block is given by

   $A_T(i) = \gamma_i A_T(i)$

$A_T(i)$ gives the Motion attention Index of frame i. Motion with high intensity attracts more attention. Frames with larger attention index are given much importance.

.

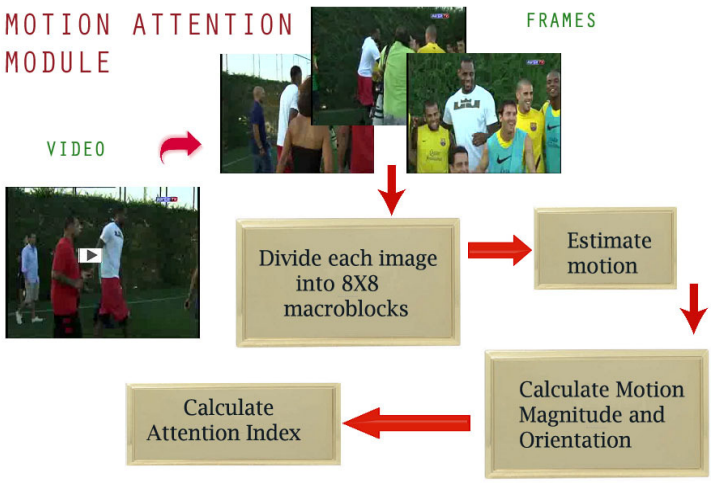

Figure 3. Operation flow of Motion Attention detection Model

4.3 Motion Priority Fusion Module

This module fuses the weighted static and dynamic attention indexes of each frame calculated by following aforesaid modules. The design of the module is given below in Figure 4.

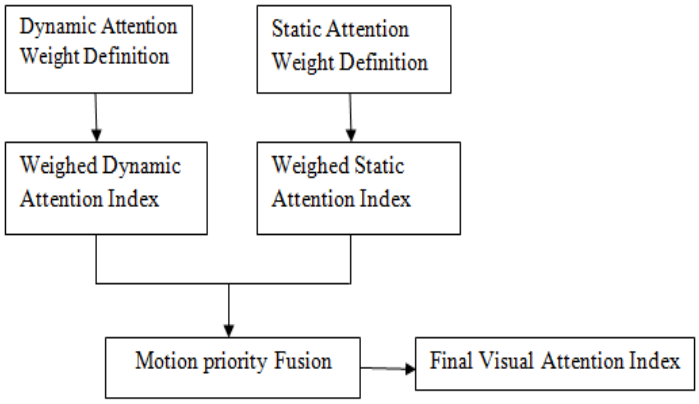

Figure 4. Architecture of Motion Priority Fusion Module

The algorithm applied in Motion priority fusion Module is as follows:

1 Input the Static and Dynamic visual Attention indexes of each frame

2 Define weights of dynamic attention and static attention for each frame using following formulae.

$W_T = A'_T.\exp(1-A'_T)$

$W_s = 1-W_T$

where $A'_T = \text{Max}(A_T) - \text{Mean}(A_T)$

3 Calculate final Visual Attention Index

$VAI = W_s A_s + W_T A_T$

4 Output the Total Visual Attention Index

4.4.Adaptive Summarization Module

The video is summarized with the important frames depending on the importance of the frame determined with the Visual Attention Index of each frame using Adaptive summarization algorithm. The simplified architecture of the Adaptive summarization module designed and implemented in this work is shown in figure 5

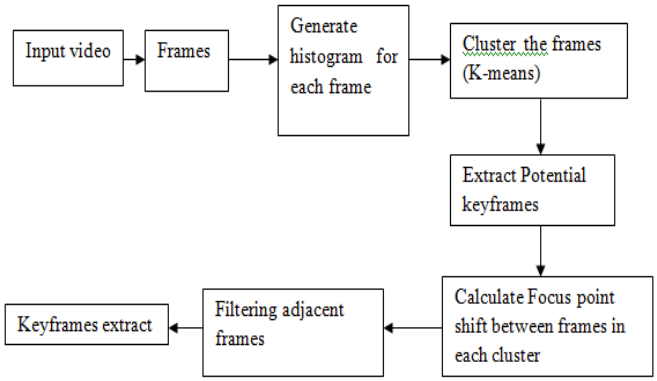

Figure 5. System Architecture of Adaptive summarization Module

The adaptive summarization algorithm when applied over all the frames of the video gives the set of frames which have high Visual attention index and do not contain redundant information, which could be used in the Video Summary.

1 Input the video and get all the frames
2 Calculate Color histogram for each frame
3 Cluster the frames using K-Means clustering algorithm based on Euclidean distance.
4 Select Representative keyframes (frames having high VAI) from each cluster
5 Compare potential keyframes by calculating focus point shift between them.
$$D_{ij} = | \Sigma (A_n^i - A_n^j)|$$
6 Extract 20 frames that have first 20 large D_{ij} values.
7 Filter out redundant frames from the selected potential keyframes using the inequality
$$D_{key} > D_{ave} + \delta D_{div} \quad (\delta = 1.5 \text{ here})$$
where D_{ave} is the average VAI difference of the keyframe, D_{div} is the standard deviation.

Thus by using the above algorithm, the most representative keyframes are extracted from the pool of video frames.

5. EXPERIMENTAL RESULTS

The system proposed above is implemented using Matlab2011 and tested using random videos downloaded from internet. The details of the videos used for testing the system and the no. of keyframes extracted are given in Table 1. The Video summaries are obtained and are subjected to user based evaluation first i.e. the videos along with the video summaries are presented to the users. The summary of the videos produced are presented before 20 users, who are later made to watch the video and their feedback on the effectiveness of the summary is obtained. The feedback is presented in the Table 2.

The evaluation is based on three main criteria. Content Coverage refers to the extent to which the summary effectively conveys the content of the video. Presentation indicates how effective the presentation is. It refers to the entertainment quotient of the summary. Total effectiveness that is rated out of 10 is based on total satisfaction provided by the summary.

Table 1. Video Details

Name of the Video	**URL**	**Total Number of Frames**	**Total No. of Keyframes Extracted**
Serj-Orders-a-coffee	http://www.youtube.com/watch?v=u06IIqYY0iU	447	10
I'm normally not a praying Man	http://www.youtube.com/watch?v=MILArKLKUEk	219	5
Minecraft in a nutshell	http://www.youtube.com/watch?v=7Dild9s2c94	135	7
Family-guy-Im beautiful	http://www.youtube.com/watch?v=VTenT9PtM2w	120	7
Late-for-work	http://www.youtube.com/watch?v=kfchvCyHmsc	193	10
Compliment Taser	http://www.youtube.com/watch?v=q081Q8CZUnk	240	9

Table 2. User based Evaluation Results

Name of the video	Content coverage	Presentation	Total Effectiveness (Rated out of 10)
Serj-Orders-a-coffee	Very Good	Fair	8
I'm normally not a praying Man	Good	Fair	7
Minecraft in a nutshell	Very Good	Good	8
Family-guy-Im beautiful	Good	Good	8
Late-for-work	Very Good	Very good	8
Compliment Taser	Good	Good	7

The summaries produced by the system for the videos listed above are shown in the following Figure 6

Figure 6. Summaries of Test Videos

The Figure 6 shows summaries of the videos conveying the content of the video through the key images obtained using the system explained. The user will be able to understand the content of the video by looking at the summary without watching the video fully.

Precision and Recall of the video summaries produced by the system are presented in the Table 3 below. Precision is the fraction of right key frames of all keyframes extracted and recall is the fraction of right key frames of all keyframes present. Here, the system is set in such a way that each video has maximum representative 10 keyframes in the summary.

Table 3. Precision and Recall

Video No.	**Name of the Video**	**Precision**
1	Serj-Orders-a-coffee	0.778
2	I'm normally not a praying Man	0.6
3	Mine Craft in a nut shell	0.714
4	Family-guy Im not beautiful	0.856
5	Late-for-work	1
6	Compliment Taser	0.8

The Precision and Recall of all the videos are represented in the Figure 7 below:

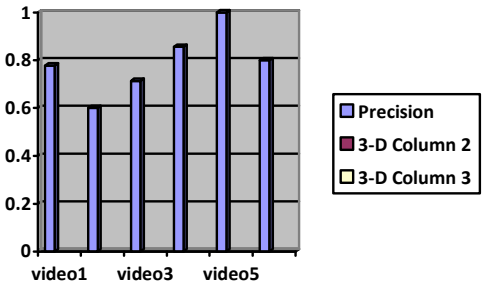

Figure 7. Precision Plot

Precision indicates how many keyframes are correct out of all the keyframes extracted. In video 1,two frames out of nine are redundant. So the precision is

5. CONCLUSION AND FUTURE ENHANCEMENTS

Thus the work undertaken helps us understanding that key frames extracted by the techniques that take both color and motion information into account are more relevant than the ones that use either color or motion information only to extract the keyframes. The effectiveness of the keyframes extracted using both color and motion criteria is more when compared to that of the ones extracted by other methods using either only color or only motion based techniques. Hence to bring about a summarization which is in line with the high level human intuition both the color conspicuousness and motion involved in a scene should be taken into account. From the user response it is found that, our system produces a very effective summary. However in the system proposed, the number of keyframes in the summary is independent of the duration and dynamicity involved in the video. Summarization is efficient if the video is segmented using a suitable method that takes both color and motion into account and then summarized by allocating appropriate number of keyframes to each shot depending on the importance of the segment. So the effectiveness can be further improved by determining the sufficient amount of keyframes required to summarize the video based on the dynamicity of the video. Also the system can be enhanced by allocating more keyframes representing shots that involve more action. This requires an efficient shot boundary detection method. These are the enhancements that could be added to the system in future. The system has been tested with web videos of very short duration(less than one minute) that have no major change in background. However, long videos with multiple shots and scenes will need appropriate shot segmentation step.

REFERENCES

[1] Zhonghua Sun, Kebin Jia ,Hexin Chen , "Video Key Frame Extraction Based on Spatial-temporal Color Distribution",International Conference on Intelligent information hiding and multimedia signal processing, August 15-18,2008, Harbin, China.

[2] Nuno Vasconcelos ,Andrew Lippman, "A Spatiotemporal Motion Model for Video Summarization", IEEE computer society conference on computer vision and pattern recognition, pp. 361-366, 25-25 June, 1998

[3] Ajay Divakaran, Regunathan Radhakrishnan and Kadir A. Peker "Motion Activity-Based Extraction Of Key-Frames From Video Shots", International Conference on Image processing, Sep. 22-25,2002, Volume 3, USA.

[4] Sathish kumar l.Varma and Sanjay N.Talbar, "Video summarization using dynamic threshold", International conference on Emerging trends in Engineering and Technology, Oct. 14-16, 2010.

[5] Jiang Pen And Qin Xiao Lin, "Key Frame Based Video Summary Using Visual Attention Clues" , IEEE Transactions On Multimedia 2010,pp. 64-73, Volume 17, Number 2, April.

[6] Michael Foster (1891). A Text-book of physiology. Lea Bros. & Co. p. 921.

[7] Aroh Barjatya, "Block Matching Algorithms for Motion Estimation" IEEE, DIP 6620 Spring, Paper (2004) Final Project Paper, pp.1-6.

[8] P.Geetha and Vasumathi Narayanan, "A survey of Content based video Retrieval", Journal of Computer Science, pp. 474-486, Vol 4 (6), 2008.