Video Shot Detection using Saliency Measure

Amudha J
Associate Professor,
Department of CSE,
Amrita School of Engineering,
Karnataka, India,

Radha D
Assistant Professor,
Department of CSE,
Amrita School of Engineering,
Karnataka, India,

Naresh Kumar P
PG Scholar,
Department of CSE,
Amrita School of Engineering,
Karnataka, India,

ABSTRACT
Video shot boundary is an early stage of content based video analysis and is fundamental to any kind to of video application. The increased availability and usage of online digital video has created a need for automated video content analysis techniques. Major bottle neck that limits a wider use of digital video is the ability of quickly finding desired information from a huge database. Manual indexing and annotating the video material are both computationally expensive and time consuming. In this paper we design a novel approach for shot boundary detection using visual attention model by comparing the saliency measures. The approach is robust to a wide range of digital effects with low computational complexity

Keywords
Shot detection, saliency measure and visual attention model.

1. INTRODUCTION
The latest developments in multimedia technology, combined with a considerable growth in computer performance and expansion of the Internet, have provided people with access to a tremendous amount of video information. Video applications, expanding at a considerable rate, have initiated an increasing demand for innovative technologies and tools to index, browse, and retrieve video data efficiently. Developing for automatic indexing, retrieval, and management of video, content-based video retrieval has become the subject of much research throughout the last decade. Structural analysis of video is a fundamental stage in analyzing video content and developing techniques for efficient access, classification, retrieval, and browsing of vast video databases. Among the several structural levels (i.e., frame, shot, scene, etc), shot level organization has been deemed suitable for browsing and content-based retrieval.

Due to the increase in the requirements of automation of shot boundary detection processes in distance learning, telemedicine, interactive television, digital libraries, multimedia news, video restoration and geographical information system, many shot detection methods on computational techniques are available and some techniques are combination of spatial and frequency domain. Every technique has its own pros and cons. false positives and multiple detection of shot boundaries during flashing effects causes decline in performance. Increase in performance is proportional to the complexity involved in the system. For any video application, the various structural levels, like shot histogram methods can be achieved by block-matching [13] methods.

Another method proposed by Zabih et al, performs a Shot Boundary Detection (SBD)[9] that uses an edge change ratio(ECR) method. In this method a canny edge detector is detection, key frame extraction [1], annotation [2] and object recognition [3] are treated as different components. This leads to a massive computational effort for the application to be developed which can be improvised by reuse of the features extracted by a visual attention model to the various structural levels. So, to decrease the computational complexity for detecting shot boundaries which are the fundamental step for any video processing application, we are implementing video shot detection using visual attention by comparing measures obtained from the saliency regions of frames. In this paper an attempt is made to study the performance of shot boundary detection using saliency measure and the results are found to be satisfactory

The rest of the paper is organized as follows. A brief survey of previous work is presented in section 2. Section 3 explains the architecture of the system, the shot boundary detection classification is explained detail in section 4. Experimental results are presented in section 5 and conclusion and future scope are given in section 6.

2. RELATED WORK
In this paper, a detailed survey of various shot detection methods has been discussed and the significance of using visual attention system to video shot detection has been highlighted. Previous works on video-shot detection can be categorized under any of the approaches

2.1 Statistical based techniques
There have been intensive researches on the Statistical approach. Various techniques based on pixel [4,5,6], histogram [7, 8], edge [9], motion vectors [8], statistical measure [10], sliding window method [11], graph partitioning [12] have been proposed.

The pixel-based method is sensitive to local and global movements, which can be dealt by image smoothening method [6]. Another way of comparing the mean and standard deviation of the pixels in regions [10] is comparatively slow due to the usage of the statistical formulas whereas it is tolerant to noise.

Color histogram based methods confine to the ratio of different color components and is however invariant to local and global movements. The major disadvantage is it is unable to detect differences existing in the shot within the same scene. A better tradeoff between pixel and color

used to get accurate results than histogram based methods. However this is less sensitive to motion compared to chromatic scaling. Sliding window method [11] studies the discontinuity values more closely to improve the robustness to camera/object motions and reduces the threshold selection

problem to an extent.Graph partitioning [12] methods are able to detect efficiently abrupt cuts and all types of gradual transitions, such as dissolves, fade and wipes with very high accuracy

Different methods like Luminance Histograms and Hough Transform [14] combination works well in detecting shot cuts and gradual transitions in MPEG compressed video sequences. These methods accomplish the gradual shot transition detection better.

2.2 Computational techniques

Independent Component Analysis (ICA) [15] detects a data driven feature which shows that it can effectively detect both abrupt transitions and gradual transitions

Another scheme which is based on rough-fuzzy set [16, 17, 18] performs better in shot boundary detection. In this scheme, 12 candidate features classified under 5 types are usually extracted. There are chances of false detections resulted by irregular camera operations during gradual transitions and lot of flash effects in a shot. An algorithm, Principle Coordinate System [19] performs well for shot boundaries in gradual changes, illumination changes or camera motion.

2.3 Combination of spatial and frequency domain based Techniques

TRECVID 2006 [20] uses the correlation to compute successive frame similarities, wavelet transforms and Support Vector Machine (SVM) classification which increases the accuracy and performance even in various effects like wobbly camera, fire, explosion and synthetic screen split manipulations.

For detecting abrupt changes in hard cut transitions, a probability function is analysed which is very complex using QR- Decomposition and Gaussian Transition Detection algorithm[21] to give better performance than most of the methods.

2.4 Visual attention based techniques

Not much work has been done in applying visual attention models to shot boundary detection. The Structural Similarity (SSIM) [1] combines three components, namely luminance-l, contrast-c and structure-s. Saliency map is obtained by purely computational technique and used as weighting function in calculation of structural similarity index. The shots are determined by applying a threshold to the structural similarity index

Major advantage of the algorithm is that it is robust to dissolving digital video effects used during shot transition. All the above techniques have its own pros and cons involved in the model. However this paper explores the use of visual attention model features for SBD which haven't investigated much and can be further reused for higher level tasks like key frame extraction [1], annotation [2] and object recognition [3]. The proposed architecture explains the use of visual attention model for shot boundary detection for both abrupt and gradual transition which hasn't been explored yet. The results prove that the system is able to significantly identify gradual transition patterns efficiently than many other approaches.

3. ARCHITECTURE

The proposed architecture identifies the shots in a given video sequence as shown in Figure 1.

First the video sequence is divided into frames and each frame is given to the visual attention model which outputs a saliency map. All the consecutive saliency maps from the frames are compared with two statistical metrics mean and variance. Shots are identified based on the threshold values and further analysis on the patterns of gradual transitions has been studied.

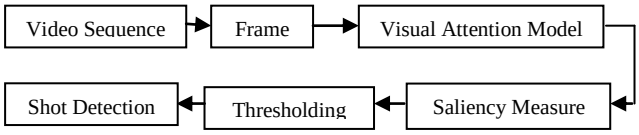


Figure 1: Work flow of Shot Boundary Detection

3.1 Visual Attention Model

Visual attention is an important biological mechanism which can rapidly help human to capture the interested region within eye view and filter out the minor part of image. By means of visual attention, checking for every detail in image is unnecessary due to the property of selective processing.

A Visual attention system [10] approach is a bottom-up approach where two different features are computed like colour, intensity. For each feature, the saliencies are computed on different scales and for different feature types e.g. red, green, blue, yellow and intensity. Thereafter, the maps are fused step-by-step, thereby strengthening important aspects and ignoring others.

For each feature, we first compute an image pyramid from which we compute scale maps. These maps are fused into feature maps representing different feature types and these again are combined to a single saliency map. Figure 2 shows the Visual attention model used for shot boundary detection

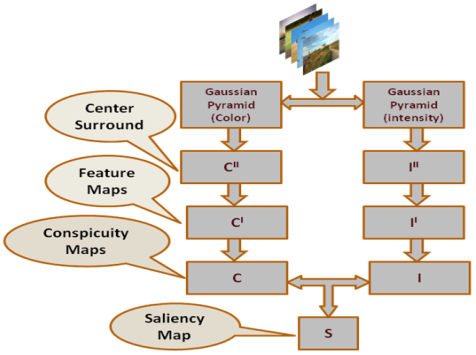


Figure 2: Visual Attention Model used for SBD

The Gaussian pyramid is a simple structure for representing images at more than one resolution. The base of the pyramid contains a high-resolution representation of the image frame being processed; the apex contains a low-resolution approximation. The first level of the image pyramid is the original image itself

$$P(I)_1 = I$$

The next level of the image pyramid

$$P(I)_{n+1} = S \downarrow G_\sigma (P(I_n)) \quad (1)$$

Where,

$P(I)_n$ – Pyramid of the image I at level n.
$S\downarrow$ – Down sampling
G_σ –Gaussian filter with standard deviation σ

3.1.1 Construction of pyramids for different channels

With r, g and b being red, green and blue channels that have been extracted from the input frames, the intensity of each frame is calculated as shown in eq. (2)

$$I=(r + g + b)/3 \qquad (2)$$

The colour channels r, g, b are normalized by the intensity I in order to decouple hue from the each image I. Each input frame I is then sub-sampled into a Gaussian pyramid and each pyramid level is decomposed into channels for red (R), green (G), blue (B), yellow (Y), intensity (I), where the value of R, G, B, Y are shown in equations (3),(4),(5) & (6).

$$R = r- (g + b) / 2 \qquad (3)$$
$$G = g- (r +b) / 2 \qquad (4)$$
$$B = b - (r + g) / 2 \qquad (5)$$
$$Y = r + g - 2 (| r - g | + b) \qquad (6)$$

(Negative values are set to zero)

The channels like Red, Green, Blue, Yellow and Intensity channels are constructed for each level of pyramid. In figure 3, feature map of colour and intensity are shown.

3.1.2 Construction of feature maps from the channels

Each feature is computed in a centre-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast in a given feature rather than the amplitude of the channel. Thus the "Centre-surround" differences detect the spatial discontinuities for each feature. Centre-Surround operations are implemented in the model as across scale difference between a fine and coarse scale for a given feature. The centre of the receptive feature corresponds to the pixels at the level c = 2 in the pyramid and the surround corresponds to the pixels at the level s=c+ l where l=1, 2. Therefore, two combinations are possible i.e., one centre with two surrounds. Hence we compute three feature maps for colour and intensity. Across scale difference between two maps, denoted "Θ" below, is obtained by interpolation to the finer scale and point-by-point subtraction.

The intensity feature map encodes for the modulus of the image luminance contrast. That is the absolute value of difference of intensity between the centre and the surround is shown in eq. (7)

$$I(c, s) = |I(c) \Theta I(s)| \qquad (7)$$

The colour feature map RG and BY corresponds to the double-opponency cells in primary visual cortex and are then the spatial contrast centre surround differences across the normalized colour channels. Each of the three-red/green, feature map is created by first computing (red-green) at the centre, then subtracting (green-red) from the surround and finally outputting the absolute value. Accordingly maps RG(c, s) are created in the model to simultaneously account for red/green and green/red double opponency. RG and BY centre surround differences are shown in eq. (8) & (9).

$$RG (c, s) = | (R(c) – G(c)) \Theta (G(s) – R(s))| \qquad (8)$$
$$BY (c, s) = | (B(c) – Y(c)) \Theta (Y(s) – B(s)) | \qquad (9)$$

3.1.3 Conspicuity Maps

The feature maps are then summed across scales combined into two conspicuity maps one for intensity, and colour .The conspicuity maps for colour and intensity shown in eq. (10) & (11)

$$C = (RG(c, s) + BY(c, s)) \qquad (10)$$
$$I = I(c, s) \qquad (11)$$

3.1.4 Saliency Map

The two conspicuity maps are summed into the saliency map 'S' is shown in eq (12).

$$S=((1/2)*((I)+(C)) \qquad (12)$$

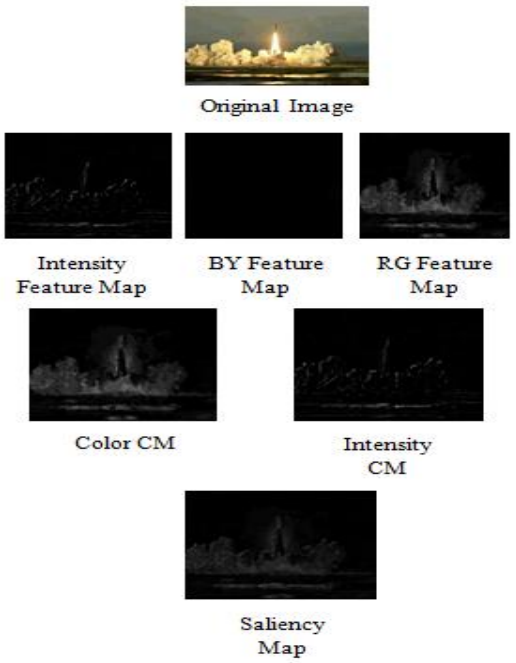


Figure 3: Stages of the visual attention model for an image frame of Ksc_launch.avi

In the same way the saliency maps are computed for all the frames and respective mean and variance values are calculated for further process of shot boundary detection.The figure 3 shows the various stages of the visual attention model on a frame taken from Ksc_launch.avi.

3.2 Analysing shot behaviour

3.2.1 Types of shot transitions

Shot boundaries can be broadly classified as abrupt and gradual transition and further classification as shown in figure.4

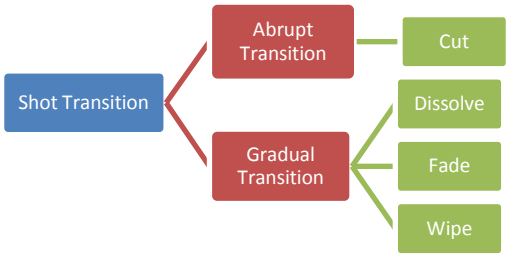


Figure 4: Classification of Shot Transitions

Abrupt Transition:

Cut: It simply means replacing one shot instantly with the next. In video editing and live switching, cuts are fast and efficient. Once a scene has been established, cuts are the best way to keep the action rolling at a good pace as shown in Figure 5 (a). Other types of transition can slow the pace or even be distracting.

Gradual Transition:

Gradual Transitions are shot transitions which occur over multiple frames resulting in smooth transitions from one shot to other

Dissolve: A gradual fade from one shot to the next is known as a crossfade, mix or dissolve. Crossfade have a slower, more relaxed feel than a cut as shown in Figure 5(b).
The speed of the crossfade transition can vary between a few frames (for relatively fast-paced content) to several seconds. Slow or incomplete crossfade can also be used to create layered video effects.

Fade: A video fade is when a shot gradually fades to (or from) a single color, usually black or white. A fade is different to a crossfade, which is a transition directly between two shots rather than one shot to a color as shown in Figure 5 (d) and 5 (f).

Wipe: In a video wipe, one shot is progressively replaced by another shot in a geometric pattern. There are many types of wipe, from straight lines to complex shapes. Wipes often have a colored border to help distinguish the two shots during the transition. Split screens often use a wipe. A horizontal line wipes from left or right into the middle of frame, revealing the new shot in that half. Wipes are a good way to show changing location or viewpoint as shown in Figure 5(h).

Further classification of shots into different classes are analysed and studied with the different transition patterns of mean and variance as shown in figure 5(c),(e),(g),(i)

4. SHOT BOUNDARY DETECTION

To find shot boundaries in a video file we require a threshold to differentiate the frames which are similar and dissimilar frames. If it is dissimilar then there should be a transistion or special effects in the sequence of frames. Those which has more threshold difference in consecutive frames are poped out as shot boundaries.

4.1 Thresholding

For detetecting shot boundary with the metrics mean and variance, three threshold levels are considered. These thresholds are fixed by experimental observation of different shots. F_i and F_{i+1} is the i^{th} and $(i+1)^{th}$ frames of the given video file, M and V are Mean and Variance respectively.

There are three different cases of threshold levels are shown with the graphs for the Ksc_launch Video

CASE 1

$M(f_i) > 2$ && $((V(f_i) > (V(f_{i+1})+(V(f_i)/4))$ | | $(V(f_{i+1}) > (V(f_i)+(V(f_{i+1})/4)))$

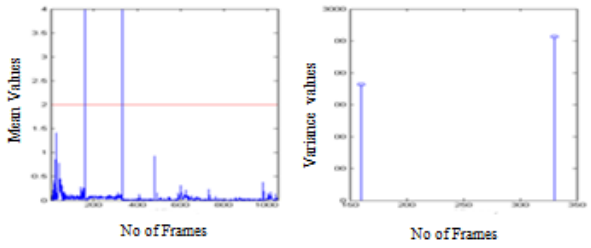


Figure 6: Shot boundary detection with threshold value greater than 2

In the Figure.6, there are two frames above the threshold value greater than 2 and both the detected frames are satisfing the variance condition.

CASE 2:

$M(f_i) >1$ && $((V(f_i) > (V(f_{i+1})+(V(f_i)/3))$ | | $(V(f_{i+1}) > (V(f_i)+(V(f_{i+1})/3)))$

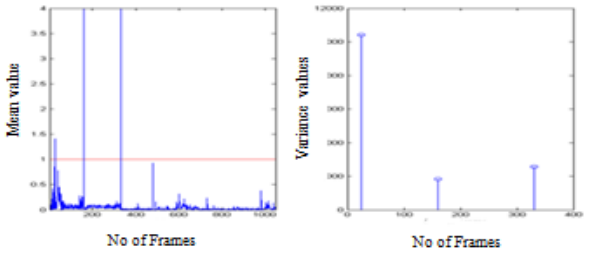


Figure 7: shot boundary detection with threshold value greater than 1

In the figure.7, there are three frames above the threshold value greater than 1 but the detected frames are only one because the variance values of either side of the detected frame is not satisfied the condition

CASE 3

$M(f_i) > 0.5$ && $((V(f_i) > (V(f_{i+1})+(V(f_i)/2))$ | | $(V(f_{i+1}) > (V(f_i)+(V(f_{i+1})/2)))$

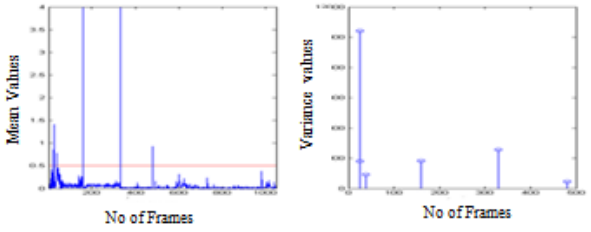
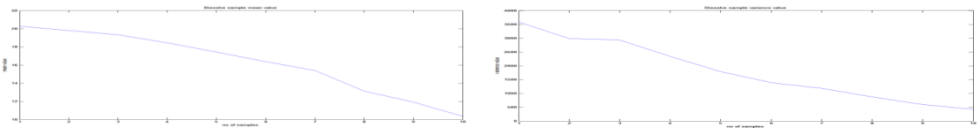


Figure.8: shot boundary detection with threshold value greater than 0.5

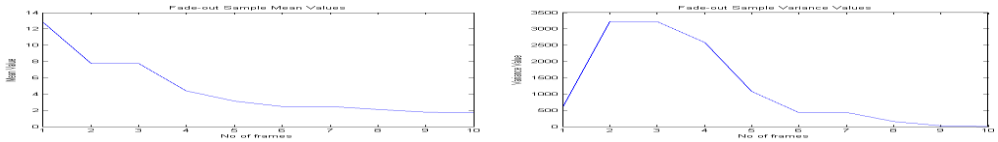
(a) Example of successive abrupt frames from Ksc_launch.avi

(b) Example of successive dissolve frames from GLGC.avi



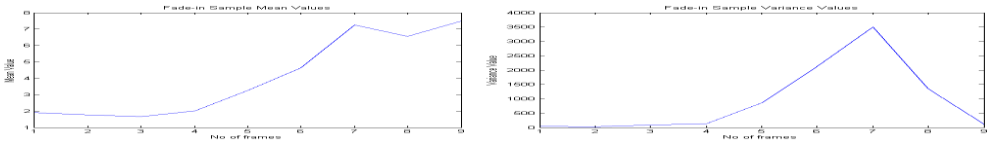
(c) Mean and Variance of frames shown in (b)

(d) Example of successive fade-out frames from you sang to me.avi



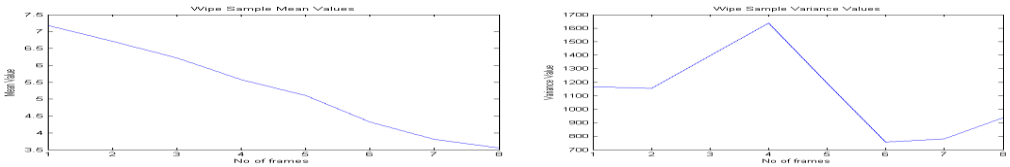
(e) Mean and Variance of frames shown in (d)

(f) Example of successive fade-in frames from you sang to me.avi



(g) Mean and Variance of frames shown in (f)

(h) Example of successive wipes frames from space.avi



(i) Mean and variance of frames shown in (h)

Figure 5: Various Shot transitions

In the Figure.8, there are more than seven frames which are above the threshold value greater than 0.5 but the detected frames are only three because the variance the detected frame is not satisfied the condition values of either side of the detected frame is not satisfied the condition.

These are the shot boundaries without filtering the multiple detection of frames within the shot. Filtering the multiple detection of frames within the frame rate and these detected frames are tested for further classification and to which class it belongs or it is a false detection.

4.2 Filtering shot boundaries

Given the sequence of shot boundary frame numbers, from which the multiple frame numbers within the frame rate need to be filterd for next level of classification or to choose which frames to be considered as shot boundary frames by further analysis within the frame rate need to be filterd for next level of classification or to choose which frames to be considered as shot boundary frames by further analysis

$S_1,S_2,S_3,S_4.....\underbrace{S_{230},S_{242},S_{249},S_{254},S_{260}}, S_{290}$ .....

$S_1,S_2,S_3,S_4.....S_{230}, S_{290}.....$

Shots filtered from S230 to S260, by assuming the frame rate is 30fps, usually the multiple detection within the frame rate will happen only when gradual transitions occur.

5. EXPERIMENTAL RESULTS

For the experimental analysis a dataset with 5 videos comprising of different transitions is considered as shown in Table 1. The shot detection results obtained on the test cases of 12061 frames of AVI testing material with 162 shots and their transitions are shown in Table 2

Table: 1 List of Transitions in various video files

S No	File Name	Cut Transitions	Gradual Transitions		
			Fades	Dissolves	Wipes
1	Shuttle-flip	x	x	X	x
2	Ksc_launch	3	x	X	x
3	Space	x	x	X	1
4	You sang tome	1	7	6	x
5	GLCN	132	x	7	x

The evaluation of the proposed shot boundary detection algorithm, the precision and a recall rates were calculated. The precision rate indicates the percent of correctly detected shot boundaries among all the detected shot boundaries and the recall rate reflects the percent of correctly detected shot boundaries among all the shot boundaries existed in the video.

$$\text{Recall} = \frac{\text{Hit}}{\text{Hit} + \text{Miss}}$$

$$\text{Precision} = \frac{\text{Hit}}{\text{Hit} + \text{False}}$$

Hit denotes the number of correct detected shot boundaries, Miss is the number of missed shot boundaries and
False is the number of falsely detected shot boundaries. Table 3 list the perfomance of the proposed algorithm for the test video files and result of the Ksc_Launch video file is shown in Figure.9

Table: 2 List of shots detected in various video files

S.NO	Video file Name	Actual Shots	Total Shots detected	True Shot detections	False Shot detections	False positives
1	Shuttle-flip	1	1	1	X	X
2	Ksc launch	4	5	4	1	X
3	Space	2	2	2	X	X
4	You sang to me	15	10	9	1	6
5	GLCN	140	100	83	17	X
	TOTAL	162	118	99	19	6

Table: 3 Precision and Recall values for video files

S.No	Video file Name	Precision	Recall
1	Shuttle-flip.avi	1.0	1.0
2	Ksc_launch.avi	0.8	1.0
3	Space.avi	1.0	1.0
4	You sang tome.avi	0.90	0.6
5	GLCN.avi	0.83	0.59
	TOTAL	0.84	0.61

In first and third cases, precision and recall is maximum due to the absence of gradual transitions but in second case the recall is maximum and precision is 0.8 due to one false detection. False detection occurred because of blurred frame in the sequence. Recall is moderate for *you sang to me.avi* and *GLCN.avi* files in fourth and fifth cases. because of the multiple detections are identified, when the gradual transitions takes place within the frame rate and all the detected boundaries are almost true other than the multiple detections within the frame rate. So, that the precision rate is high for the same files

In the previous shot detection using structural similarity [32], the precision and recall are high but they have tested with only on neurosurgical video but where as in our project we have taken different cases of videos, some of them have only cuts and others are combination of gradual transitions, flashing effects and zooming.

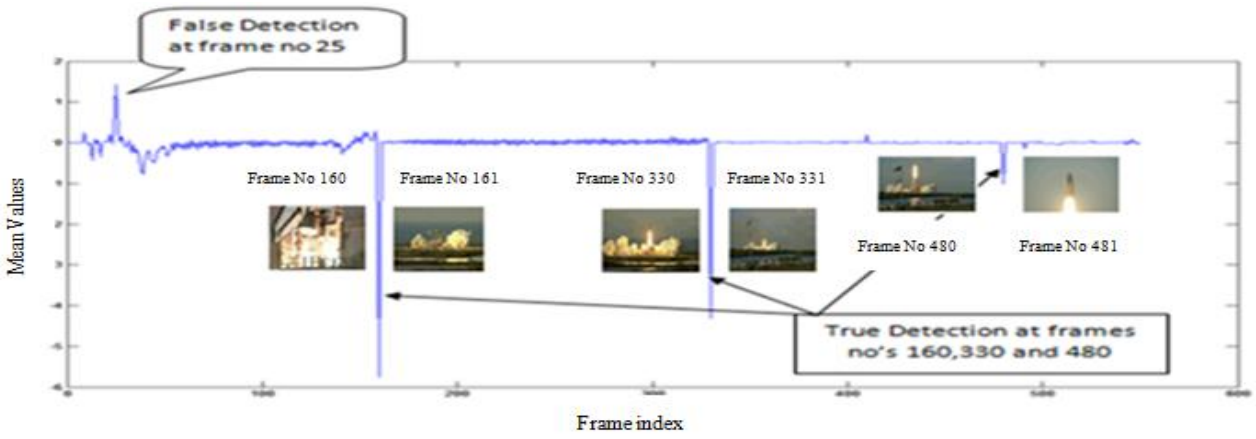


Figure 9: Results of the Ksc_launch video file

6. CONCLUSION AND FUTURE SCOPE

The proposed model for video shot detection using saliency measure is a novel approach for detecting shot boundaries. Visual attention model is used to decrease the complexity for obtaining saliency map instead of using computational methods, and also only two metrics are used for calculating shot boundaries. Further classification of shots into different classes based on the patterns can be studied and analysed. Analysis of patterns can be done on machine learning tool for classifying them to specific digital transition.

7. REFERENCES

[1] Engine Mendi, Coskun Bayrak, "Shot Boundary Detection and key Frame Extraction using Salient Region Detection and Structural similarity", ACMSE'10, April 15-17, 2010, Oxford, MS ,USA.

[2] Amudha J, K.P. Soman and Vasanth K (2008) "Video Annotation. Using Saliency", International conference on Image processing. Computer vision and Pattern Recognition" Vol 1, pp.191-195.

[3] Amudha J, K.P. Soman and Y. Kiran (2011), "Feature Selection in Top Down Visual Attention Model with WEKA", International Journal of Computer application, Foundation of Computer Sciences, USA, Vol.24, No.4, pp. 38-43

[4] A. Hampapur, R. Jain, and T. Weymouth, ''Digital video segmentation,'' *Proc. ACM Multimedia 94*, pp. 357–364, San Francisco, CA (1994).

[5] B. Shahraray, ''Scene change detection and content-based sampling of video sequences,'' in *Digital Video Compression: Algorithms and Technologies, Proc. SPIE* 2419, 2–13 (1995).

[6] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, ''Automatic partitioning of full-motion video,'' *Multi-media Systems*, 10–28, (1993).

[7] A. Nagasaka and Y. Tanaka, ''Automatic video indexing and full video search for object appearances,'' in *Visual Database Systems II*, E. Knuth and L. Wegner, Eds., pp. 113–127, Elsevier Science Publishers (1992).

[8] H. Ueda, T. Miyatake, and S. Yoshizawa, ''IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system,'' *Proc. CHI. 1991*, pp. 343–350 ACM, New York (1991).

[9] R. Zabih, J. Miller, and K. Mai, ''A feature-based algorithm for detecting and classifying scene breaks,'' *Proc. ACM Multimedia 95*, pp. 189–200, San Francisco, CA (1995).

[10] R. Kasturi and R. Jain, ''Dynamic vision,'' in *Computer Vision: Principles*, R. Kasturi and R. Jain, Eds., IEEE Computer Society Press, Washington (1991).

[11] Shan Li, Moon-Chuen Lee, 2005. An improved sliding window method for shot change detection. Proceeding of the 7th IASTED International Conference Signal and Image Processing, Aug. 15-17, IIonolulu, IIawaii, USA, pp: 464-468

[12] Cernekova.Z, N. Nikolaidis and I. Pitas, 2006. Temporal video segmentation by graph partitioning. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, May 14-19, 2: 209-212, Doi: 10.1109/ ICASSP. 2006.

[13] R. Kasturi and R. Jain, Dynamic Vision, *Computer Vision: Principles*, R Kasturi and R. Jain (Eds.), IEEE Computer Society Press, Washington DC, 1991, pp. 469-480.

[14] Edmundo Saez, José I. Benavides, Nicolas Guil, "Combining Luminance and Edge based Metrics for Robust Temporal Video segmentation", International Conference on Image Processing (ICIP), IEEE 2004.

[15]Jian Zhou, Xiao-Ping Zhang, "Video Shot Boundary Detection Using Independent Component Analysis", IEEE 2005

[16] Gao X. and Tang. X , 2002. Unsupervised video shot segmentation and model-free anchorperson detection for news video story parsing. IEEE Trans. Circuits Syst. Video Technol., 12: pp.765-776

[17] Gao X. and X. Tang, 2000. Automatic parsing of news video based on cluster analysis. In Proceedings of 2000 Asia Pacific Conference on Multimedia Technology and applications, Kaohsiung, Taiwai, China, Dec.17-19, pp: 17-19.

[18] Han Bing, Gao Xin-bo, Ji Hong-bing, 2003. An efficient algorithm of gradual transition for shot boundary segmentation. 3rd International Symposium on Multispectral Image Processing and Pattern recognition (MIPPR'03), Beijing, 9:956-961.

[19] Alper YILMAZ, Mubarak Ali Shah, "Shot Detection Using Principal Coordinate System" University of Central Florida, USA, 2010.

[20] Nithya Manickam, Neela Sawant, Aman Parnami, Srikanth.L., Sarath chandran, "TRECVID 2006", Indian Institute of Technology, Bombay.

[21] Ali Amiri and Mahmood Fathy, "VideoShot Detection Using QR-ecomposition and Gaussian Transition Detection", EURASIP journal on Advances in Signal Processing, Volume 2009, Article ID 509438, doi:10. 1155/2009/509438.