Shot Boundary Detection Using Structural SIMilarity Index

Srilakshmi B., Sandeep R.
Cambridge Institute of Technology, K. R Puram, Bengaluru, India

Abstract - Due to easy availability of video capturing and editing tools, low cost storage media devices and broadband data connection, the digital videos are becoming widely used. However, the increasing availability of digital video has not been accompanied by an increase in its ease of accessibility. Searching for videos with desired content from such a large collection is tedious and time consuming. One of the possible solution is, to make the videos available in an elementary form called shots in the initial step. A shot is defined as an unbroken sequence of frames taken from camera. The accurate shot boundary detection helps to organize the video contents into meaningful parts. Therefore, the Shot Boundary Detection (SBD) is the first step towards video indexing and content based video management.

Shot transitions can be either abrupt (cut) or gradual (fades, dissolves, wipes). In this work, a new shot boundary detection (SBD) method is proposed using Structural SIMilarity (SSIM) Index. The abrupt cuts are identified using SSIM and gradual transitions (fades) are identified using standard deviation plot of the frames in the video. The proposed method only needs mean, standard deviation and co-variance of the frames as basic input parameters for detecting cuts and gradual transitions. The performance of the proposed method is comparable with that of the existing global and local histogram method for SBD.

Keywords - *Shot, Structural similarity index, Abrupt transitions, Gradual transitions.*

I. INTRODUCTION

Video processing is an incipient area for investigation that has magnetized many researchers to focus their studies on shot boundary detection (SBD) in digital video. Multimedia applications have been expanded over the last decade and the need for their efficient management is necessary. Videos have become very popular in many areas such as communications, education and entertainment. A huge collection of video clips, live TV programs and movies can be found on the internet. Searching the video, retrieving relevant information from the huge database is time consuming process. Also viewers want better control over the video data. As more digital videos are produced, the need to search and retrieve visual content in video has increased. Hence, many video browsing, indexing and summarization applications are developed. These applications require the videos to be available in an elementary form called shots. It is better to divide the video into smaller manageable parts so that retrieving relevant information will be an easy task. The detection of shots is essential to video analysis. It segments the video into its basic components. A shot is a series of interrelated consecutive pictures taken contiguously by a single camera and representing a continuous action in time and space. A frame is the basic unit in the video sequence. Videos are organized in a hierarchical structure of stories, scenes, shots and frames as shown in Fig 1.1.

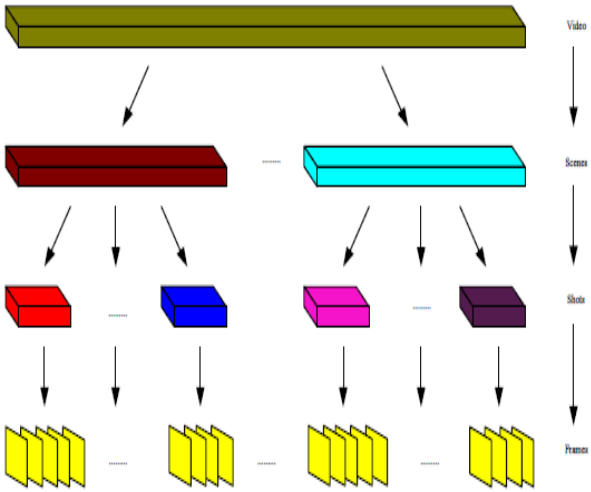


Fig.1.1: Hierarchical structure of a video clip

The hierarchical structure of the video [1] starts with the lowest level which consists of frames and is represented by images. A shot can be described as a sequence of frames that is continuously interpreted by the same camera and in the same action sequence. A scene is a collection of one or more shots focusing on one or more objects of interest. In other words it is a set of images (frames) taken from a single camera. A digital video sequence consists of group of scenes. A shot boundary separates two consecutive shots when one shot changes to another shot. Shot boundaries or shot transitions can be classified into two types.

1. Abrupt transition
2. Gradual transitions

A. Abrupt transitions

Abrupt transition is a transition from one shot to another. The visual content is suddenly transformed between shots, immediately after the last frame in the previous shot, and the first frame of the next shot. The most commonly used transitions that are found in films and on television production are cuts. The cut transition in a video sequence

happens when two different sequential frames belong to two different shots. Detecting abrupt transition is easier than detecting a gradual transition. An example of cut transition cut transition [2] is shown in Fig. 1.2.

Fig.1.2: Cut transition

B. Gradual transitions

A gradual shot transition occurs when the change takes place over a sequence of frames. The most common gradual transitions are fades (in-out), dissolves and wipe.

1) Fade: This is a shot transition with the first shot gradually disappearing (fade out) before the second shot gradually appears (fade in). The frames in fade-out transition go on fading out until the original content of the shot completely vanishes resulting in a totally black frame. This effect is usually used in movies to end a scene very smoothly. The sample of fade-out transition [1] is shown in Fig. 1.3.

Figure 1.3: Fade-out transition

The frames in fade-in transition starts appearing from a completely dark sequence. The each successive frame becomes brighter till the shot starts. This effect is popularly used in movies to start a scene smoothly. The sample of fade-in transition [1] is shown in Fig. 1.4.

Figure 1.4: Fade-out transition

2) Dissolve: A dissolve transition is a combination of a fade-out transition and fade-in transition where the first shot is replaced by another shot. The ending frames of the first shot go on fading-out slowly, and at the same time the frames in the second shot starts appearing slowly (fading-in). Thus, the transition from one shot to another is very smooth. In dissolve transition, the last few frames of the disappearing shot temporally overlap with the first few frames of the appearing shot. This effect is popularly used in the video industry for smooth scene and shot changes. An example of dissolve [1] is shown in Fig. 1.5.

Figure 1.5: Dissolve transition

3) Wipe: This transition is not a very smooth transition. One scene gradually enters across the view while another gradually leaves. The frames of the first shot in this transition are replaced by frames in the second shot in steps. Initially, no frames in the second shot are visible. Gradually in steps, the parts of frames in the second shot replace the respective parts of the frames in the first shot. An example of wipe [3] is shown in Fig. 1.6.

Figure 1.6: Wipe transition

In dissolve, fade-out and fade-in transitions, it is difficult to find a clear distinction between two consecutive frames, thus often becoming hard to detect. Another problem that makes dissolve detection difficult is that it is often confused with motion.

II. LITERATURE SURVEY

In the past decade, significant work has been done in the area of video processing to partition a given video into particular shots. The basic idea of most of the techniques is to measure and compare the similarities between consecutive frames. Some of the existing SBD methods are

1. Pixel comparison based SBD
2. Histogram comparison based SBD
3. Temporal video segmentation based SBD
4. Walsh-Hadamard transform based SBD

The following subsections explains the methods mentioned above.

A. Pixel comparison based SBD

Pair-wise pixel comparison evaluates the differences in intensity or color values of corresponding pixels in two successive frames. The simplest way is to calculate the absolute sum of pixel differences and compare it against a threshold [4]. The frame difference equations for gray level and color images are given in Eq. 2.1 and Eq. 2.2 respectively.

$$D(i,i+1)=\sum_{x=1}^{x=X}\sum_{y=1}^{y=Y}\frac{\left(p_i(x,y)-p_{i+1}(x,y)\right)}{XY} \quad (2.1)$$

$$D(i,i+1)=\sum_{x=1}^{x=X}\sum_{y=1}^{y=Y}\sum_{c}\frac{\left(p_i(x,y,c)-p_{i+1}(x,y,c)\right)}{XY} \quad (2.2)$$

where *i* and *i +1* are two successive frames with dimension X × Y . $p_i(x,y)$ is the intensity value of the pixel at the coordinates (x, y) in frame i, $p_{i+1}(x,y)$ is the intensity value of the pixel at the coordinates (x, y) in frame i+1, *c* is index for the color components (R,G,B) and $p_i(x,y,c)$ is the color component of the pixel at (x, y) in frame i. A cut is detected if the differences *D(i, i+1)* is above a specified threshold T . The main disadvantage of this method is that, it is not able to distinguish between a large change in a small area and a small change in a large area. For example, cuts are misdetected when a small part of the frame undergoes a large, rapid change. Therefore, methods based on simple pixel comparison are sensitive to object and camera movements. Also it is sensitive to noise. A simple improvement is to count the number of pixels that change in value more than some threshold and compare the total against a second threshold [5], [6] and [7].

$$DP(i,i+1,x,y)=\begin{cases}1, & \text{if } \left|p_i(x,y)-p_{i+1}(x,y)\right|>T_1\\ 0, & \text{otherwise}\end{cases} \quad (2.3)$$

$$\mathrm{D(i,i+1)}=\frac{\sum_{x=1}^{X}\sum_{y=1}^{Y}DP(i,i+1,x,y)}{XY} \quad (2.4)$$

Where *DP(i,i +1, x,y)* is a partial match between the k^{th} blocks in *i* and *i+1* frames.

B. Histogram Comparison based SBD

The reduction of sensitivity to camera and object movements can be done by comparing histograms of successive frames. Histogram differences are most widely used in shot detection. The idea behind histogram approach is that, two frames with unchanging background and unchanging objects will have little difference in their histograms. Histograms are invariant to image rotation and change slowly under the variations of viewing angle and scale [8].

1) Global Histogram: In global histogram, instead of intensity values, the gray level histograms are compared [5] [6] [9]. The frame difference equation for global histogram is given by

$$D(i,i+1)=\sum_{j=1}^{j=n}(|H_i(j)-H_{i+1}(j)|) \quad (2.5)$$

where $H_i(j)$ is the histogram value for the gray level j in the frame *i*, *j* is the gray value and *n* is the total number of gray levels. The difference between two frames across a cut is enhanced by using χ^2 test to compare the (color) histograms $H_i(j)$ and $H_i(j+1)$ of the two successive frames *i, i + 1.* The frame difference equation for χ^2 test is given in equation by

$$D(i,i+1)=\sum_{j=1}^{j=n}\frac{(|H_i(j)-H_{i+1}(j)|)^2}{H_{i+1}(j)} \quad (2.6)$$

χ^2 test not only enhances the difference between two frames across a cut but also increases the difference due to camera and object movements.

2) Local Histogram: Histogram-based approaches [10] are simple and more robust to object and camera movements but they ignore the spatial information and hence fail when two different images have similar histograms. Block-based comparison methods make use of spatial information. The frame-to-frame difference of frame *i* and frame *i+1* is computed using following equations.

$$D(i,i+1)=\sum_{k=1}^{k=b}DP(i,i+1,k) \quad (2.7)$$

$$DP(i,i+1,k)=\sum_{j}^{n}(|H_i(j,k)-H_{i+1}(j,k)|) \quad (2.8)$$

Where *DP(i, i+1, k)* is a partial match value between the k^{th} blocks in *i* and *i+1* frames. In [11] color local histogram comparison is done.

C. Temporal video segmentation based SBD

Chen Yinzi [12] proposed temporal video segmentation method based on the detection of shot abrupt transitions and gradual transitions. Based on the user requirement, the method generates different video summarization for each user.
The individual steps to generate video summarization are as follows.

1) Color space transformation: First the video frame data are abstracted from original video. It is more natural for human visual system to describe a color in HSV color space than in YUV color space. Hence the color transformation from YUV to HSV is done.

2) Color quantization: In this step the quantization of continuous HSV color values into dicrete intervals is done using following equations.

$$h' = \left\lfloor \frac{h}{\Delta h} \right\rfloor$$
$$s' = \left\lfloor \frac{s}{\Delta s} \right\rfloor \quad (2.9)$$
$$v' = \left\lfloor \frac{v}{\Delta v} \right\rfloor$$

where h', s', and v' denote the quantization intervals of H, S and V dimensions. (h', s', v') are the quantized color value.

3) Feature extraction: Color histogram [13] is used for feature extraction. Normalize the color histogram.The following equation is for color histogram of the frame f.

$$\mathrm{HIST}(\mathrm{f}) = (\mathrm{hist}[0], \mathrm{hist}[1], \ldots . \mathrm{hist}[\mathrm{L}-1]) \quad (2.10)$$

where L is the dimensionality of the color histogram.

4) Detection of abrupt and gradual transition: To compute histogram intersection distance number of methods are available in literature [14]. Comparing with others, histogram intersection distance needs less computational burden. The color histogram comparison is done by using histogram intersection distance. If

$$D(f[i-1], f[i]) > AT \quad (2.11)$$

abrupt transition is observed between $(i-1)^{th}$ frame and i^{th} frame. If

$$\sum_{i=s}^{t-1} D(f[i], f[i+1]) > GT \quad (2.12)$$

gradual transition happens from sth frame to tth frame.
Here, *AT*=Abrupt transition threshold, *GT*=Gradual transition threshold and *D(f[i − 1], f[i])* is the histogram intersection distance between successive frames in the video.

5) Summary generation: The video is segmented into shots. Then an appropriate number of frames are selected from each shot depending on its importance. The selected frames are keyframes. Using the keyframes, video summary is generated.

D. Walsh-Hadamard transform based SBD

Priya and Domnic in [15] has proposed a new SBD method using color, edge, texture, and motion strength as vector of features and Walsh Hadamard Transform (WHT) matrix.
The procedure for shot boundary detection is as follows.

1) Basis vectors for feature extraction: Image transforms [16] are designed to have one of the properties of isolating the various frequencies of the image. Low frequencies correspond to the important image features, high frequencies correspond only to the details of the image. WHT is simple and attractive. Hence, the WHT matrix is considered for feature extraction.

2) Feature extraction: Each row of WHT matrix is called a 1D Walsh Hadamard basis vector. The 2D WHT kernels (WHTK) or basis images are generated by tensor product of 1D basis vectors of WHT matrix. The four features (color,edge,texture and motion strength) are extracted by projecting the frames on selected basis vectors of WHT kernel and WHT matrix. The motion vector is computed using fast block matching algorithm in [17].

3) Construction of continuity signal: Construction of continuity signal is done by using similarity/dissimilarity between successive frames in the video sequence using a distance metric D. The distance metric used is city block distance measure. The amount of information contributed by a certain feature are calculated by using Kullback- Leibler (KL) measure [18]. The information content of a feature value f_{ij} is calculated using the following equation.

$$D_{KL}(i, j) = \sum_{tf \in TF} p(tf \mid f_{ij}) \log \left(\frac{p(tf \mid f_{ij})}{p(f_{ij})} \right) \quad (2.13)$$

Where $D_{KL}(i, j)$ is the average mutual information between the target feature TF and feature value f_{ij} . Based on the importance of features the feature weights are calculated. Feature weights are calculated using feature weighting method [19], using the following equation.

$$w(f_i) = \frac{\sum_{j|i} R.D_{KL}(i, j)}{-\sum_{j|i} p(f_{ij}) \log_2 p(f_{ij})} \quad (2.14)$$

The continuity signal is calculated using the following equation.

$$\phi = w_1\alpha + w_2\beta + w_3\gamma + w_4\lambda \quad (2.15)$$

*4) Shot transition identification technique:*After obtaining continuity signal the shot transitions are identified using Procedure Based Identi-fication (PBI) technique [15].

III. STRUCTURAL SIMILARITY INDEX

The most widely used quality metrics are mean squared error (MSE) and peak signal-to-noise ratio (PSNR). In [20] and [21] a new framework was proposed for image quality measures. Structural SIMilarity (SSIM) index has been proposed in [22] as an objective quality measure to compare the similarity between two images X and Y . This method has been proved as better, when compared with metrics such as MSE and PSNR [23].
The block diagram of similarity measurment system [22] is shown in Fig. 3.1.

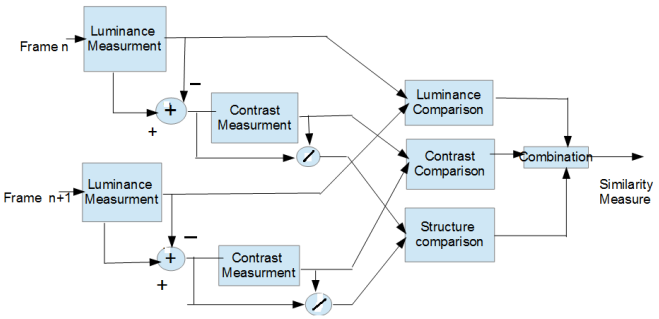


Fig.3.1: Structural Similarity Measurment System

where, n and n+1 are successive frames in the video.
The similarity measurement is done in three steps and are explained below.
1. Luminance comparison
2. Contrast comparison
3. Structure comparison

A. Luminance comparison

The luminance is an estimate of the mean intensity of the frame and is defined by the following equation.

$$\mu_{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (3.1)$$

where N is the number of pixels in the frame.
The luminance comparison function $l(\mathbf{X},\mathbf{Y})$ is a function of μ_X and μ_Y and is calculated as follows.

$$l(\mathbf{X},\mathbf{Y}) = \left(\frac{2\mu_{\mathbf{X}}\mu_{\mathbf{Y}} + C_1}{\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2 + C_1}\right) \qquad (3.2)$$

where μ_X is mean intensity of the frame $\boldsymbol{X}$, μ_Y is mean intensity of the frame $\mathbf{Y}$ and C_1 is the constant assumed to be zero (because some dissimilarity exists between successive frames in the video).

B. Contrast comparison

The contrast is an estimate of the standard deviation of the frame and calculated by the following equation.

$$\sigma_{\mathbf{X}} = \left(\frac{1}{N-1}\sum_{i=1}^{i=N}(x_i - \mu_{\mathbf{X}})^2)^{\frac{1}{2}}\right) \qquad (3.3)$$

The contrast comparison function c($\mathbf{X}$,$\mathbf{Y}$) is a function of σ_X and σ_Y and is calculated as follows.

$$c(\mathbf{X},\mathbf{Y}) = \left(\frac{2\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}} + C_2}{\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2 + C_2}\right) \qquad (3.4)$$

where σ_X is standard deviation of the frame X, σ_Y is standard deviation of the frame Y and and C_2 is the constant.

C. Structure comparison

Structure comparison is conducted after luminance subtraction and variance normalization. Structure comparison is calculated by the following equation.

$$s(\mathbf{X},\mathbf{Y}) = \left(\frac{\sigma_{\mathbf{XY}} + C_3}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}} + C_3}\right) \qquad (3.5)$$

The covariance between successive frames is denoted by σ_{XY} is calculated as follows.

$$\sigma_{\mathbf{XY}} = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{\mathbf{X}})(y_i - \mu_{\mathbf{Y}})^{\frac{1}{2}}\right) \qquad (3.6)$$

C_3 is the constant. Lastly, structural similarity index between successive frames is combination of luminance, contrast and structure comparisons and is defined by the following equation. To have equal weightage of three comparisons here α, β and γ are assumed to be 1.

$$SSIM(\mathbf{X},\mathbf{Y}) = \left[l(\mathbf{X},\mathbf{Y})^{\alpha}\right]\left[c(\mathbf{X},\mathbf{Y})^{\beta}\right]\left[s(\mathbf{X},\mathbf{Y})^{\gamma}\right] \qquad (3.7)$$

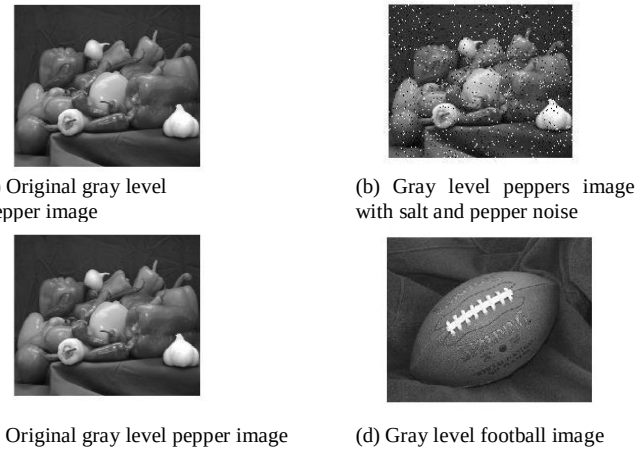
The final equation for SSIM is given below.

$$SSIM(\mathbf{X},\mathbf{Y})) = \frac{(2\mu_{\mathbf{X}}\mu_{\mathbf{Y}} + C_1)}{(\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2 + C_1)}\frac{(2\sigma_{\mathbf{XY}} + C_2)}{(\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2 + c_2)} \qquad (3.8)$$

($\mathbf{X}$,$\mathbf{Y}$) represents successive frames in the video. In this thesis, C_1, C_2 are neglected because some dissimilarity will exists between successive frames in the video. Hence it is expected that there is less chance for μ_X, μ_Y to be zero. The modified equation for SSIM as follows.

$$SSIM(\mathbf{X},\mathbf{Y})) = \frac{(2\mu_{\mathbf{X}}\mu_{\mathbf{Y}})}{(\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2)}\frac{(2\sigma_{\mathbf{XY}})}{(\mu_{\mathbf{X}}^2 + \mu_{\mathbf{Y}}^2)} \qquad (3.9)$$

Here two images are taken and structural similarity is calculated between them. The structural similarity between two images is calculated using Eq. 3.9. The two images are shown in Fig. 3.2(a) and Fig. 3.2(b). The two images are perceptually same. Hence, the structural similarity is high and SSIM was found to be 0.9551. Next, two different images are taken and are shown in Fig. 3.2(c) and Fig. 3.2(d). SSIM was found to be 0.2837.



(a) Original gray level pepper image
(b) Gray level peppers image with salt and pepper noise
(c) Original gray level pepper image
(d) Gray level football image

Fig.3.2: Various test images to calculate SSIM

IV. PROPOSED ALGORITHM

Based on assumption that human visual perception is highly adapted for extracting structural information from a scene, a new method structural similarity index has been proposed for image quality assessment in [22]. Here, the same method has been modified to detect shot boundaries in video applications. The aim of the proposed algorithm is to provide a simple and better approach for video shot detection.
The block diagram of proposed algorithm is shown in Fig. 4.1.

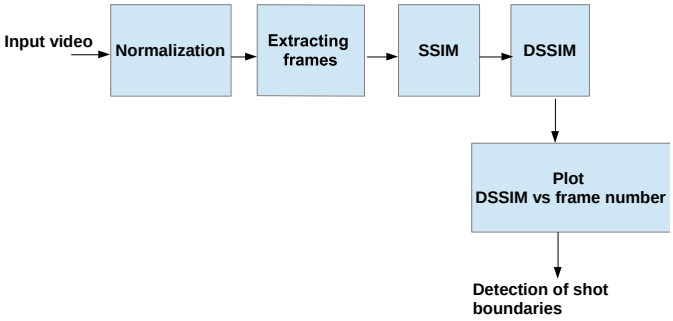


Fig.4.1: Block diagram of proposed algorithm to detect shot boundaries

In the block diagram, SSIM stands for Structural SIMilarity index and DSSIM stands for DiSSIMilarity measure.

The step by step procedure for detecting shot boundaries are listed as follows.

1. Extract the frames from the given video sequence.
2. Calculate the luminance, contrast and structure for individual frames in the video.
3. Calculate the luminance, contrast and structure comparisons.
4. Calculate the SSIM between the successive frames.
5. Calculate DSSIM between the successive frames.
6. Plot the graph of dissimilarity measure versus the frame number.
7. Sudden transitions from the above plot are counted as shot boundaries

The modified equation for SSIM is as follows.

$$SSIM(\mathbf{X},\mathbf{Y})) = \frac{(2\mu_X\mu_Y)}{(\mu_X^2+\mu_Y^2)}\frac{(2\sigma_{XY})}{(\mu_X^2+\mu_Y^2)} \qquad (4.1)$$

The DSSIM is calculated by using the following equation.

$$DSSIM = \frac{1}{1-SSIM} \qquad (4.2)$$

The block diagram of proposed algorithm to detect fades is shown in Fig. 4.2.

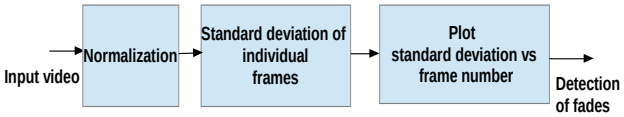


Fig.4.2: Block diagram of proposed algorithm to detect fades

The step by step procedure for detecting fade (in-out) are listed below.

1. Extract the frames from the given video sequence.
2. Calculate the standard deviation of individual frames in the video.
3. Plot the graph of standard deviation versus the frame number.
4. The gradual decrease in the intensity (downward trend) represents fade-out. In a similar fashion,

the gradual increase in the intensity (upward trend) represents fade-in.

V. SIMULATION RESULTS

In order to validate the effectiveness of the proposed SBD method based on structural similarity index, number of video samples have been tested. For comparison purpose, 50 test video sequences are considered from [23] [24] and [25]. The simulations are carried out using MATLAB software. The input videos are in two formats (such as mpg, avi). It is to be noted that thresholds used for detecting the shot boundaries are different for different methods. Fixing the common threshold for all three methods is a challenging task. So, following methodology has been adopted.

The plots of frame difference versus frame number, total frame difference versus frame number and Dissimilarity measure versus frame number are considered for global histogram, local histogram and SSIM methods respectively. Gradual transitions are also detected based on standard deviation of individual frames. By observing the sudden transitions from the plot, the shot boundaries are detected in all the three cases.

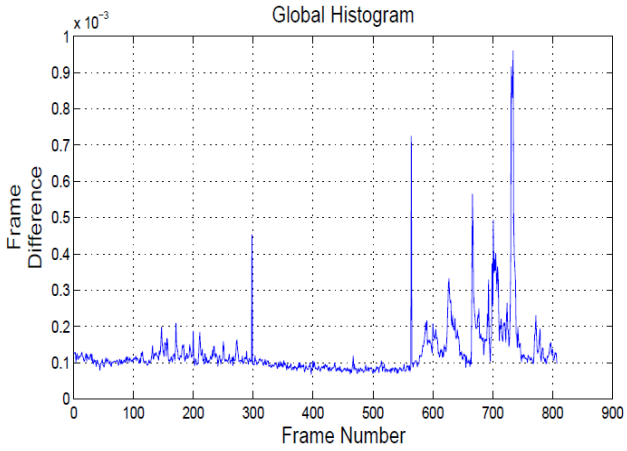


Fig.5.1: Simulation results for global histogram method

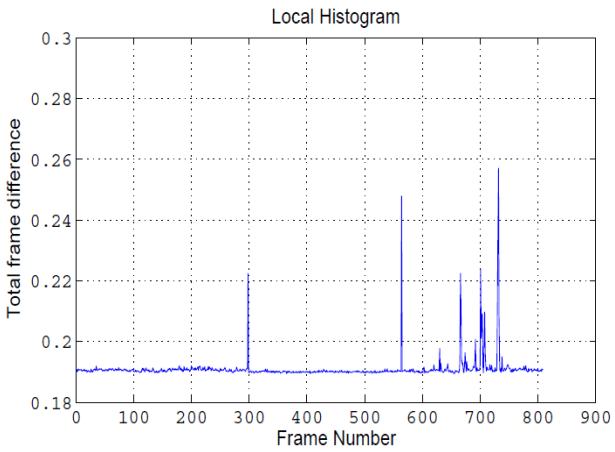


Fig.5.2: Simulation results for local histogram method

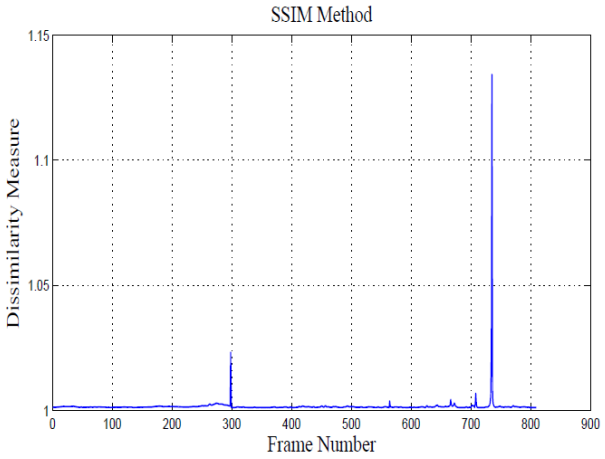


Fig.5.3: Simulation results for SSIM method

The plots shown in Fig. 5.1, Fig. 5.2 and Fig. 5.3 for three methods are for the test sequence clip8.avi. The plots shown in Fig. 5.4, Fig. 5.5 and Fig. 5.6 for three methods are for the test sequence Mergedvideo5.avi. The plots corresponding to fade-in and fade-out transitions are shown in Fig. 5.7

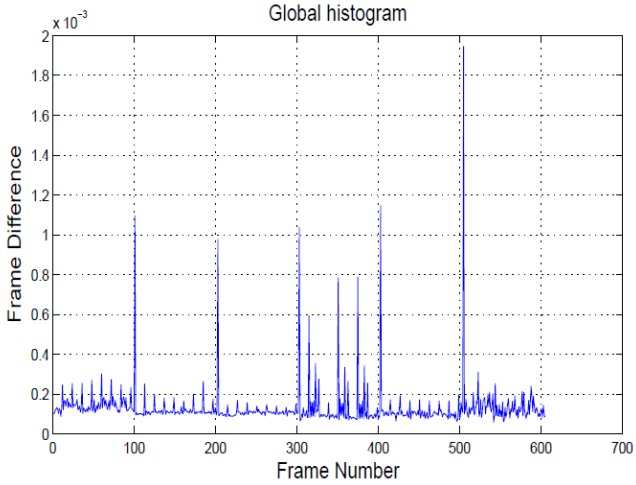


Fig.5.4: Simulation results for global histogram method

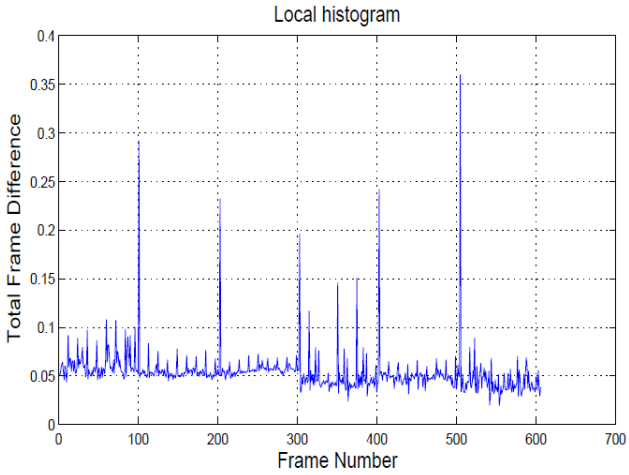


Fig.5.5: Simulation results for local histogram method

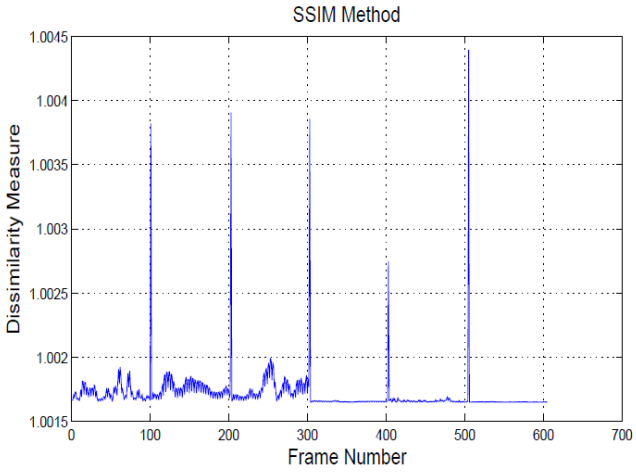


Fig.5.6: Simulation results for SSIM method

Fig.5.7: Simulation results for SSIM method

A transition that is detected correctly is called a hit and denoted by N_c, a cut that is present but was not detected is called a missed hit and denoted by N_m. Lastly, if the method assumes a cut, but actually no cut is present, it is called a false hit and denoted by N_f . From the simulation results the parameters N_c, N_m and N_f are calculated for 50 videos and results are listed in Table 5.1 and Table 5.2.

For comparison purpose, parameter called success rate is defined as follows.

$$Success\ rate = \frac{N_{Success}}{N_{Total}} \qquad (5.1)$$

Where $N_{success}$ = Number of test cases with correctly identified number of shot boundaries present in the video and

N_{Total} = Total number of test cases in the database.

It is observed that success rate in global histogram is 50%, in local histogram is 38% and in proposed SSIM method it is 76%. So, the performance of SSIM method is better than the local and global histogram methods for SBD. Precision recall curve is used to evaluate the performance of the proposed algorithm. Precision is defined as

$$\Pr ecision = \frac{N_c}{N_c + N_f} \qquad (5.2)$$

Recall is defined as

$$\text{Re}call = \frac{N_c}{N_c + N_m} \quad (5.3)$$

where N_c= Number of correct shots retrieved, N_m= Number of missed shots and N_f= Number of false positive shots.

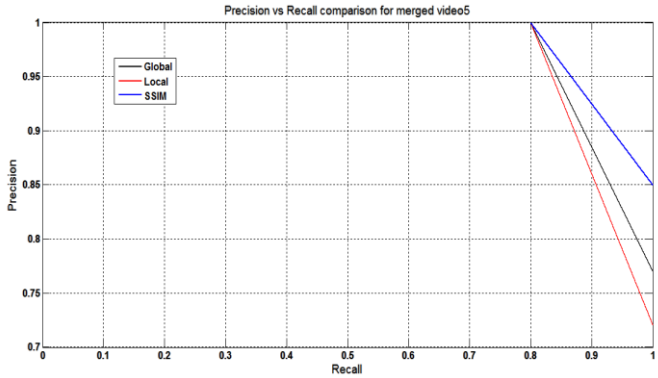


Fig.5.8: Precision recall curve for global histogram, local histogram and SSIM methods

VI. CONCLUSION AND FUTURE WORK

Video shot detection is an important step in the video indexing, retrieval and summarization applications. This report presented the method for video shot boundary detection using SSIM. The conclusion and future scope of the work is as follows.

A. *Conclusion*

Video shot detection can be done by detecting shot transitions. In this thesis, a new method for shot boundary detection (SBD) based on structural similarity index (SSIM) was proposed. The proposed method is very simple and its performance is comparable to the existing global and local histogram methods. It is observed that success rate in SSIM method is 76%. The gradual transitions (fade-in and fade-out) were detected by observing the standard deviation plot of the frames.

B. *Future Scope*

It is to be noted that proposed SSIM method is based on manual identification of number of shots by observing sudden transitions from the simulation results. This process will be easy for the videos associated with less number of shot boundaries (<20). But, if the number of shot boundaries are in the order of hundreds or more then it will be very cumbersome to identify manually. So, in future the method can be extended for automatic identification.

Video Name	Actual Shot Boundaries	Global Histogram			Local Histogram			SSIM		
		Nc	Nm	Nf	Nc	Nm	Nf	Nc	Nm	Nf
2.mpg	2	2	0	0	2	0	0	3	0	1
13.mpg	2	3	0	1	3	0	1	4	0	2
22.mpg	3	4	0	1	3	0	0	3	0	0
31.mpg	2	2	0	0	2	0	0	2	0	0
50.mpg	2	2	0	0	2	0	0	2	0	0
55.mpg	9	5	4	0	7	2	0	9	0	0
74.mpg	6	5	1	0	4	2	0	4	2	0
88.mpg	3	3	0	0	3	0	0	3	0	0
102.mpg	2	2	0	0	2	0	0	2	0	0
111.mpg	6	7	0	1	6	0	0	5	1	0
124.mpg	3	4	0	1	2	1	0	3	0	0
132.mpg	3	5	0	2	5	0	2	3	0	0
136.mpg	5	3	2	0	3	2	0	5	0	0
139.mpg	1	2	0	1	2	0	1	2	0	0
154.mpg	4	8	0	4	6	0	2	4	0	0
174.mpg	4	6	0	2	5	0	1	4	0	0
177.mpg	10	10	0	0	11	0	1	10	0	0
199.mpg	2	2	0	0	2	0	0	2	0	0
Clip8.avi	2	5	0	3	5	0	3	2	0	0
Clip90.avi	2	2	0	1	1	0	0	1	0	0
Clip195.avi	2	4	0	2	1	1	0	2	0	0
Clip272.avi	3	3	0	0	2	1	0	3	0	0
Clip302.avi	3	3	0	0	2	1	0	2	1	0
Clip358.avi	2	4	0	2	3	1	1	2	0	0
Clip360.avi	4	5	0	1	2	2	0	12	0	0

Table 5.1: Comparison of global histogram, local histogram and SSIM methods

Video Name	Actual Shot Boundaries	Global Histogram			Local Histogram			SSIM		
		Nc	Nm	Nf	Nc	Nm	Nf	Nc	Nm	Nf
Clip412.avi	1	1	0	0	1	0	0	1	0	0
Clip549.avi	7	7	0	0	8	0	1	8	0	1
Clip673.avi	5	5	0	0	5	0	0	5	0	0
Clip703.avi	2	2	0	0	2	0	0	2	0	0
Clip752.avi	1	1	0	0	1	0	0	1	0	0
Clip804.avi	2	2	0	0	2	0	0	2	0	0
Mergedvideo1.avi	9	9	0	0	9	0	0	9	0	0
Mergedvideo2.avi	5	5	0	0	8	0	3	5	0	0
Mergedvideo3.avi	7	8	0	1	8	0	1	7	0	0
Mergedvideo4.avi	11	11	0	0	11	0	0	11	0	0
Mergedvideo5.avi	5	8	0	3	8	0	3	5	0	0
Mergedvideo6.avi	12	13	0	1	14	0	2	12	0	0
Mergedvideo7.avi	15	16	0	1	17	0	2	17	0	2
Mergedvideo8.avi	10	10	0	0	10	0	0	9	1	0
Mergedvideo9.avi	9	10	0	1	10	0	1	9	0	0
Mergedvideo10.avi	16	16	0	0	16	0	0	17	0	0
Mergedvideo11.avi	17	18	0	1	18	0	1	17	0	0
Mergedvideo12.avi	15	15	0	0	14	1	0	15	0	0
Mergedvideo13.avi	15	17	0	2	17	0	2	16	0	1
Mergedvideo14.avi	14	14	0	0	15	0	1	14	0	0
Mergedvideo15.avi	14	16	0	2	16	0	2	15	0	1
Mergedvideo16.avi	16	18	0	2	17	0	1	16	0	0
Mergedvideo17.avi	10	9	1	0	9	1	0	10	0	0
Mergedvideo18.avi	15	15	0	0	15	0	0	15	0	0
Mergedvideo19.avi	12	12	0	0	12	0	0	12	0	0

Table 5.2: Comparison of global histogram, local histogram and SSIM methods (Continued)

REFERENCES

[1] K. I. Koumousis, V. E. Fotopoulos, and A. N. Skodras, "A new approach to gradual video transition detection," in *16th Panhellenic Conference on Informatics, PCI 2012, Piraeus, Greece, October 5-7, 2012*, 2012, pp. 245–249.

[2] H. Muurinen, "Video segmentation and shot boundary detection using self-organizing maps," Master's thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, February 2007.

[3] E. ASAN, "Video shot boundary detection by graph theoretic approaches," Master's thesis, THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF MIDDLE EAST TECH-NICAL UNIVERSITY, SEPTEMBER 2008.

[4] T. Kikukawa and S. Kawafuchi, "Development of an automatic summary editing system for thaudio-visual resources," *Transactions on Electronics and Information J75-A*, pp. 204–212, 1992.

[5] A. K. H.J. Zhang and S. Smoliar, "Automatic partitioning of full motion video, in multimedia systems," vol. 1, pp. 10–28, 1993.

[6] Y. T. Akio Nagasaka, "Automatic video indexing and full-video search for object appearances," *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, pp. 113–127, 1992.

[7] A. Hanjalic, *Content-based Analysis of Digital Video*, illustrated ed. Springer Science & Business Media, 2004.

[8] M. J. Swain, "Interactive indexing into image databases," in *Storage and Retrieval for Image and Video Databases (SPIE)*, 1993, pp. 95–103.

[9] Y. Tonomura, "Video handling based on structured information for hypermedia systems," in *Proceedings*

of ACM International Conference on Multimedia Information Systems. ACM, 1991, pp. 333–344.

[10] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 477 – 500, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S09 23596500000114

[11] D. Swanberg, C. Shu, and R. Jain, "Knowledge-guided parsing in video databases," in *Storage and Retrieval for Image and Video Databases (SPIE)*, 1993, pp. 13–24.

[12] C. Yinzi, D. Yang, G. Yonglei, W. Wendong, Z. Yanming, and W. Kongqiao, "A temporal video segmentation and summary generationmethod based on shots' abrupt and gradual transition boundary detecting," in *Communication Software and Networks, 2010. ICCSN '10. Second International Conference on*, Feb 2010, pp. 271–275.

[13] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[14] J. W. Jianjun Dou and C. Liu, "Histogram based color image retrieval," *Infrared and Laser Engineering*, vol. 34, pp. 84–87, 2005.

[15] G. Lakshmi Priya and S. Domnic, "Walsh 2013;hadamard transform kernel-based feature vector for shot boundary detection," *Image Processing, IEEE Transactions on*, vol. 23, no. 12, pp. 51875197, Dec 2014.

[16] D. Salomon, *Data Compression: The Complete Reference*, 2007, with contributions by Giovanni Motta and David Bryant.

[17] M. Santamaria and M. Trujillo, "A comparison of block-matching motion estimation algorithms," in *Computing Congress (CCC), 2012 7th Colombian*, Oct 2012, pp. 1–6.

[18] E. T. Jaynes, "Information theory and statistical mechanics, I and II," *Physical Reviews*, vol. 106 and 108, pp. 620–630 and 171–190, 1957.

[19] C.-H. Lee, F. Gutierrez, and D. Dou, "Calculating feature weights in naive bayes with kullback- leibler measure," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, Dec 2011, pp. 1146–1151.

[20] Z. Wang, "Rate scalable foveated image and video communications," Ph.D. dissertation, The University of Texas at Austin, Dec 2001.

[21] Z. Wang, A. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 4, May 2002, pp. IV–3313–IV–3316.

[22] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, April 2004.

[23] http://media.xiph.org/video/derf/, "xiph database," 2015, test video sequences.

[24] http://www.reefvid.org/, "Reefvid database," 2015, test video sequences.

[25] http://www nlpir.nist.gov/projects/tv2011/pastdata/copy.detection /201/, "Trecvid 2011 database," 2011, test video sequences.

Srilakshmi B. completed B.Tech in Electronics and Communication Engineering from Narasaraopeta Engineering College in the year 2006 . Currently, she is pursuing the M.Tech. degree in Signal processing, Cambridge Institute of Technology (Affliated to VisvesarayaTechnological University) Bengaluru,India. Her current research interests include signal processing, video segmentation.

Sandeep R. received the B.A. degree in Hindi from Mysore Hindi Prachar Parishad, the B.E. degree in Electronics and Communication Engineering and the M.Tech. Degree in Electronics Engineering from Visvesvaraya Technological University in the year 2001, 2006 and 2009 respectively. Currently, he is pursuing Ph.D. in the department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India and working as Associate Professor Cambridge Institute of Technology (Affliated to Visvesaraya Technological University) Bengaluru, India. He has both academic and industry experience. His current research interests include perceptual image hashing, perceptual video hashing, biometric hashing and biomedical image processing.