AUDIO-VISUAL CONTENT-BASED VIOLENT SCENE CHARACTERISATION

By

LaxmiKantaNayak

Roll No. 111EC0181

8th Semester, B.TECH(ELECTRONICS & COMMUNICATION ENGG.)

Guided by

Prof. LakshiPrasad Roy

Department of Electronics & Communication Engineering

National Institute of Technology Rourkela

National Institute of Technology, Rourkela

DECLARATION

I proclaim that the project work entitled “AUDIO-VISUAL CONTENT-BASED VIOLENT SCENE CHARACTERISATION” is a record of my original work done under Prof. Lakshi Prasad Roy, National Institute of Technology, Rourkela. I have followed the guidelines provided by the Institute while preparing the thesis. I have confirmed to the norms and guidelines in the Ethical Code of Conduct of the Institute. Throughout this project wherever contributions of others are involved, every endeavor has been made to acknowledge this clearly with due reference to literature. This project work is being submitted in the partial fulfilment of the requirements for the degree of Bachelor of Technology in Electronics and Communication Engineering at National Institute of Technology, Rourkela for the academic session 2011–2015.

Laxmi Kanta Nayak

111ec0181

Department of Electronics and Communication Engineering

National Institute of Technology, Rourkela

CERTIFICATE

This is to certify that the Thesis report entitled "**AUDIO-VISUAL CONTENT-BASED VIOLENT SCENE CHARACTERISATION**", submitted to the National Institute of Technology, Rourkela by Mr. Laxmi Kanta Nayak, Roll No: 111EC0181 for the award of Bachelor of Technology in Electronics & Communication Engineering is a bona-fide record of research work carried out by him under my supervision and guidance.

The candidate has fulfilled all the prescribed requirements. The report which is based on candidate's own work has not been submitted elsewhere for a degree/diploma. In my opinion, the report is of standard required for the award of a Bachelor of Technology in Electronics & Communication Engineering.

Supervisor

Prof. Lakshi Prasad Roy

ACKNOWLEDGEMENTS

On the submission of my Thesis report of "**AUDIO-VISUAL CONTENT-BASED VIOLENT SCENE CHARACTERISATION**", I would like to express my sincere gratitude to my supervisor **Prof. Lakshi Prasad Roy** for his guidance. I would like to thank him for his encouragement, and support throughout the course of this work from the core of my heart. It was an invaluable learning experience for me to be one of his students. As my supervisor his insight, observations and suggestions helped me to establish the overall direction of the research and contributed immensely for the success of this work. His immense knowledge, technical skills and human values have been a source of inspiration to me.

LAXMI KANTA NAYAK

111EC0181

ABSTRACT

In this thesis, I show a novel strategy to portray and list violent events in TV dramas and movies. Our objective is to distinguish violent scenes from the normal scenes and confine brutal occasions inside a film to bolster "abnormal state" feature indexing. Specially, I have investigated multiple “audiovisual” signals to create a perceptual connection for reasonably important vicious scene recognizable proof. Potential applications are programmed obstructing of violent scenes in motion pictures observed by kids, hiding viciousness utilizing information filtering or data sifting and genre classification of advanced feature database.

Table of Contents

1 INTRODUCTION

1.1 Background and motivation

As violence in movies is regarded harmful for children, we propose distinctive algorithms to detect violence scene. Movies are extremely accessible to the viewing audience these days due to a few progressions in innovation such as digitalization of movies and players. One sort extremely mainstream to masses are the Video CDs accessible everywhere. These scenes contain wide range of film genres including horror, suspense cum thriller and action which depict violence. The Video-indexing schemes rely on two types visual feature analysis such as low-level and high-level. These schemes allow searching for image/video containing visual properties like colour, motion, texture, sketch of video objects[1]. All of the existing work on “Detection of violent content” relies on using single source of information. An action scene is described by temporarily localized properties of video shots which have little or no repeating similar visual contents[2], but they are not sufficient to determine violent actions . For instance, it is very difficult to differentiate violence from some highly active sports video using these methodologies. On the other hand, audio based violence detection was autonomously performed on soundtrack[1].

1.2 AUDIO CONTENT ANALYSIS

1.2.1 Fundamental properties of Audio

The substance of sound must be respected from two focuses: first regarding quantifiable properties, from the angle of material science, e.g. amplitude or waveform, and second, as for properties of human cognition, for example, loudness or harmony[3].

1.2.2 Physical property

Sound is described as a change in air pressure which is demonstrated as a waveform consists of sinusoidal waves of different amplitude, frequency and phase. Experiments with different sounds have demonstrated that the human ear cannot differentiate phases, but it is well known that we hear amplitude variations as variations in loudness , and frequency variations as variations in pitch[3].

More fascinating than the waveform itself, is regularly its composition of sinusoidal waves and their amplitudes and frequencies which is likewise called Fast Fourier Transform[4].

1.3 Substance based segmentation

So as to perceive the substance of sound, it is important to first structure the sound stream. This is like deciding substance in still pictures: effective article division is the premise for further handling. Our initial phase in substance based sound division is to recognize music, discourse, quiet and other sound groupings, on the grounds that treatment of substance contrasts in a general sense for each of these. For instance, if a self-assertive bit of sound is discovered to be discourse, discourse acknowledgment and speaker acknowledgment can be performed on it. On the off chance that it is discovered to be music, note, bar or topic limits may be extricated, and central frequencies can be resolved.

How can the general classification into silence, speech, music and other sounds be accomplished? First, we divide the audio stream into similar segments. This is performed both in the temporal and in the frequency domain. Humans determine silence on a relative scale: a loudness of '0' dB is not very common in any natural environment. Therefore, an automatic recognition of silence must be based on comparison of loudness levels along a timeline and with an adaptive threshold. In that way, silence can be distinguished from other sound classes.

Speech and music are discernable simply by the spectrum that they cover: speech frequencies lie in the scope of 100 to 7000 Hz, and music frequencies between around 16 and 16000 Hz.

1.4 Difference between audio & audio-visual signal

Normally, audio signals have their frequency spectrum between 20Hz to 20 kHz. However, audio-visual signals have frequency range between 20Hz to 4.5MHz. Audio-visual signal can be considered as combination of both audio & pictures .In audio-visual signal, audio have a sampling rate of typically 44 kHz. Similarly, video have typically 22 frames per second.

1.5 Violence detection in audio part

For detection of violence scenes in audio visual signal, we have to consider separately for audio & video part that means we divide audio visual signal into audio & video cues. In audio part of the signal, we detect the high pitch instances i.e,the time instances where the energy entropy of the audio sample exceeds our predefined threshold value. Similarly, for video we analyse frame by frame.

In general violent scenes are accompanied by exceptional non discourse sounds (ex:- blast, shouting, gunfire and so on.). Also characterising feature of violent scenes in the background music. Especially, we focus the class (i.e, vicious or peaceful soundtrack) to which the sound track has a place by viewing the given sound track in light of Gaussian demonstrating strategy.

In our experiment, we have extracted the audio part from the audio-visual signal by means of a software entitled as “Any Video Converter”. This software converts video of .avi format to audio of .wav format which is readable by MatlabR2012b.We have taken a video song from a Bollywood movie & then analyze its Amplitude Vs time curve and Energy versus time. We have the file named as “E:/abe.wav”.

For more efficient approach, we consider energy entropy of the audio part. In that case, we consider the sudden change in energy level of the audio signal as another feature associated with violence. It is figured by separating each (audio analysis) frame into segments of K samples each[1]. The signal energy is calculated over each of these segments and normalized by the overall frame energy. Then, the energy entropy, I, of the frame is given by:

$$I = \sum_{i=1}^{j} \sigma^2 log_2 \sigma_i^2$$

Where , J is the total segments in a frame and σ is the normalized energy of the ‘i’th short segment in a frame (In our experiments, we use K=15 and **J=64**

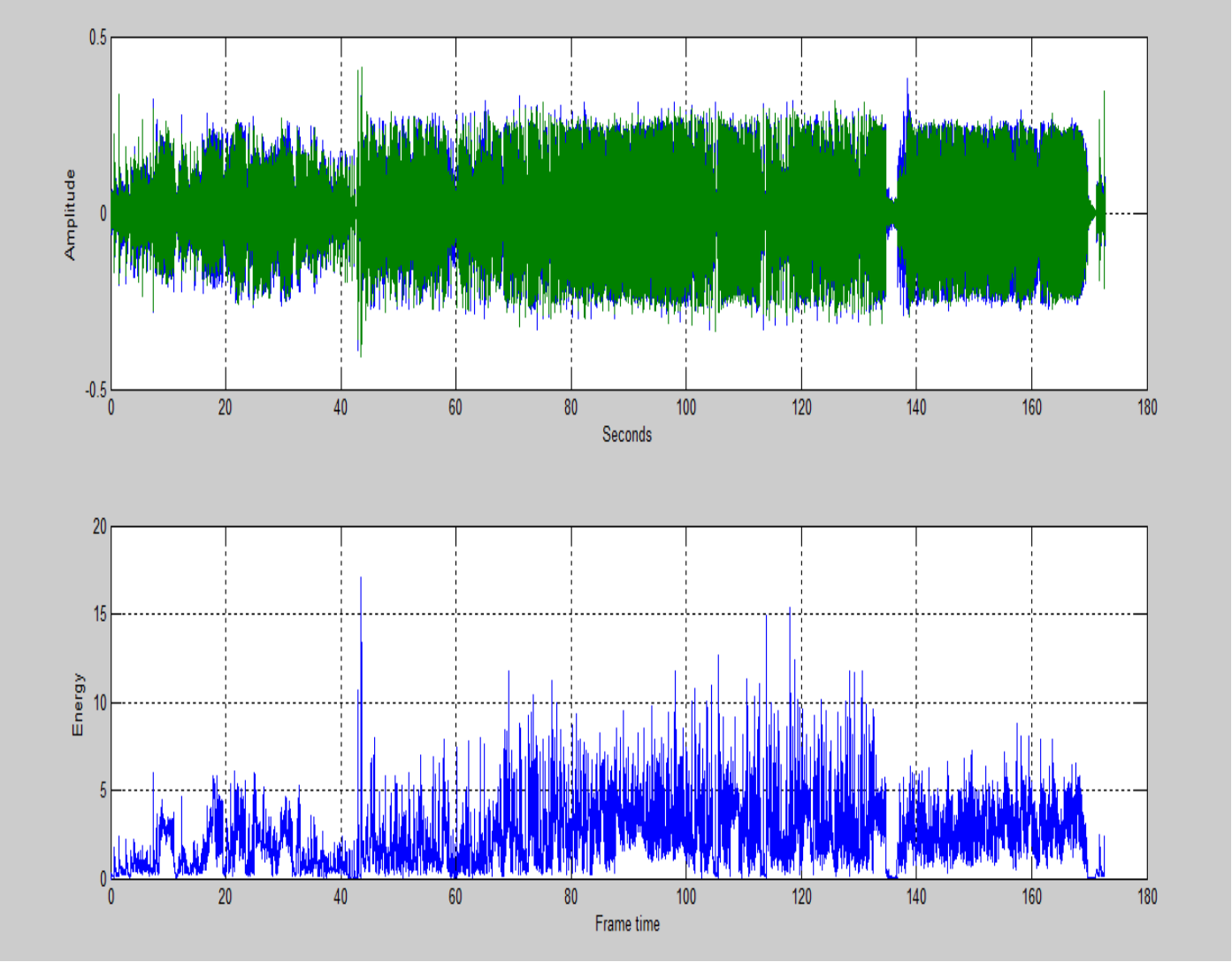
With **46** msec frame).

1.6 Matlab Code

```
filename= ‘E:\abe.wav’;

frame=100;

[y fs]= wavread(filename);

Time=length(y)/fs;

dt=1/fs;

dtf=1/frame;

t=0:dt:(time-dt);
```

```
n=time*frame;
p=zeros(1,n);
m=fs/frame;
tf=0:dtf:(time-dtf);
for j=1:n
 p(j)=0
for i=1:m
p(j)=p(j)+power(y(i+m*(j-1)),2);
end
end
subplot(2,1,1); plot(t,y);
xlabel('Seconds');ylabel('Amplitude');
grid on;
subplot(2,1,2); plot(t,p);xlabel('Frame time');ylabel('Energy');grid on;
```

1.7 Variation of signal and energy with time



[Plot of amplitude & energy against time of the audio part of the audio-visual signal]

[Figure 1.1]

2 VIDEO PART ANALYSIS

2.1 Visual feature analysis

We introduce in this area some visual highlights created from clearly vicious activities and occasions which are effortlessly found in regular movies. In any case, we must observe that roughness is subjective and type subordinate. There can exist numerous other diverse sorts of violence related occasions in movies. Some of these roughness occasions may not be successfully identified by a blend of the highlights that we exhibit here.

2.2 Spatio-temporal dynamic activity

Most activity scenes include quick and critical developments of persons and items (e.g., battling or pursuing scenes). Such dynamic movement is viably spoken to by progressive transient feature shots in a brief while term[5]. Specifically, the spatial variety of development inside the shot and the worldly varieties in shot term can build a "visual musicality" as a rule movies.

To productively estimate these spatial temporary dynamic activity highlights, we first concentrate the movement arrangement from a feature. This succession is registered by a temporary high-pass filtering of the 1-D wavelet change along the time development of the force at every pixel in the spatially lessened coarse edges from 2-D wavelet decay of the first feature[1]. The movement arrangement delivered by the feature grouping division approach which catches just the spatial variety of noticeable moving questions inside every feature shot. We process the spatio-transient dynamic activity of every video shot from movement grouping as follows:

$$\text{Dynamic Activity} = \frac{1}{T}\sum_{i=b+1}^{e}\left(\sum_{m,n} m_i^k(m,n)\right),$$

where $m_i^k(m,n)$ is 'i'th frame of motion sequence within the kth video shot beginning at 'b'th frame and ending at 'e'th frame, and 'T' is the associated shot length

(i.e, T = e - b). In this way, the shorter length and more movement every shot has, the higher value its motion density demonstrates[6].

To describe the dynamic activity scene, we distinguish an arrangement of time-obliged progressive feature shots whose dynamic activities are higher than the mentioned threshold label. Note again that this movement highlight can distinguish just the presence of activity substance inside the given progressive shots. To verify if "rough" occasions are included, we likewise abuse other visual and sound highlights[7].

2.3 Violence detection using dynamic activity of frames

At first, we analysed the frames (i.e, still images) of the video of any format say (.mp4, .avi, .mpg) using MatLabR2012b. Here, we have taken a bomb blast video from a Hollywood movie. We took 25 frames as one shot that means each shot duration is one second.

Now calculate the dynamic activity of each shot by summing the dynamic activity of each shot by summing the dynamic activities of each frame involved in that shot. Then , we plot the dynamic activity of video versus time duration of the video. Compare the audio & video part of each shot that means compare (Energy Vs Time) of the audio with Dynamic Activity of the video.

We have analysed the video saved as " E:\Child.mp4" in our experiment which is downloaded from youtube .Its URL is given below:

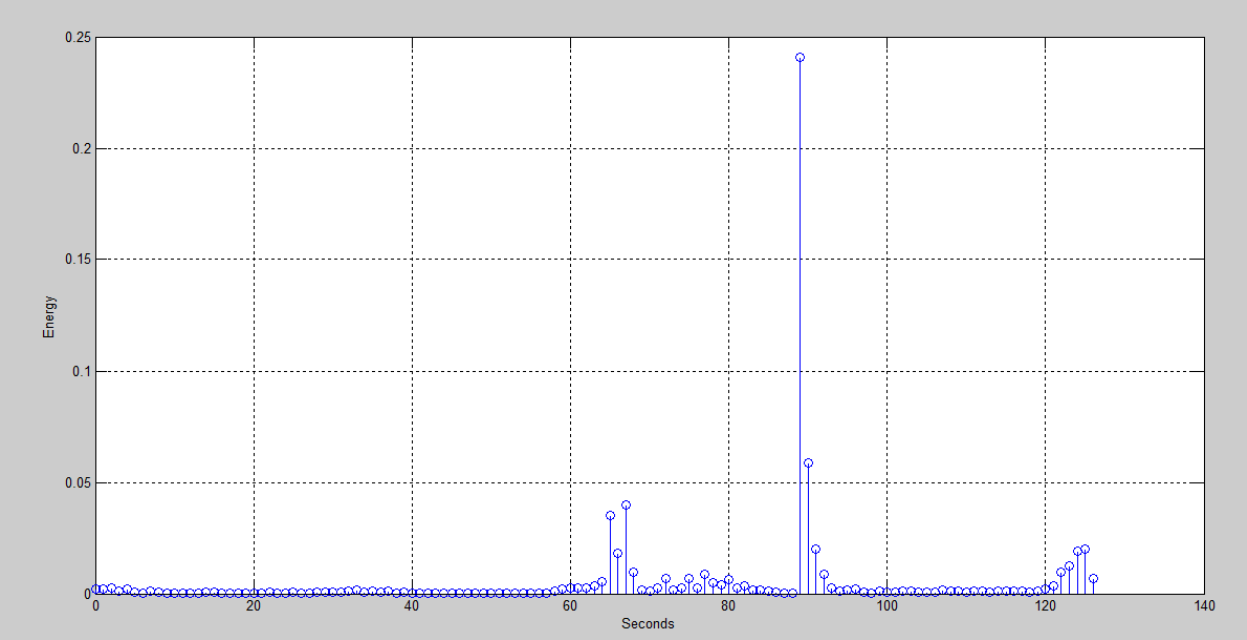
"**https://www.youtube.com/watch?v=VJivXSErhB8**" .We have compared both audio and video part.

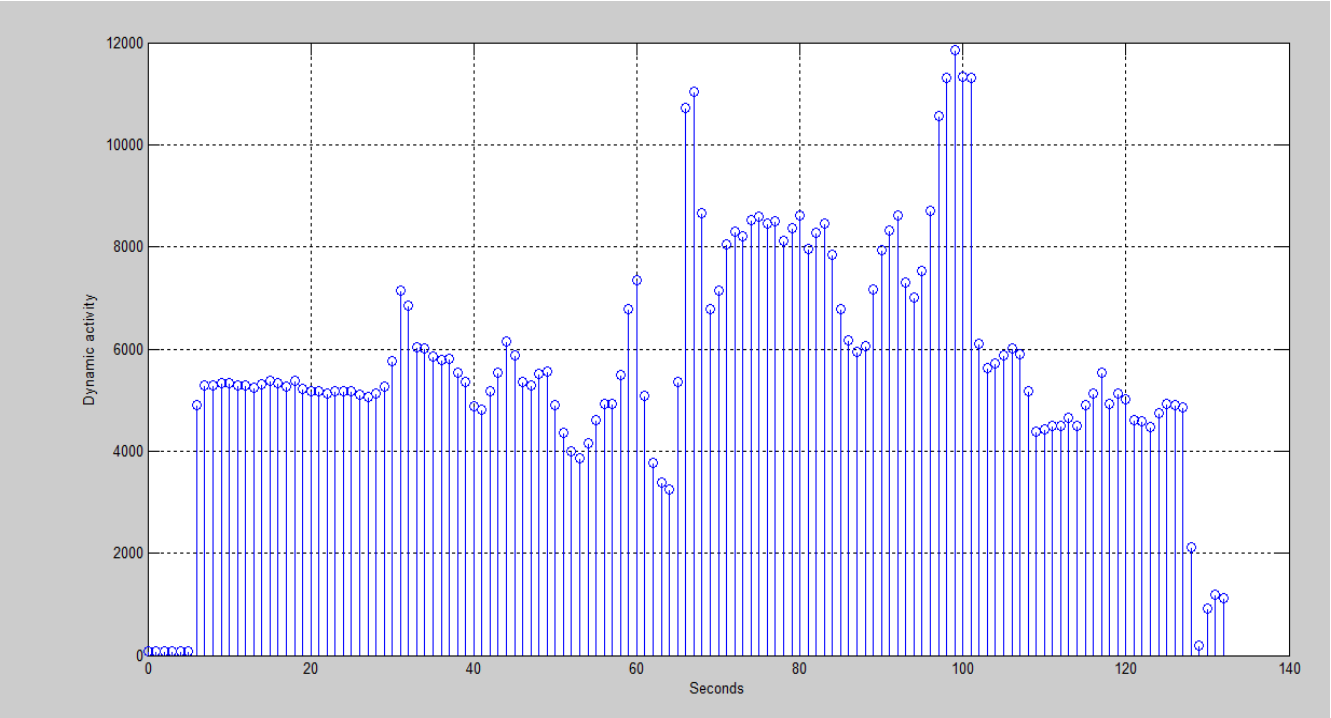
2.4 Matlab code

```
filename = 'E:\Child.mp4';

d = VideoReader(filename);

[m, n, k] = size(d);
```

```
nframes = d.NumberOfFrames;

time = nframes/23;

t=0: 1:(time-1);

m=d. Height;

n=d. Width;

f=23;

for i=1:time

for  q=1:23

  vd (: , : ,q)=double(rgb2gray(read(d,(q+f*(i-1)))));

end

s(i)=sum(sum(sum(vd(:,:,1:23))))/230000;

end

stem(t,s); xlabel('Seconds'); ylabel('Dynamic activity') ;grid on;
```

2.5 Results & graphs





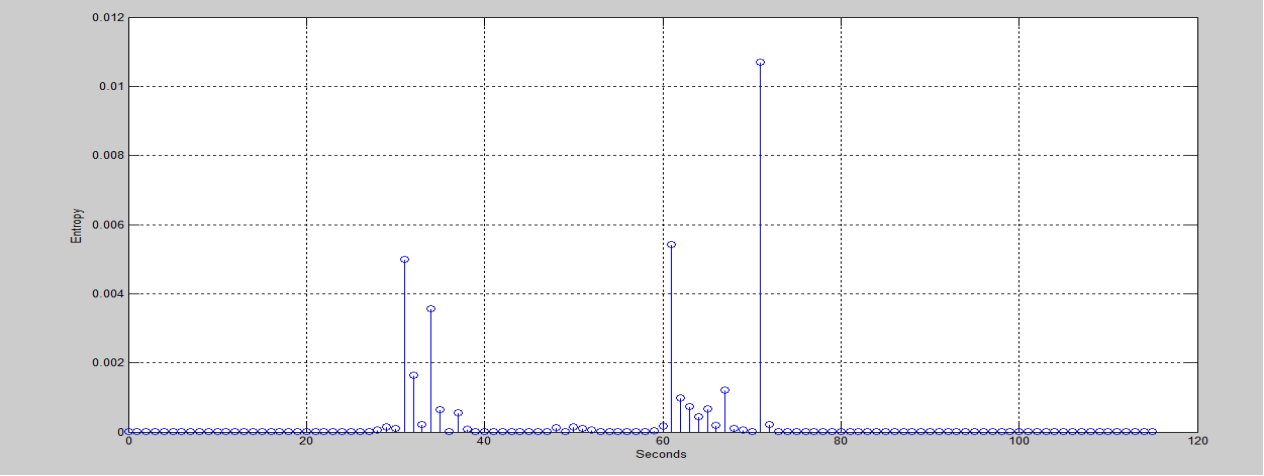
[Plot of energy of the audio & dynamic activity of video of the mentioned audio-visual signal]

[Figure 2.1]

2.6 Analysis of dynamic activity frame by frame

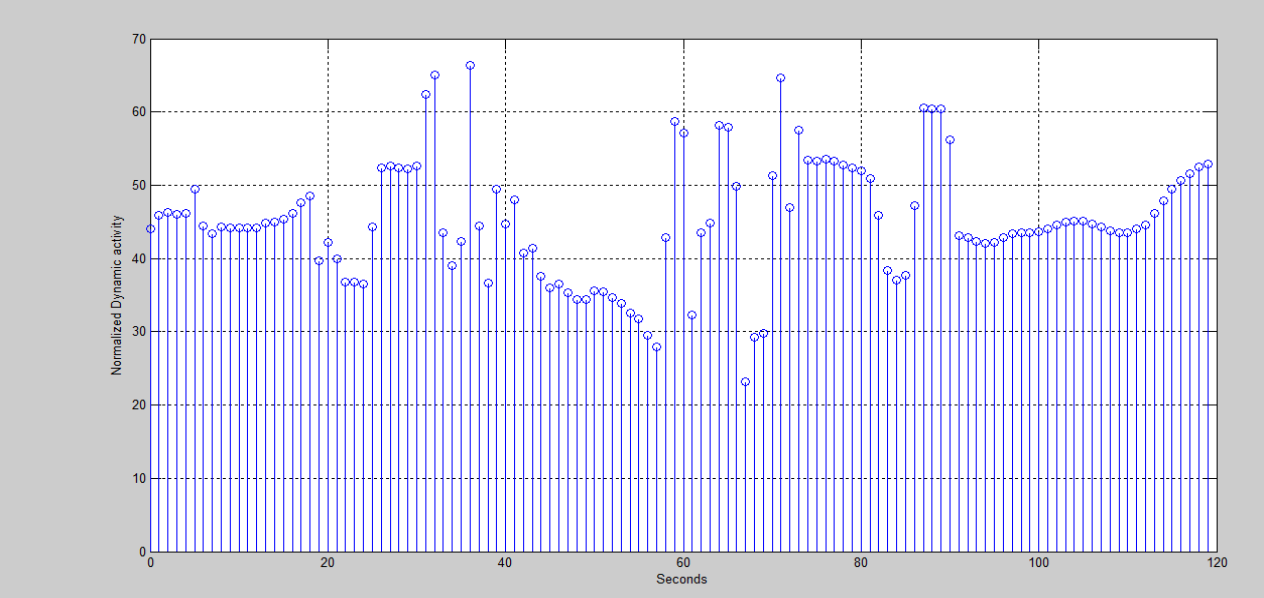
From the above figure, we can see that there is an abrupt change in dynamic activity levels of two consecutive time instances. Hence, we opted for frame to frame analysis of dynamic activity of audio-visual signal. Here, we have taken another video from “Drive-Motel Scene (HD)” which is of duration 116 seconds .Here, we have analyzed the audio-visual signal from 24second to 40 second elaborately.

2.7 Entropy of the audio & Dynamic activity of the video



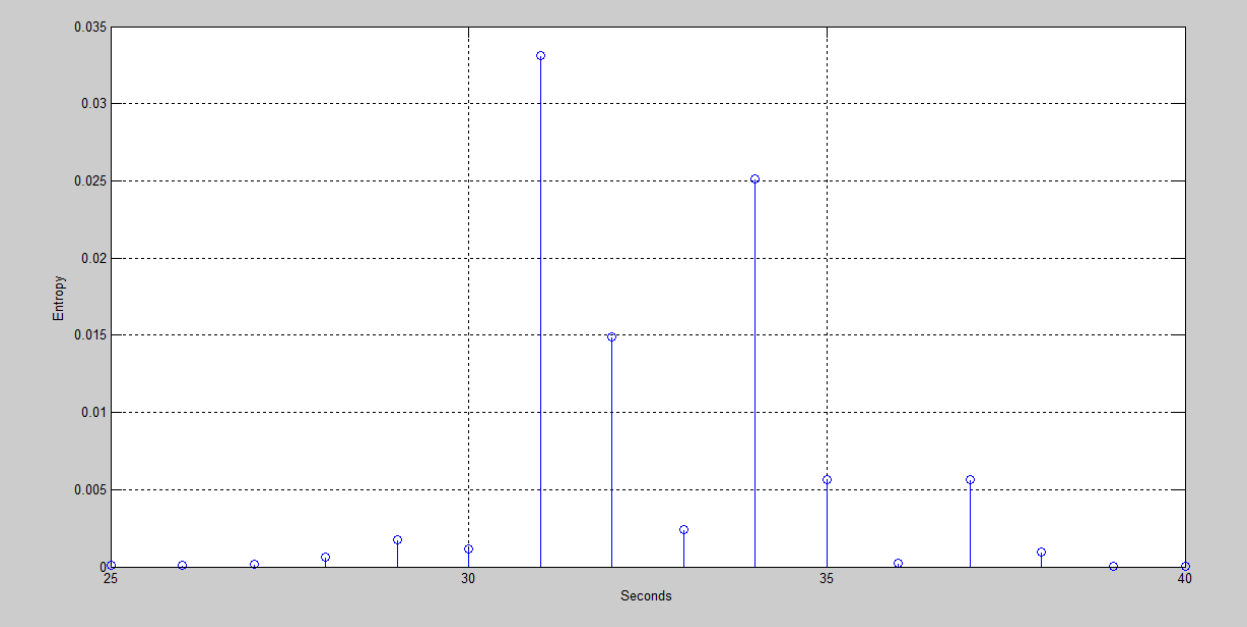
[Energy entropy of the above mentioned video]

[Figure 1.3]



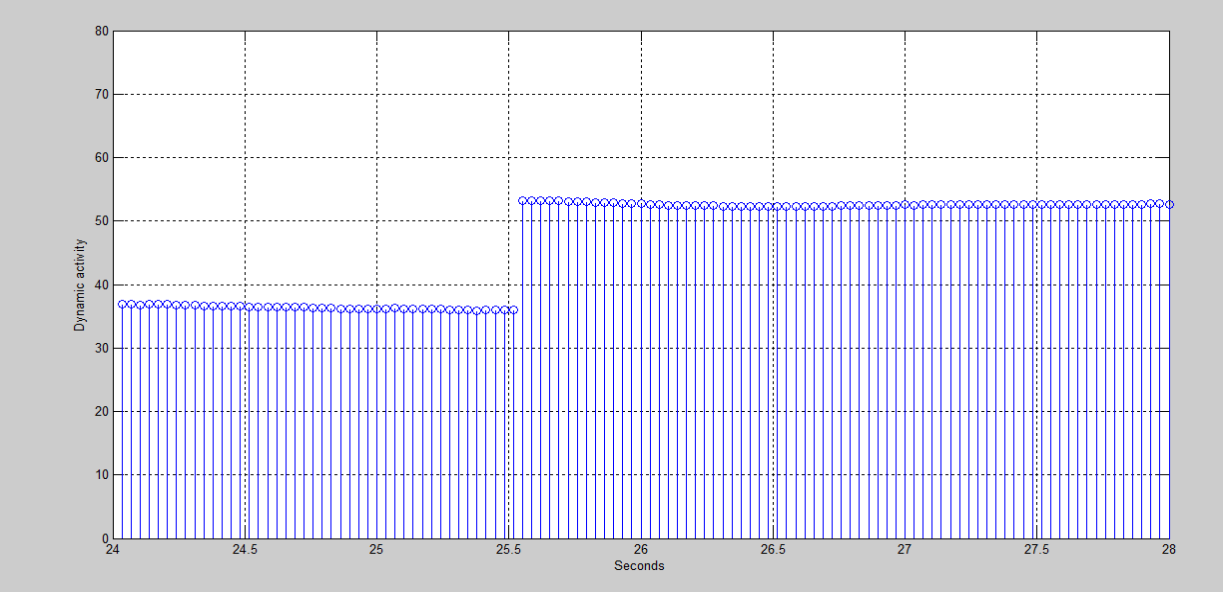
[Dynamic activity of the video]

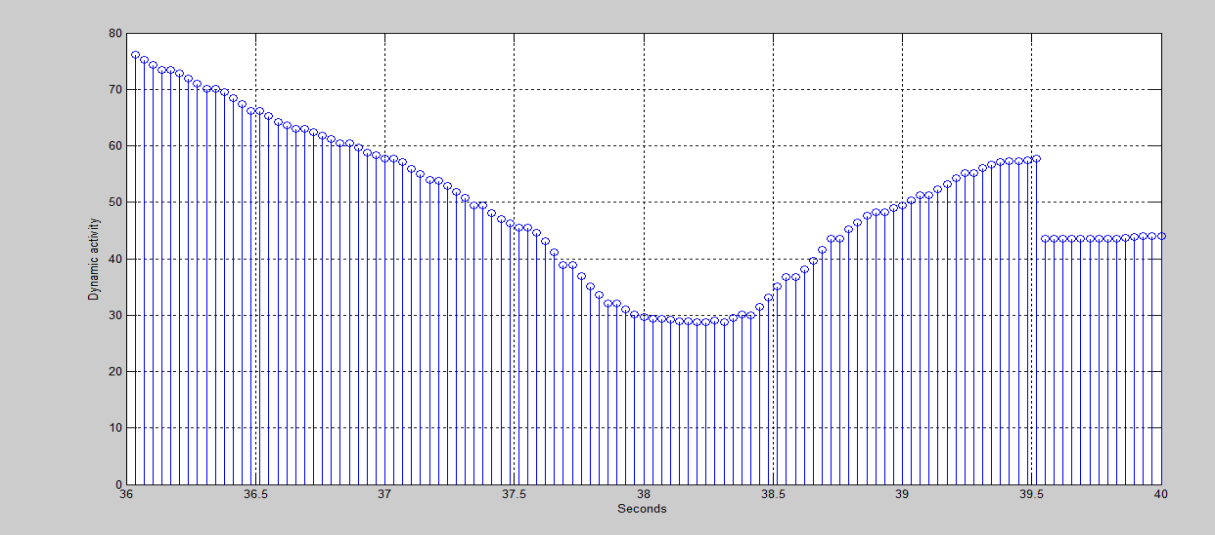
[Figure 2.2]



[Energy entropy of the audio-visual signal from 25 sec to 40 sec]

[Figure 2.3]





[Dynamic activity of each frame from 28sec to 40 sec]

[Figure 2.4]

3. BLOOD DETECTION

Under our meaning of viciousness, the assignment of fierce scene recognition is divided into action scene identification and bloody frame identification .While past methodologies tended to on shot level of video structure just, our methodology chips away at more semantic complete scene structure of video. The input digital video is initially divided into a few scenes. In view of film making attributes of activity scene, a few highlights of the scene are removed to nourish into the bolster feature machine for order.

At last, the blood and motion intensity data are coordinated to figure out if the activity scene has violence substance. Examination results demonstrate that the proposed methodology lives up to expectations sensibly well in recognizing the vast majority of the savage scenes.

3.1 Bloody frame detection

For every recognized activity scene, we further verify if it contains the bloody frame or not. As scene is made out of a few shots, key frames can be obtained from every shot[8]. Key frame is the frame which can speak to the notable substance of the shot. At that point, we figure out if one of these key frames has bleeding substance. We first distinguish the presence of human and blood in the key frame. There are some advanced procedures to identify particular color object, for example, fire and skin. Be that as it may, for the blood pixel discovery undertaking, the most straightforward strategy is to characterize blood-shading bunch choice limits for RGB color space.

3.2 Threshold range

Ranges of threshold values for each color space component are defined and the image pixel with values that fall within these predefined ranges is detected as blood pixel. According to our observation, there are two types of blood color: bright red and dark red. The ranges of RGB color space are defined as (1) $170 > R > 80$ & $G < 5$ & $B < 5$

(2) 200 > R> 120 & G< 90 & B < 90

In our experiment, we also consider the threshold range for hsv color space .the range for bright red is 0.05<hue<0.97 & saturation>0.3 & value>0.01

A key frame is detected as bloody frame if it satisfies the following conditions[9]:

I. It contains both face and blood areas.

II. The average motion intensity of each blood and face region is greater than a threshold.

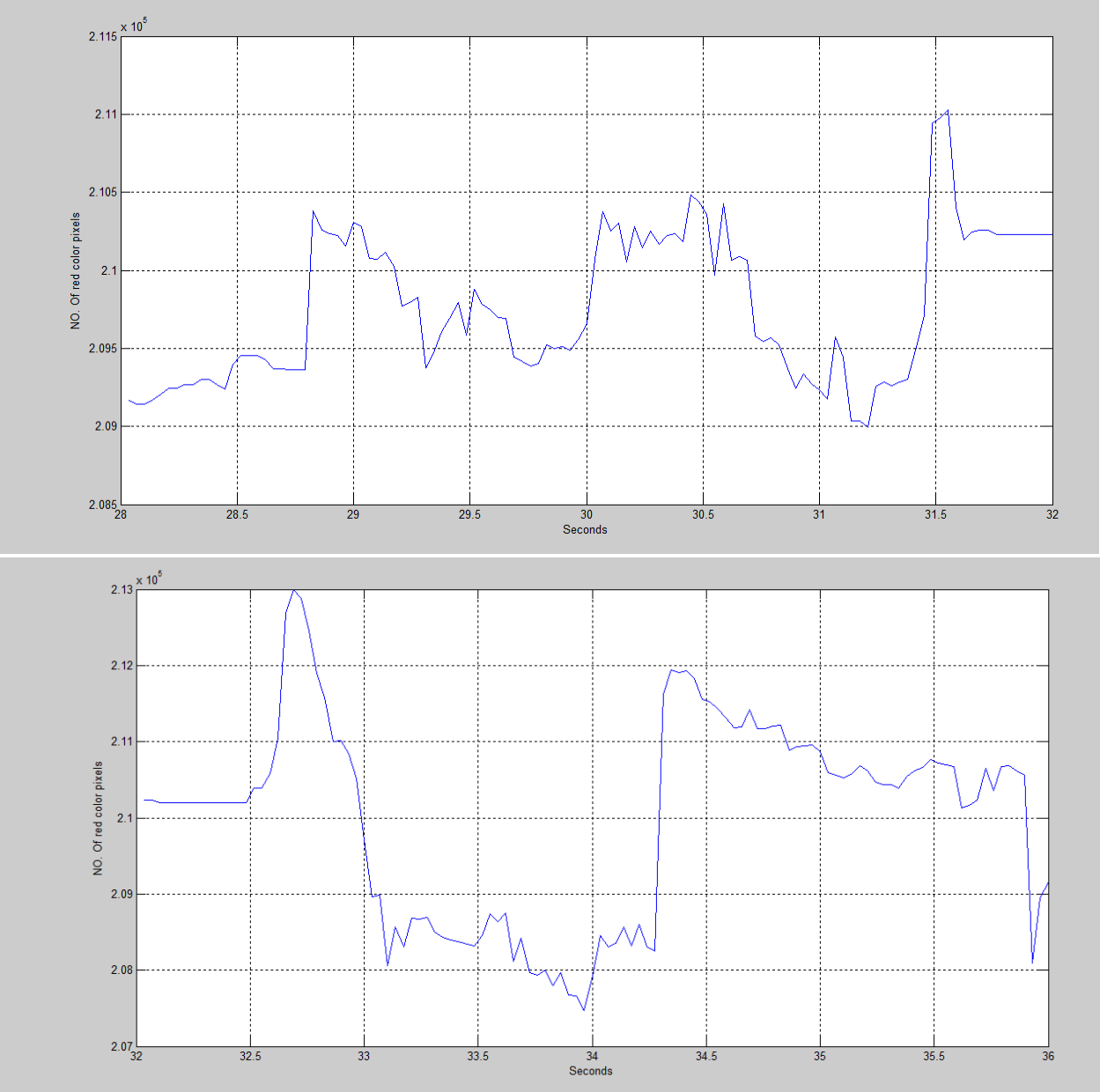
We have done our experiment on the previous mentioned video (Drive-Motel Scene).First, we detected the frames having maximum red color pixel. Then using “Color Threshold” in MATLAB R2014b, we detect the blood part in the frame.

3.3 Matlab code

```
xyloObj = VideoReader('E:\Drive.mp4');

nFrames = xyloObj.NumberOfFrames;

time=nFrames/29;

vidHeight = xyloObj.Height;

vidWidth = xyloObj.Width;

yy= VideoReader('E:\Drive.mp4');

 p=116;

mov(1:116) = ...

 struct('cdata',zeros(vidHeight,vidWidth, 3,'uint8'),...

        'colormap',[]);

count=[zeros(p,1)];

for k = 1 : 116

    mov(k).cdata = read(xyloObj,(k+928));

    h=rgb2hsv(mov(k).cdata);
```

```
    hu(:,:,1)=h(:,:,1);
    sa(:,:,1)=h(:,:,2);
    va(:,:,1)=h(:,:,3);
    n=0;
   for I = 1: vidHeight
        for j = 1:vidWidth
    if ((0.05< hu(I,j) < 0.97)&&(0.3 < sa(I,j) <1)&&(0.01 < va(I,j)<1))
                n=n+1;
           end
        end
 end
 count(k,1)=n;
end
t = 929:1044; plot(t/29,count);
grid on; xlabel ('Seconds');ylabel('NO. Of red color pixels');
```

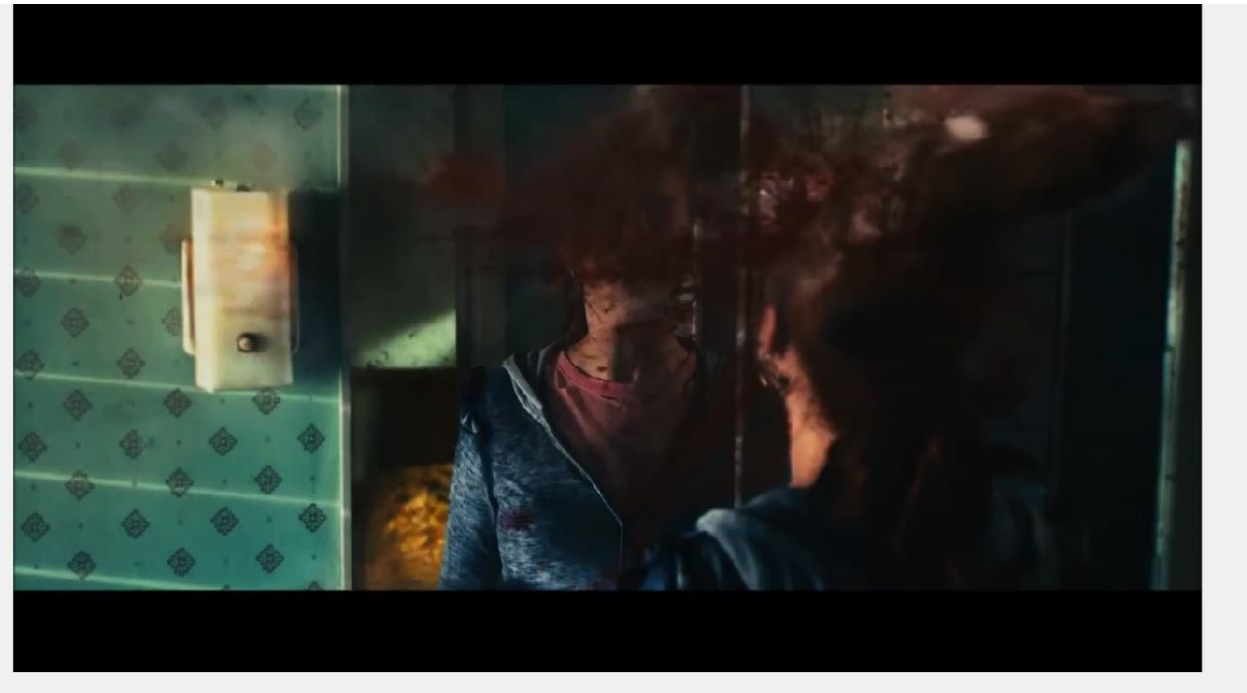
3.4 Variation of number of red color pixels with respect to frames



[No. of Red color pixels in frames from 28sec to 36 sec]

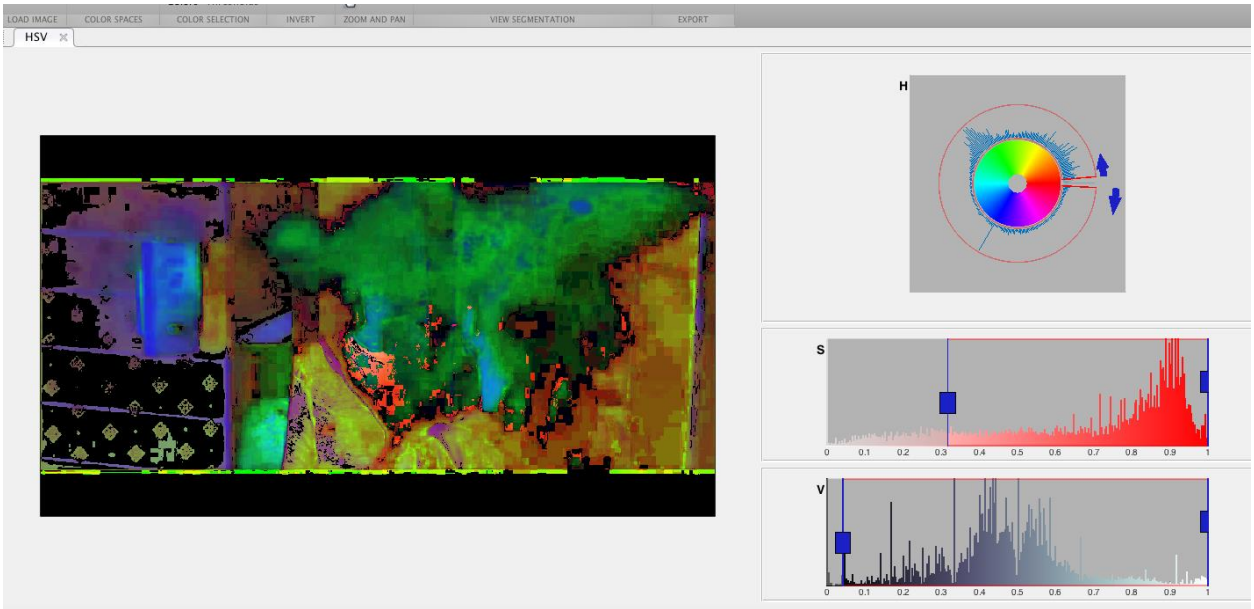
[Figure 3.1]

3.5 Images of blood detection



[Image of 983th frame]

[Fig. 3.2]



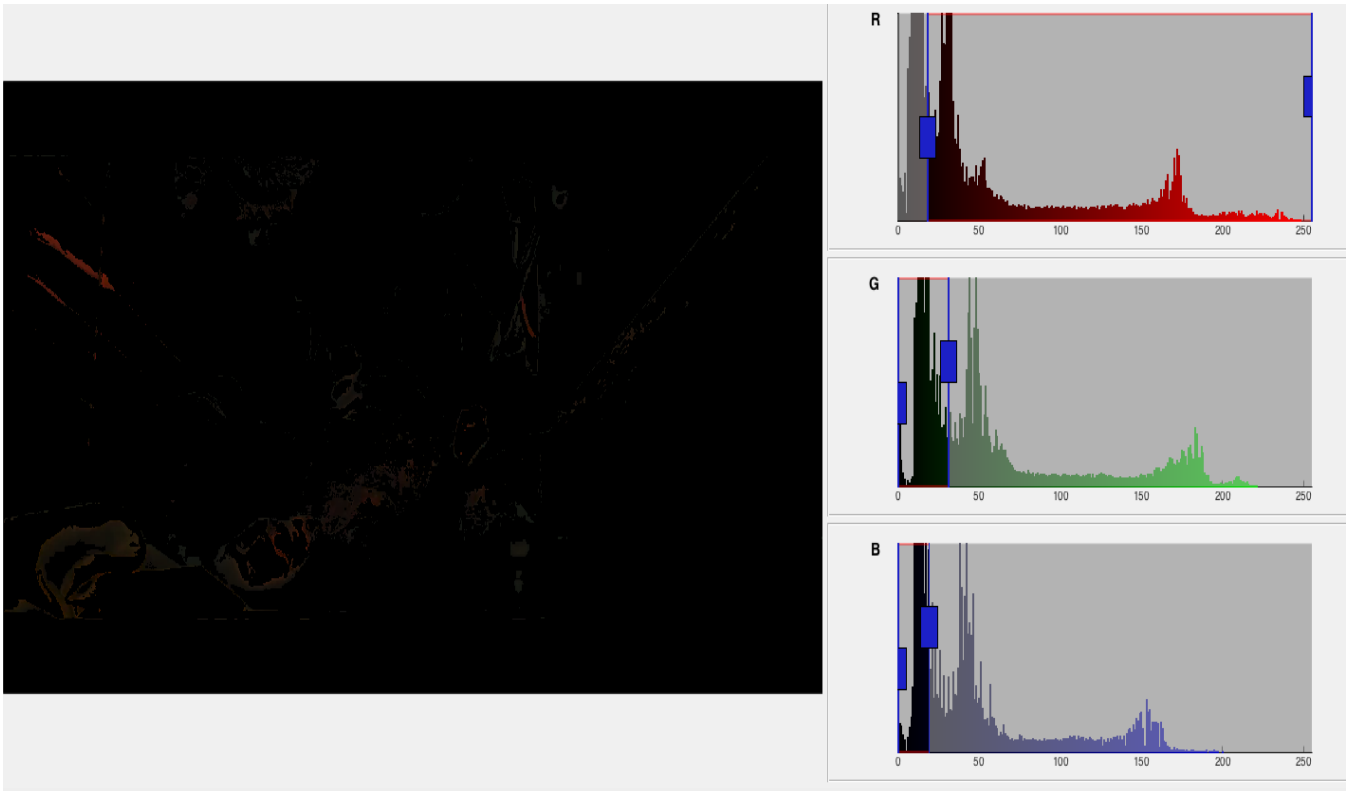
[Blood content in the above shown Frame]

[Figure 3.3]

From the plot result, we have scene that at time instant 32.75second there is maximum no. of red color pixel. We analyze the corresponding frame i.e, 983^{th} frame. The green region in the above plot shows the blood regions. Similarly, we analyze 2088^{th} frame.

[Image of 2088^{th} Frame]

[Figure 3.4]



[Blood Content in the above mentioned Frame]

[Figure 3.5]

4 CONCLUSION

In this piece of work we have presented a technique to detect the violence contents in an audio-visual signals by means three major factors: energy entropy of the audio part, dynamic activity of video & blood content. We integrate these three factors to detect vicious events that means if the energy entropy at any time is maximum and also the dynamic activity is also in peak ,then we go for blood detection in that frame. If we get blood content in that frame, then we are ensured of occurrence of violence at that moment.

BIBLOGRAPHY

[1] J . Nam and A. H. Tewfik ‘‘combined Audio and Visual Streams Analysis for Video Sequence Segmentation,” *IEEE ICASSP,* 1997.

[2] B. F. Kawin, “How Movies Work ”, University of California Press, Ltd., London, England, 1992.

[3] S. Pfeiffer, S. Fischer and W. Effelsberg, “Automatic Audio Content Analysis,” *Proc. of ACM Multimedia Conference,* 1996.

[4] D. Sinha and A. H. Tewfik, “Low Bit Rate Transparent Audio Compression using Adapted Wavelets,” *IEEE Trans. on Signal Processing,* vol. 41, no. 12, Dec. 1993.

[5] A.H. Tewfik. "Audio-visual content-based violent scene characterization", Proceedings 1998 International Conference on Image Processing ICIP98 (Cat No 98CB36269) ICIP- 98, 1998.

[6] Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, D. T. Lee, Shih-Fu Chang “Joint Audio-Visual Words for Violent Scenes Detection in Movies”, Dept. of Computer Science and Information Engineering, National Taiwan University,2012.

[7] B. L. Ye0 and B. Liu, “Rapid Scene Analysis on Compressed Videos,” *IEEE Trans. on Circuits and Systems For Video Technology,* vol. 5, no. ***6,*** pp.533-544, Dec. 1995.

[8] Liang-Hua Chen, Hsi-Wen Hsu and Li-Yun Wang Department of Computer Science , 2011 Eighth International Conference Computer Graphics, Imaging and Visualization.

[9] Christine T. Clarin, Judith Ann M. Dionisio Michael T. Echavez, Prospero C. Naval, Jr. Computer Vision & Machine Intelligence Group, Department of Computer Science College of Engineering University of the Philippines- Diliman “DOVE: Detection of Movie Violence using Motion Intensity Analysis on Skin and Blood”.