**“Video Shot Detection and Summarization”**

*A BACHELOR’S MINI PROJECT*

*submitted in partial fulfillment*

*of the requirements for the completion of the* ***6th*** *semester*

*of the*

**UNDER GRADUATE PROGRAM**

*in*

**INFORMATION TECHNOLOGY**

**(B.Tech in IT)**

*Submitted by*

*Prateek Porwal(IIT2010035)*

*Shahbaz Khan(IIT2010039)*

*Rakesh Kumar(IIT2010076)*

*Under the Guidance of:*

**Prof.Uma Shankar Tiwary**

IIIT-Allahabad

****

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY**

**ALLAHABAD-211012**

**May, 2013**

**CANDIDATE’S DECLARATION**

I hereby declare that the work presented in this project entitled “**Video Shot**

**Detection and Summarization**”, submitted in the partial fulfillment of the completion

of the semester 6th of Bachelor of Technology (B.Tech) program, in Information

Technology at Indian Institute of Information Technology, Allahabad, is an authentic

record of my original work carried out under the guidance of **Prof.U.S Tiwary**

due acknowledgements have been made in the text of the project to all other material

used. This semester work was done in full compliance with the requirements and

constraints of the prescribed curriculum.

Place: Allahabad **Name of the students**

Date: Prateek Porwal

Shahbaz Khan

Rakesh Kumar

**CERTIFICATE FROM SUPERVISOR**

I/We do hereby recommend that the mini project report prepared under my/our

supervision by Prateek Porwal, Shahbaz Khan and Rakesh Kumar titled “**Video Shot**

**Detection and**  **Summarization**” be **accepted** in the partial fulfillment of the

requirements of the completion of 6th semester of Bachelor of Technology in

Information Technology **for Examination**

Date: **Guide’s name**

Place: Allahabad Prof.U.S.Tiwary, IIITA

Committee for Evaluation of the Thesis

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENTS**

The author would like to express his sincere gratitude to **Prof.U.S.Tiwary**

Place: Allahabad **Name of the students**

**Prateek Porwal(IIT2010035)**

**Shahbaz Khan(IIT2010039)**

**Rakesh Kumar(IIT2010076)**

Date:

B Tech 3rd Year, IIITA

**ABSTRACT**

Video shot detection is the processor-intensive task of splitting a video into continuous

Shots with hard or shot cuts as the boundaries. After shots are detected those

shots are summarized based on key-frames extraction techniques. Those detected shots

have been widely used in video summarization for video analysis, indexing and

browsing. Video summarization has been categorized based on two approaches

namely : static video summarization and motion-based video summarization. Many

algorithms have been proposed for detecting video shot boundaries and classifying

shots and shot transition types. This report presents comparison on some of the

important algorithms for video shot detection like histogram difference ,statistical

difference ,pixel difference ,edge change ratio ,adaptive threshold ,compression

differences ,principal co-ordinate system and some summarization techniques like

sampling-based Key-frame extraction and shot-based key-frame extraction,

color-based approach, motion-based approach and segment-based key-frame extraction.

We have proposed the histogram difference technique for video shot detection.

**Table of Contents**

1. **Introduction……………………………………………………………………1**
   1. Currently existing technologies…………………………………
   2. Analysis of previous research in this area……………………
   3. Problem definition and scope.……………………………………
   4. Formulation of the present problem………………………………
   5. Organization of the thesis…………………………………………
2. **Description of Hardware and Software Used…………………………**
   1. Hardware…………………………………………………………
   2. Software………………………………………………………………
3. **Theoretical Tools – Analysis and Development………………………**

3.1 ……………………………………………………………………….

3.2 ………………………………............................................................

3.3 …………………………………………………………………………

3.4 ………………………………………………………………………

1. **Development of Software ………………………………………....................**
   1. ……………………………………………………………………………
   2. ………………………………………………………………………
   3. …………………………………………………………………
   4. ……………………………………………….........................................
2. **Testing and Analysis…………………………………………………………**
   1. ………………………………………………………………………
   2. …………………………………………………………………………
3. **Conclusions…………………………………………………………………**
4. **Recommendations and Future Work……………………………………**

**Appendix - Explanation of the Source Code………………………………**

**References……………………………………………………………………**

1. **INTRODUCTION**
   1. **Currently existing technologies**

**Shot detection**

Shot Detection in a video sequence is a process of identifying visual discontinuities along the time domain. During this process, it is required to extract visual features that measure the degree of similarity between frames in a given shot. Shot is defined as a sequence of frames captured from a single camera operation. The shot detection approaches are classified based on the features and methods used to solve the problem and other approaches are described.

1. **Histogram di**ff**erences**

Histograms are the most common method used for comparing images. And this technique can be extended to find the image whose histogram varies significantly from the previous image histogram. Thus we can detect shot boundaries. The histogram method computes 64 bin gray level histograms of the two images and Euclidean distance measure is used to find the histogram difference. If this distance between the two histograms is above a threshold, a shot boundary is assumed.

1. **Region based Histogram di**ff**erences**

In this method each image is divided into 16 blocks in a 4x4 pattern. For every image, a 64-bin gray-scale histogram is computed for each region. A Euclidean distance measure is used to find the difference between the region histograms of two consecutive images. If the distance is above a threshold, the region count for that image is incremented. If the region count is more than some predefined threshold, a shot boundary is assumed.

**1.2 Analysis of previous research in this area**

1. **Statistical di**ff**erences**

In this method, each image is divided into 16 blocks. Then, for each block the mean and the standard deviation is found. A Euclidean distance measure is used to find the mean and standard deviation differences between the corresponding blocks of two consecutive images. Similar to the previous method, two thresholds are defined. If the region count is more than the second threshold, a shot boundary is assumed.

**2.Pixel di**ff**erences**

This is the simplest method for determining shot boundaries. The difference between corresponding pixels of two consecutive images is computed. If the difference is greater than some threshold, than a shot boundary is assumed. All the above approaches can be improved to give good results by using other color spaces like RGB, HSV and YIQ instead of grayscale. And besides Euclidean distance, Chi-square distance measure can be used.

**3.Adaptive threshold**

The above approaches use static thresholds, instead Yusoff use adaptive threshold setting, by using statistics of the dissimilarity measure within a sliding window. Three different models, constant covariance model, proportional variance model and Dugad models are used for setting the threshold. Adaptive thresholding gives a better rate of detection shots than the static threshold method.

1. **Principal Co-ordinate System**

This is a new approach given by Alper Yilmaz . An eigen space decomposition of the RGB color space is used to describe the frames in a more descriptive coordinate system. The method is based on calculation of principal coordinate system from spatial information determined from RGB color space by applying eigen space decomposition. In this method covariance matrix of the color space is computed, then the eigenvectors are determined for that matrix and the maximum eigen valued eigenvector is used to determine the shot boundaries. This approach is well suited for real time surveillance

1. **Compression Differences**

Little et al used differences in the size of JPEG compressed frames to detect shot boundaries as a supplement to a manual indexing system. Arman, Hsu, and Chiu found shot boundaries by comparing a small number of connected regions. They used differences in the discrete cosine transform (DCT) coefficients of JPEG

compressed frames as their measure of frame similarity, thus avoiding the need to decompress the frames. A further speedup was obtained by sampling the frames temporally and using a form of binary search to find the actual boundary. Potential boundaries were checked using a color histogram difference method.

**Video summary**

These are the different approaches of video summary. One of the approach in a video summary is to select keyframes (a set of salient images) from the original video sequence. Based on the way the key frames are constructed, there are 3 classes of keyframe extraction:

1. **Sampling-based Key-frame Extraction**

Most of earlier work in video summarization choose to select key-frames by randomly or uniformly sampling the video frames from the original sequence at certain time intervals, which was applied in the Video Magnifier, MiniVideo system. Although this is probably the simplest way to extract key-frames, the drawback is that such an arrangement may cause some short yet important segments to have no representative frames while longer segments could have multiple frames with similar content, thus failing to capture and represent the actual video content.

1. **Shot-based key-frame Extraction**

More sophisticated work tends to extract key-frames by adapting to the dynamic video content. Since a shot is defined as a video segment within a continuous capture period, a natural and straightforward way of key-frame extraction is to use the first frame of each shot as its key-frame . However, while being sufficient for

stationary shots, one key-frame per shot does not provide an acceptable representation of dynamic visual content, therefore multiple key-frames need to be extracted by adapting to the underlying semantic content. However, since computer vision still remains to be a very difficult research challenge, most of existing work chooses to interpret the content by employing some low-level visual features such as color and motion, instead of performing a tough semantic understanding. Based on the features that the approaches are into 4 different classes: color-based approach, motion-based approach, mosaic-based approach.

1. **Color-based Approach**

The method proposed by Zhang , the keyframes are extracted in a sequential fashion

for each shot. Particularly, the first frame within the shot is always chosen as the

first keyframe, then the color-histogram difference between the subsequent frames

and the latest keyframe is computed. Once the difference exceeds a certain threshold,

a new keyframe will be declared. A similar work is also reported by Yeung and Liu.

One possible problem with above extraction methods is that there is a probability

that the first frame is a part of transition effect at the shot boundary, thus

strongly reducing its representative quality. In Huang from University of Illinois

at Urbana-Champaig propose to extract the keyframes using an unsupervised

clustering scheme. Basically, all video frames within a shot are first clustered into

certain number of clusters based on the color-histogram similarity comparison where

a threshold is predefined to control the density of each cluster. Next, all the clusters

that are big enough are considered as the key clusters, and a representative

frame closest to the cluster centoid is extracted from each of them. Ferman and

Tekalo reported a similar work in this. Because the color histogram is invariant to

image orientations and robust to background noises, color-based keyframe

extraction algorithms have been widely used. However, most of these works are

heavily threshold-dependent, and cannot well capture the underlying dynamics

when there is lots of camera or object motion.

1. **Motion-based Approach**

Motion-based approaches are relatively better suited for controlling the number of frames based on temporal dynamics in the scene. In general, pixel-based image differences or optical flow computation are commonly used in this approach. In Wolfs work , the optical flow for each frame is first computed, and then a simple motion metric is computed. Finally by analyzing the metric as a function of time, the frames at the local minima of motion are selected as the keyframes. A domain specific keyframe selection method is proposed where a summary is generated for video-taped presentations. Toklu and Liou from Siemens Corporate Research reported their work where 3 different operation levels are suggested based on the available machine resources: at the lowest level, pixel-based frame differences are computed to generate the temporal activity curve since it requires minimal resources; at level 2, color histogram-based frame differences are computed to extract color activity segments, and at level 3 sophisticated camera motion analysis is carried out to estimate the camera parameters and detect the motion activity segments. Keyframes are then selected from each segment and necessary elimination is applied to obtain the final result.

**5.Mosaic-based Approach**

A limitation of above approaches is that it is not always possible to select the keyframes that can represent the entire video content well. For example, given a camera panning/tilting sequence, even if multiple keyframes are selected, the underlying dynamics still couldnot be well captured. In this case, the mosaic-based approach can be employed to generate a synthesized panoramic image that can represent the entire content in an intuitive manner. Mosaics, also known as salient stills, video sprites or video layers, are usually generated in the following 2 steps:

1. Fitting a global motion model to the motion between each pair of successive frames;
2. Compositing the images into a single panoramic image by warping the images with the estimated camera parameters.

**6. Segment-based Key-frame Extraction**

One major drawback of using one or more key-frames for each shot is that it does not scale up for long videos since scrolling through hundreds of images is still time-consuming, tedious and ineffective. Therefore, recently more and more people begin to work on higher-level video unit, called as video segment. It could be a scene, an event, or even the entire sequence. In this context, the segment-based key-frame set will surely become more concise than the shot- based keyframe set. Uchihashi first cluster all video frames into a predefined number of clusters, and then the entire video is segmented by determining to which cluster the frames of a contiguous segment belong. Next an importance measure is computed for each segment based on its length and rarity, and all segments with their importance lower than a certain threshold will be discarded. The frame that is closest to the center of each qualified segment is then extracted as the representative keyframe, with the image size proportional to its importance index. Finally, a frame-packing algorithm is proposed to efficiently pack the extracted frames into a pictorial summary.

* 1. **Problem definition and scope**

Shot detection: given a video V consisting of n shots, we have to find the beginning and end of each shot

Also known as shot boundary detection or transition detection It is fundamental to any kind of video analysis and video application since segmentation of a video into its basic components: the shots.

Those detected video shots are summarized further for video processing indexing and browsing.

* 1. **Formulation of the present problem**

* 1. Organization of the thesis

Our thesis of project development comprises of two parts:-

1)Video Shot detection:-We studied several shot detection algorithms for the development of our project.Out of them we proposed to use histogram difference because this technique is far more better than other techniques in terms of detection of shots accurately and relative errors.

We started doing code for the histogram difference in opencv on codeblocks compiler.

2)Video Summarization:-shot detection technique is further utilized for video summarization for indexing,browsing and video processing.After shots are deteted correctly,key-frames of shots are combined together in a proper sequence in order to create the summary of the entire original video using key-frame based extraction technique.

1. **Description of Hardware and Software Used**

**2.1 Hardware**

Hardware specification:

32 bit CPU processor

2 GB RAM

**2.2 Software**

Software specification:

Windows operating system 32 bit

Open CV 2.3.0

Codeblocks 10.0.5

1. **Theoretical Tools – Analysis and Development**

Some of the applications of shot detection are:

i) Managing video databases for indexing, browsing, searching and

summarization.

ii) Digital Video Library as a networked Internet application allowing for

Storage, cataloging, retrieval and uni-casting video sequences. Scene,

event or object detection.

Video Summarization: Video summary is an abstract of a longer video document

in a shorter period of time. It is a method of presenting the content of a video in a

condensed manner without changing the original video.

**Terminologies:**

**Hard cuts:** A cut is an instantaneous transition from one scene to the next. There

are no transitional frames between 2 shots.

**Fades:** A fade is a gradual transition between a scene and a constant image (fade-out) or between a constant image and a scene (fade-in).

**Dissolves:** A dissolve is a gradual transition from one scene to another, in which the first scene fades out and the second scene fades in.

**Wipe:** another common scene break is a wipe, in which a line moves across the screen, with the new scene appearing behind the line.

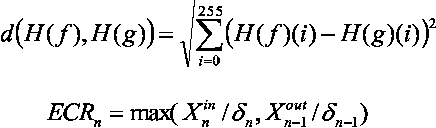
**Schema of Cut Detection:**

Calculate a time series of discontinuity feature values f(n) for each frame. Suppose we use function d(x,y) to measure the dissimilarity between frame x and y. The discontinuity feature value for frame n is f(n)=d(n- 1 ,n).

Pick the cuts position from f(n) based on some threshold techniques. **Features to Measure Dissimilarity:**

1. Intensity/color histogram

Edges/contours: Based on edge change ratio (ECR). Let σn be the number of edge pixels in frame n, and Xnin and Xn-1 out. the number of entering and exiting edge pixels in frames in frames n and n-1, respectively. The edge change ratio ECRn between frames n-1 and n is defined as:



1. Edges/contours

How to define the entering and exiting edge pixels Xn in and Xn-1

Suppose we have 2 binary images en-1 and en. The entering edge pixels Xnin are the fraction of edge pixels in en which are more than a fixed distance r from the closest edge pixel in en-1. Similarly the exiting edge pixels are the fraction of edge pixels in en-1 which are farther than r away from the closest edge pixel in en.

**Thresholding**

1. Global threshold

A hard cut is declared each time the discontinuity value f(n) surpasses a

global thresholds.

1. Adaptive threshold

A hard cut is detected based on the difference of the current feature values f(n) from its local neighborhood. Generally this kind of method has 2 criteria for a hard cut declaration:

- F(n) takes the maximum value inside the neighborhood The difference between f(n) and its neighbors’ feature values is bigger than a given threshold.

_Pic4

**4 .DEVELOPMENT OF SOFTWARE**

1. **TESTING AND ANALYSIS**
2. **CONCLUSIONS**
3. **RECOMMENDATIONS AND FUTURE WORK**

**Appendix - Explanation of the Source Code**

**REFERENCES**

[1]M. Mills, “A magnifier tool for video data,” Proceedings of ACM Human Computer Interface, pp. 93–98, May 1992.

1. Y. Taniguchi, “An intuitive and efficient access interface to real-time in-coming video based on automatic indexing,” Proceedings of ACM Mul­timedia, pp. 25–33, November 1995.
2. K. Otsuji, Y. Tonomura, and Y. ohba, “Video browsing using brightness data,” in Proceedings of International Society for Optical Engineering, vol. 1606, 1991, pp. 980–985.
3. S. W. Smoliar and H. J. Zhang, “Content-based video indexing and retrieval,” IEEE transactions on Multimedia, pp. 62–72, 1994.
4. H. J. Zhang, C. Y. Low, and S. W. Smoliar, “Video parsing and browsing using compressed data,” IEEE transactions on Multimedia Tools and Applications, vol. 1, pp. 89–111, 1995.
5. Y.Pritch, A.Rav-Acha, and S. Peleg, “Non-chronological video synopsis and indexing,” IEEE Transactions on Pattern Analysis and Machine Intelligent, 2008.