Journal of Graphic Era University
Vol. 7, Issue 2, 90-107, 2019
ISSN: 0975-1416 (Print), 2456-4281 (Online)

# Video Segmentation Techniques for Instructional Videos – Survey

**Jyoti Parsola[1*], Durgaprasad Gangodkar[2], Ankush Mittal[3]**

[1]Departement of Computer Application
[2,3]Departement of Computer Science and Engineering
Graphic Era (Deemed to be University), Dehradun, Uttarakhand, India
[*]Corresponding author: jyotee.negi@gmail.com

(Received October 11, 2018; Accepted May 8, 2019)

**Abstract**
Low cost smart phones and easy internet access have caused an increase in viewership of e-learning video. Usually the memory size of mobile phones is less therefore, it becomes extremely important to reduce size of these instructional videos. Video segmentation is the fundamental task of reducing size of e-learning videos. This paper gives an overview of existing techniques used for video segmentation of e-learning videos. Most of the methods used so far for segmenting instructional video are broadly categorized into i) feature extraction based segmentation ii) motion based segmentation. The performance, comparative merits and limitations of each approach is thoroughly examined and contradicted. The analysis is beneficial for appropriate use of existing methods and for enhancing their performance or forming new methods on the basis of existing methods by combining one or two methods together.

**Keywords-** Video Segmentation, E-Learning Applications.

## 1. Introduction

With the advancement in internet technologies have equipped educationist with various means, alternative to conventional teaching method. E-Learning setting not only conquers the constraints of traditional classroom learning but, also broadens learning method by transferring the approach in whatever time or place to existence. Its objective is, to supply learners with excelling learning experience in their day to day learning environments. Mobile learning integrates advances from Electronic Learning and Mobile Computing. The essential and entire function of mobile computing technologies in mobile based learning is to construct a learning environment, where anyone is able to learn at anyplace and anytime. However, the e-learning application in-order to be effective the e-learning videos have to be clear and concise. In e-learning video the major challenge is to provide the learner, instinctive, clear and speedy access to the educational videos based on their interest (Amir et al., 2001).

Various techniques are emerging whose aim is to automate the distribution of learning material to improve the accessibility and to provide learners a good platform for learning. Some of the techniques are automatic presentation of lecture videos (Asghar et al. 2014; Liu et al., 2001), synchronization of slides to the lecture videos (Bianchi, 1998) intelligent classroom system for assisting the instructor (Mukhopadhyay and Smith, 1999). All the techniques mentioned are dependent on high tech equipment such as electronic whiteboards, moving camera etc. which are costly and malfunctioning of these equipment hinder the lecturers. Therefore, there is a need of such techniques which provide an easy medium to the lecturer to spread their knowledge in a cost efficient way. With the frequent use of mobile phones students now use it for learning purpose as well. Therefore, e-learning videos have to be memory efficient i.e. the size of the video should be minimal without losing the essential content. In this entire process, segmentation plays a key role. In this paper, we review the segmentation techniques used so far in instructional videos. We are presenting the analysis of the techniques, which are used for identifying or segmenting the region containing text, chalkboard area and power point presentation area etc.

## 2. Video Segmentation

A video consists of numerous video frame which is an image thus, an image is considered to be (Pal and Pal, 1993) two dimensional function *f*(*x*, *y*) or more appropriately to say a two dimensional matrix whose row *x* value and column *y* values represents a point which is called as pixel in digital image terms. Thus, a digital image can be represented as

$$F_{m\,X\,n} = [f(x,y)_{m,n}] \tag{1}$$

Where $m$ and $n$ is the size of the image or video frame, for sequence of images or frames it can be represented as

$$F_{m\,X\,n} = [f(x,y,t)_{m,n}] \tag{2}$$

Segmentation (Fu and Mui, 1981)of a grid *G* for a uniform predicate *Pre*( ) defined on a set of pixels G , then segmentation is a partition of set G into connected subset or regions $S_1, S_2, S_3, S_4$ such that,

$$\cup_i^n S_i = G \ with \ S_a \cap S_b = \emptyset \ \ (a \neq b) \tag{3}$$

The uniformity predicate $Pre(S_a) = true$ for all regions $S_a$ and $Pre(S_a \cap S_b) = False$ when $a \neq b$ and $S_a and\ S_b$ are neighbors. Almost all image segmentation technique proposed so far are adhoc in nature as they cannot work for all images. Thus, image segmentation techniques are to be selected based on the application.

The general framework of video segmentation of e-learning videos as shown in Figure 1 consists of raw lecture videos captured from a camera, in context to e-learning, videos frames are almost similar therefore rather than storing all frames, one frame per second is processed further for video analysis, the main function of the video analysis module is to segment video into region which consists of visual content and non-visual content area. Visual content here is referred to the text and figures in the board area. The video frames where there is change in the visual content area is saved and are forwarded to the server.

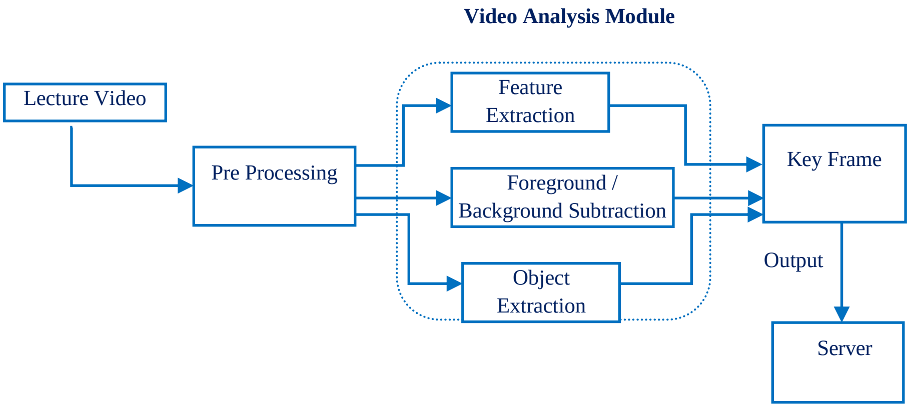


**Figure 1. General framework of video segmentation of e-learning videos**

Instructional video consists of power-point presentation, chalkboard presentation or a combination of both so extracting the meaningful content (text area) from these video is a challenging task. Moreover, the obstacles like noise or clutter present in these videos, lightning quality in the classroom, dissimilarity in the colour of chalkboard, disturbance or occlusion due to the random movement of instructor, shadow created due to the instructor makes the job more troublesome. To scrutinize this issue, video frames are first segmented into various regions based on some segmenting techniques and then board area or the region which consists of the text is identified and extracted. Finally, all the frames which consist of text area are saved for the purpose of e-learning application. A lot of research has been performed for extracting content from educational videos. Now in order to segment instructional videos it is very necessary to segment those regions in the frame which are relevant and irrelevant. It is quite obvious the chalk-board region is the relevant region in the frame. Thus our review focuses on those papers which aims in segmenting the chalk-board region.

## 3. Chalk Board Extractions of E-Learning Videos

In case of video frames data is almost similar therefore, rather than saving all the frames, frames with a certain time interval are saved for further processing. Educational videos consist of frames which are almost similar in-order to reduce the frame number and these frames are called as key frames. These videos contain the useful content only in the board area used by the teacher/ tutor. So frames which have changes in the chalk-board area are only considered and the rest of the frame is discarded. To identify board area region image segmentation is performed. Regions are divided into i) Foreground ii) Background region.

### 3.1 Foreground Region

In e-learning videos foreground is the region which is dynamic, the board area is the dynamic area which comprises of text and figures as drawn by the tutor/teacher. To identify it, is the tedious task as in e-learning videos apart from text and figures the tutor/teacher is also dynamic. The area covered by the tutor (Brejl and Sonka, 2000) is not useful for e-learning only the audio which consists of tutor/teacher narration about the topic. Therefore, foreground identification in e-learning videos is a challenge.

### 3.2 Background Region

Region which is static is the background region. The content of this region are not useful. In case of e-learning application board area is the background.

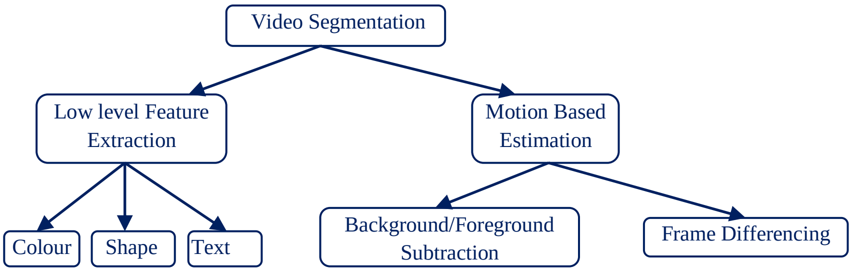


**Figure 2. Video segmentation techniques**

Our literature survey for segmentation in e-learning videos is categorized in two parts and is shown in Figure 2.

i) Low level features extraction for segmentation
  a) Color feature extraction
  b) Shape feature extraction
  c) Text detection
ii) Motion based segmentation
  a) Background/ Foreground Subtraction
  b) Frame Differencing.

## 4. Low Level Feature Extractions

### 4.1 Color Feature

One of the most important features to segment video frames is color and is one of the most widely used for image segmentation (Cheng et al., 2001). Color is the combination of red, green and blue are called the primary colors. For color feature extraction (Ju et al., 1998), assumed $P$ be the image and $\partial$ be pixels of $P$ . Then color extraction can be denoted as the function $f \rightarrow P{:}\,C$ where $C = \{c_1, c_2, c_3, \ldots c_n\}$ be the set of color, $f$ maps the pixel to $C$. According to them color histogram, a mixture of Gaussian models, color models, color coral grams etc are in the color based features. Color based feature extraction are dependent on the color spaces like HSV, RGB, YCBCR, HVC and normalized RG and YUV.

Video segmentation has to been done for segmenting board area from rest of the frame. Color is used for segmentation purpose because in lecture videos the board which is to be segmented is of a single color (green, black or white) and therefore it is easy to identify the non-board region. The blackboard region is segmented from rest of the region as discussed in Lin et al.

(2010), the frame image is converted to the $L*a*b*$ color space and then segmented into several regions by clustering pixels with similar colors based on the k-means algorithm. After extracting the blackboard region the contents are extracted.

Liu and Kender (2002) focused on lecture notes, they define semantic content as “ink pixels”, and present a low-level retrieval technique to extract this content from each frame with consideration of various occlusion and illumination effects. “Key frames” in this video genre are redefined as set of frames that cover the semantic content, and the fluctuating amount of visible ink is used to drive a heuristic real-time key frame extraction method. However the proposed work cannot identify the handwritten text accurately.

Video segmentation based on color histogram comparison of two images is been done by various researchers (Dong and Li, 2005; Chen et al., 2010). Although two histograms with different content can have same value. It can be computed with the following

$$\delta(k, k+1) = \sum_{i=0}^{i=n} |h_k(i) - h_{k+1}(i+1)| \tag{4}$$

where $h_k$ represents the frequency of the $i^{th}$ gray value of the $k^{th}$ frame of histogram, i represents the intensity of gray value, n represents the different gray value.

Chen et al. (2010), used histogram change between the two frames and detects long term change over time to identify the shot. Shot is referred as visual change during the presentation. Similarly, Haubold and Kender (2005), proposes two level of video segmentation i) Slide segmentation and ii) Text segmentation. For slide detection, it computes the histogram difference of two frames, when the difference value is large then, slide boundaries are identified and for topic segmentation, image matching is performed between the obtained slide image and video frame. The original video is segmented into various shots (Li and Dong 2006) calculated histogram difference of two frames and then for the segmented shots, used edge or contrast strength for computing the content of different video frame. Then high content frame are picked for the video capsule and the entire process is shown in Figure 3.

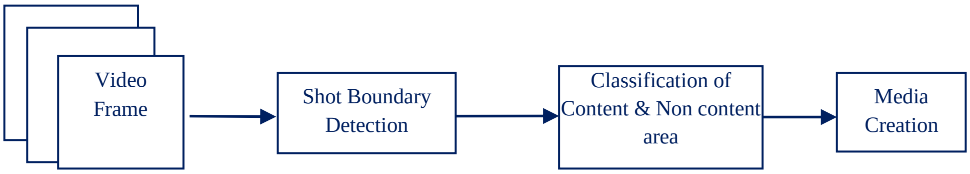


**Figure 3. Processes involved in the proposed technique in (Li and Dong, 2006)**

Similarly whiteboard color is matched with the color model (He and Zhang, 2006) to find the board area and model. Masneri and Schreer (2014) used color histogram and Wallick et al. (2005) used color classification to identify instructor and board pixels. Similarly, Dorai et al.

(2003) computed color histogram for each frame to capture the spread of pixels in the RGB color space. It uses a decision tree to classify between the slide text and text written in the whiteboard. The limitations of these approaches are that they cannot identify the blackboard area of any other color.

Various researchers rather than focusing on identifying or segmenting the writing board or the tutor based on color feature extraction they focused on the ink color. He and Zhang (2006), classifies image pixel into whiteboard background, pen strokes and occupied objects from video frames and newly pen strokes are extracted from the video frames and are white balanced for greater compression. Limitation of color feature is it fails when the outfit of the instructor is similar to that of a blackboard.

A feature selection algorithm is proposed by Liu and Kender (2002), which is used for video frame classification. Mukhqadhyay and Smith (1999) used feature extraction technique for segmentation of lecture videos which, comprises power point presentation slides. Video frames are low pass filtered and then adaptively threshold to generate a binary image. Then difference $\Delta$ is computed between the two binary images. The $\Delta$ is compared with the threshold. Threshold is set in such a manner that no slide is undetected. Choudary and Liu (2007) used mean shift algorithm proposed by (Ram and Chaudhuri, 2009) for segmenting video frames into connected regions and color of largest region is extracted and is assumed as the color of background.

### 4.2 Shape Feature

In instructional videos the region to be segmented is the chalk board which is square, rectangle in shape thus knowing this fact most of the researchers have used shape feature extraction, to segment chalk-board region form the rest of the frame. From the various shape based segmentation technique edge detection is used commonly (Pal and Pal, 1993). Edge transforms an image into an edge image benefits from change of gray tomes in image and as well as maintains the shape of the object in an image. It occurs where two regions have different intensity. Since it is low level feature therefore they are driven by local information.

According to Fu and Mui (1981), some of the motivating factors of this method are:

- Most of the information of the image lies in the boundary between different regions.
- A biological visual system appears to make use of edge detection but not of thresholding.
- For blackboard segmentation (Onishi et al., 2000), uses edge detection from spatio temporal images to identify two types of edges
    1. Dynamic edges (edges with moving objects)
    2. Static edges (edges with stationary objects).

Spatio temporal images are considered because the instructor is moving continuously. Therefore writing region is detected by the static edges.

For video segmentation, Ram and Chaudhuri (2009) applied Sobel edge detection in each two consecutive frames followed by binarization and dilation. Then CC analysis is applied to differentiate the key frames from all kind of redundancies. In this way, text and figures are segmented from a video frame.

The segmentation algorithm of Baidya and Goel (2014), is of two steps in first step the entire video is analyzed and an edge map of two frames is created and CC analysis is performed on this map and second step aims in finding the slide transition and then consider only those frames which contain non redundant values. Davila and Zanibbi (2017) used canny edge detection technique to extract whiteboard region from the rest of the frame.

Reconstruction of binary images and background removal procedure is used to increase the precision of content extracted from lecture videos after binarization. To segment the content area, the prior knowledge that the slide area is almost quadrilateral in shape consisting of two horizontal and vertical edges is considered in Liu et al. (2002). Hence, first four edges of the content area are identified using color similarity weighted least square method. Liu and Choudary (2006), proposed a framework for real time streaming of instructional videos. The technique used for content analysis is built on edge detection. However, the limitation of edge based segmentation is that it does not work for videos which have so many edges.

Some researchers have used combination of features to segment video frame. In Wang et al. (2007) a combination of features like gesture, posture and text is used to segment the video frames. They have used frame differencing as well as skin color detection for identifying gesture and posture.

Some researchers used region of interest (ROI) to segment video frame. Yadid and Yahav, (2016), identified the integrated development environment (IDE) which is used for writing the program. Identification is based on the segmentation of image into various containers where, the smallest container which covers the maximum program code is searched out.

Jeong et al. (2012), identified slide region in lecture videos is used to identify slide region and separate it from the rest of the frame, background frame and teacher. Scale invariant feature transform (SIFT) is used for slide to slide feature comparison. Let $s_i$and $s_{i+k}$ be the slide region in frame $f_i$ and $f_{i+1}$ respectively. Where $f_i$ frame is captured at $i^{th}$ time and $f_{i+1}$frame is captured at $i+1^{th}$ time. The matched shift features is denotedby $P(s_i, s_{i+1})$.

Apart from this, some researchers have used clustering technique. In Yang et al. (2011), K – means segmentation technique is used for segmenting teaching board area. K-means algorithm is an unsupervised classification. Given a set of classification $\{(c_1, c_2, \ldots\ldots c_n\}$. K means algorithm aims at partitioning n observations into k sets $(k < n) S = \{S_1, S_2, \ldots S_k\}$so as to reduce the within cluster sum of square

$$argmin \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\| \quad (5)$$

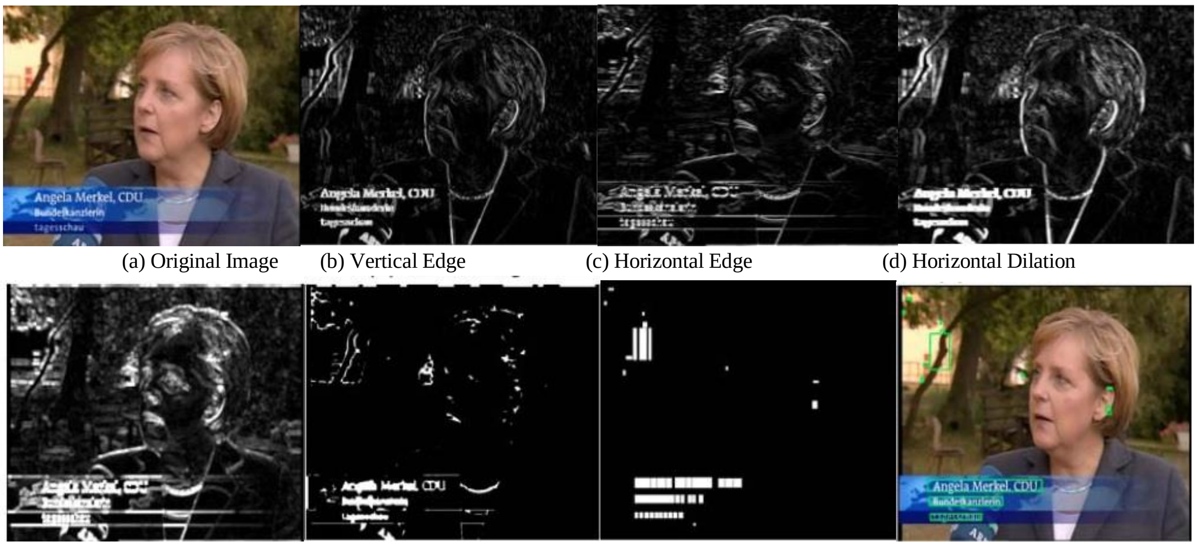
The symbol $\mu_i$ is the mean point in $S_i$. The mean value of $\mu_i$ of the maximum component of k-means segmented results is adopted to be the background color B as the following e.q.

$$B = \mu_i : i = argmax \sum_{x_j \in S_m} 1 \quad (6)$$

Further connected component technique helps to refill the board area covered by the lecturer.

### 4.3 Text Based Segmentation

Extracting text in e-learning or instructional videos is challenging and is different to the text detection in other domain. Video text detection which depends on Optical Character Recognition (OCR) (Yang et al., 2011; Baidya and Goel, 2014; Tuna et al., 2015), which have fixed fonts and clear text lines such as in enclosed video and result of text based segmentation is shown in Figure 4. However, they ail in determining the hand written text which varies in size. These OCR based techniques work well with neat and clean text but their performance is affected if the blackboard is not clean and text is handwritten and even not readable.


(a) Original Image (b) Vertical Edge (c) Horizontal Edge (d) Horizontal Dilation

(e) Vertical Dilation (f) Intersected Map (g) Projection Map (h) Detection results

**Figure 4. Results of text segmentation techniques discussed in (Yang et al., 2011)**

A handwritten text detection method is proposed (Tang and Kender, 2005). They use stroke for as primary method for image segmentation and trained multi-level perceptron neural network for character recognition.

Banerjee et al. (2014) used the idea in mind the prior knowledge of text region and then rather than applying any other feature they applied Scale Invariant Feature Transform (SIFT) as discussed by Lowe (2004), densely over the entire region then text is extracted.

Lin et al. (2004), segments instructional videos based on text based segmentation on the basis of transcribed text. Text based segmentation in instructional videos is still a challenge.

## 5. Motion Based Segmentation

In a lecture the teacher is often moving around in front of the blackboard to write, explain and emphasize something. In instructional videos instructor is non visual content so to remove these, two methods are mainly used to track foreground objects. One is to model the background as discussed by Javed et al. (2002), and subtract this background model from a frame, and the second one is to use motion estimation as proposed by Ekinci and Gedikli (2003).

To segment text area in video, Imran et al. (2012), segments the background i.e. board from the foreground. Text is analyzed from the background and key frames are extracted. Key frames are further subjected to text localization, extraction, enhancement and segregation.

### 5.1 Background Subtraction

Background subtraction is widely used approach for detecting moving objects. According to Dickson et al. (2008), for  background subtraction pixel wise difference is calculated between the current frame and the reference (background or background model) frame, its working is shown in Figure 5.

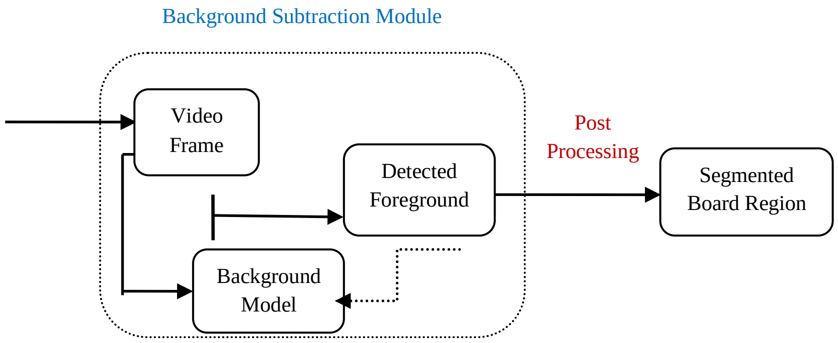


**Figure 5. Background subtraction process for e-learning application**

The reference frame is updated at a fixed interval of time so as to adapt with varying lightning conditions. The background subtraction is expressed with the following eq. (7)

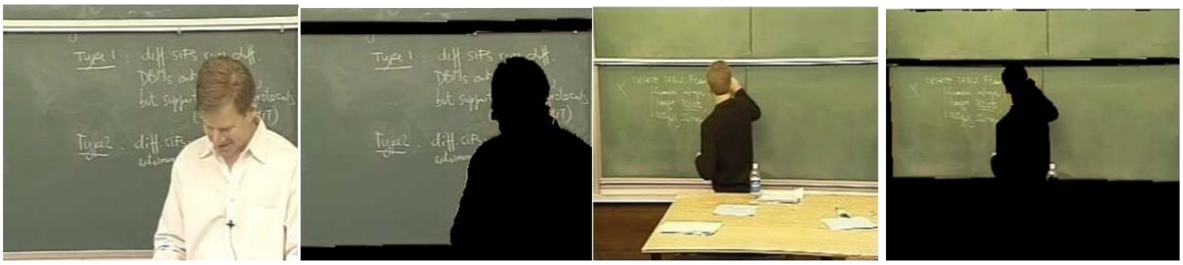
$$\beta_k(i,j) = \sum_{i,j=0}^{i,j=n} C_k(i,j) - R_k(i,j) \qquad (7)$$

after applying some image processing technique final frame containing only the board region is where$\beta$ is the background subtraction, $C_k$is the current frame $R_k$ is the reference frame, (i,

j) is the pixel co-ordinates and size of video frame is n X n and $R_k$ is updated regularly. The result of segmentation completely depends on the type of model formed for background the parameters considered for the model etc. Once foreground object is detected it consists of teacher and the change in the board content, therefore obtained.

In Dickson et al. (2006) a background subtraction model called Gaussian Mixture Model is constructed for upper region of the black board. A background subtraction with the current frame and the model is computed and binary optimal method is used to find the handwritten data.

The result of background separation is shown in Figure 6 as discussed in Choudary and Liu, (2007) after identifying the color of the background.

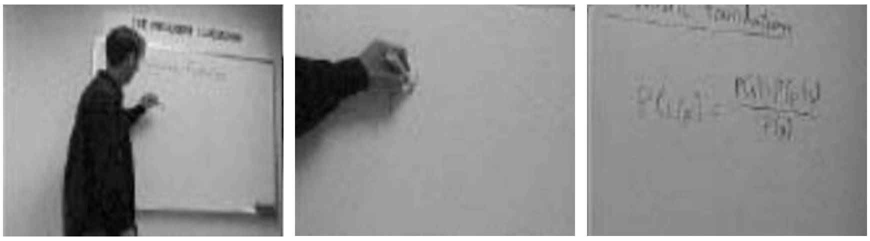


(a) Orginal frame (b) Subtracted frame (c) Original frame (d) Subtracted frame

**Figure 6. Result of background subtraction (Choudary and Liu, 2007)**

A highlight of video is created in Subudhi et al. (2017), by segmenting and recognizing the activities in instructional videos using Hidden Markov Model (HMM) the activities are classified into talking head, writing hand and slide show.

Usually when the e-learning videos are compressed for transmission the quality of the video decreases thus in Franklin and Hammond (2001) content is enhanced by first segmenting the chalk board region and then a background model is formed for background and foreground separation and combination technique This separation normalizes and denoises the foreground i.e. the instructor and thus enhances the readability of the chalk-board content as shown in Figure 7.



(a) Instructor Writing on board (b) System zooms to the writing region (c) Text area

**Figure 7. Text detection result of Franklin et al. (2001) approach**

### 5.2 Frame Differencing

It is the pixel wise difference between the current frame and reference frame.

$$S(i,j) = I_{t1}(i,j) - I_{t2}(i,j) > threshold \tag{8}$$

where $S$ is the subtracted frame and $I_{t1}$ and $I_{t2}$is the frame captured at the time t1 and t2 respectively.

Using motion estimation is not that time consuming compared to background modeling which often need updating and statistical calculations? In addition, motion estimation is not dependent on a clear background for initialization. Motion from the teacher could be enough to detect and afterwards remove the teacher. For instance, a generic solution to the segmentation problem has not been addressed earlier. The segmentation techniques would work for green blackboards, but not for whiteboards. Some of the segmentation algorithms used in other contexts is possible to use to segment the teacher in lecture videos. But the main difference between lecture video segmentation and for instance video surveillance is that there is mostly one person to segment, the person is close to camera and there is valuable background information. Also many segmentation algorithms use background subtraction to remove the background and keep the foreground. In lecture video it's opposite; the teacher has to be removed so that the content on the blackboard is clear and visible.

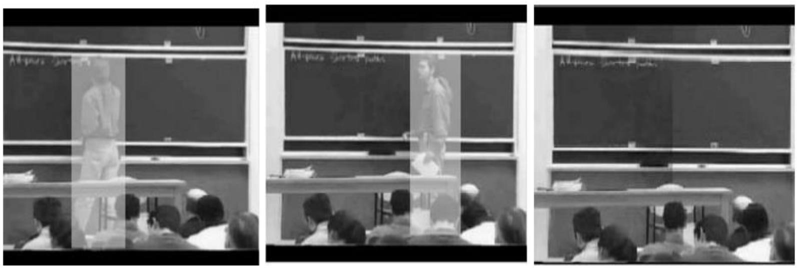
In Ma and Agam (2012), video is segmented into various scenes by identifying the transition of frames by the analysis of color histogram of videos frames and each shot is referred as shot. Frame differencing is calculated by Eq. (1) and then color histogram is computed using. The difference between two frames is identified as

$$Diff = \sum_{i,j}\left|\sum_{a,b} I_{i,j}^{x}(a,b) - \sum_{a,b} I_{i,j}^{x+1}(a,b)\right| \tag{9}$$

$$Hist_{x,y}(a,b) = \begin{cases} 0 & if\ 32 * j \le Int(a,b) \le 32 * j + 1 \\ 1 & otherwise \end{cases} \tag{10}$$

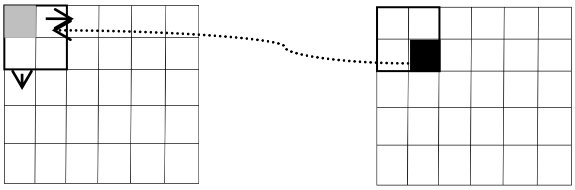
where $i \in \{RGB\}, j \in \{0,7\}, a \in [0, width), b \in [0. height)$, i indicates the color histogram and j indicates the bins of the histogram and Diff is the difference between the frame *x* and *x+1.Int(a,b)* indicates the intensity at each pixel.

In Mittal et al. (2006), educational videos are compressed by first segmenting video frames into various components containing teacher, board, background. Blackboard segmentation is achieved by canny edge detection. Horizontal and vertical edges are identified. Teacher is segmented by firstly modeling a background as shown in Figure 8, teacher is tracked into subsequent frames, further frames are divided into n*n blocks. Region covered by the teacher is masked and dilated, these processing makes it computationally expensive.



(a) One frame for background model (b) Another frame for background model (c) Another frame for background model

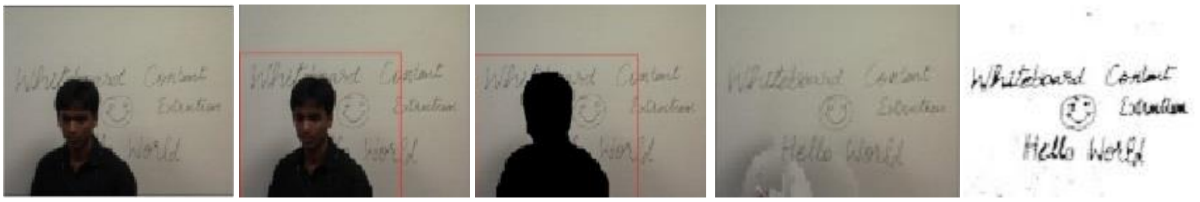
**Figure 8. Results of background modeling of Mittal et al. (2006) approach**



(a) Reference frame (b) Current frame

**Figure 9. Motion detection**

Another technique for frame differencing used is block based approach as proposed in Liu and Choudary (2006); Prabhu et al. (2008) detects content region by dividing both the frames (current frame and reference frame) into n X n blocks and is shown in Figure 9. Blocks are categorized into content and non-content block when the edge density is higher than the threshold (Liu and Choudary, 2006). Result of Prabhu et al. (2008) text detection is shown in Figure 10. It finds or removes foreground by finding a reference frames, occluded object is identified by dividing the two frames i.e. reference frame and current frame into 16 X 16 blocks. The object is identified by calculating sum of absolute difference (SAD) for all the blocks and blocks are further clustered together to find the foreground object.



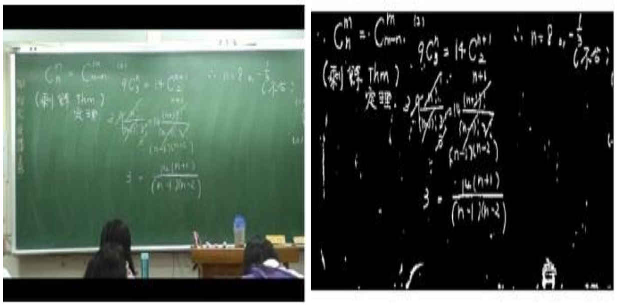
(a) Original frame (b) Moving object detected (c) Segmented Foreground (d) Pixels of foreground are replaced by pixels of reference frame (e) Final text area

**Figure 10. Content region detection result of Prabhu et al. (2008) approach**

In Yang et al. (2011), frame differencing is used to compare two frames and find out new segment. They have considered text, figures as a group of collected components (*CCs*).

Therefore, they create differential edge map of two frames and perform CC analysis on the map. To ensure that the new segment is detected a threshold (*th*) is set.

In Lee et al. (2017), temporal differencing is used to detect the ROI from the high resolution images. Since the temporal differencing cannot extract all the pixels of foreground object a graph cut technique is used for further processing. Once foreground object is detected it is segmented from the frame, then the chalkboard image is generated which does not contain instructor again to detect the change in the chalk board area temporal difference is computed finally a time shrunk video is generated. Results of text segmentation is shown in Figure 11.



(a) (b)

(b)

**Figure 11. Text segmentation results of Lee et al. (2017) approach**

In Yokoi and Fujiyoshi (2006), whiteboard content is extracted by first converting the captured image into average image and then a refined image is created by remapping color to the input image. Pixel difference is calculated to identify the lecturer for this only the corresponding pixels where the sum of color differences within the color values of all 3 color channels exceed an empirically determined threshold of 80 are considered different.

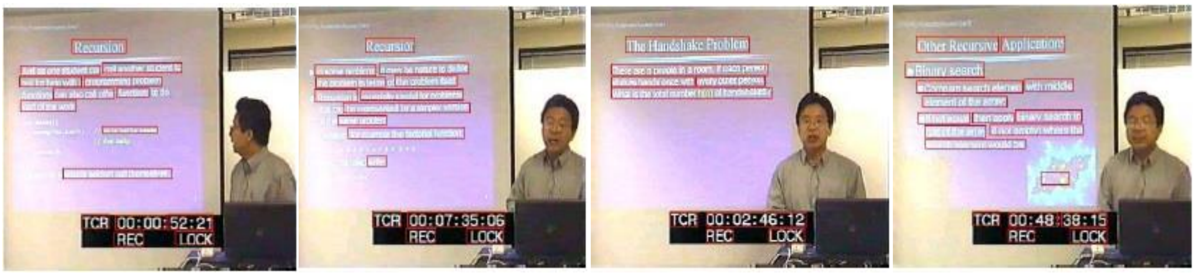


**Figure 12. Text detection results of Ngo et al. (2003) approach**

A system for analyzing and annotations of technical talks is proposed in Ngo et al. (2003). A motion estimation technique is used to identify key frames in the videos. The potential gestures like pointing or writing are tracked by active contours and are used for annotation purpose. Results of text detection in various slides are shown in Figure 12.

A lot of research is being carried out in segmentation of instructional videos. The summary of all the techniques used so far for segmentation is shown in Table 1.

**Table 1. Various segmentation techniques**

| Segmentation Technique | Method Description | Advantage | Disadvantage |
|---|---|---|---|
| Color Feature Extraction<br>(Cheng et al., 2001),<br>(Ju et al., 1998), | First extracts the color of the board , then segments the board region | Computationally inexpensive | Does not work well for classroom scenario where the teacher is wearing the same color outfit as that of the chalk board.<br>The techniques discussed so far are able to segment a particular type of color i.e.white, green or black,but fail to extract other type of boards. |
| Text Based Segmentation<br>(Baidya and Goel, 2014),<br>(Tuna et al., 2015) | A model is trained by various alphabets in lower and upper case. Text is segmented with the help of the text segmentation model, and then text region is identified. | Text is directly subtracted which contains only the visual data | Depends on how efficient the model is and how it is trained.<br>Text also contains symbols and equations which are often not recognized as text by the proposed model. |
| Shape Based Segmentation<br>(Pal and Pal, 1993)<br>(Fu and Mui, 1981) | Rectangular shape is determined to segment the board area, for this edge detection technique is used. | Efficient for determining the shape of the objects in the video frame | Does not work for videos which have so many edges. |
| Background / Foreground Subtraction<br>(Dickson et al., 2008)<br>(Subudhi et al., 2017) | A background or model is created and then pixel wise difference is calculated between the background and the current frame. Once foreground (tutor) is obtained, it is segmented from the frame and after applying image processing technique final frame which contains the text in the chalk-board is obtained. | Efficient for segmentation of videos. | Efficiency greatly depends on how a model is formed and it has to be updated frequently to adapt with the various conditions. |
| Temporal (frame) Differencing<br>(Ma and Agam, 2012)<br>(Lee et al., 2017)<br>(Yokoi and Fujiyoshi, 2006) | Pixel wise difference is computed between two frames captured in two different intervals is computed. | Easy, way to segment dynamic region of a frame based on the motion of that region. | In case of e-learning videos the tutor as well as the visual contents like(text and figures ) are also changing , therefore frame differencing is not self sufficient to segment the board region , some other image processing technique need to be used along with it. |

## 6. Conclusions

In this paper, most of the methods proposed for segmentation of instructional videos are reviewed. In general no segmentation technique is efficient enough to segment the board region from rest of the frame. However remarkable results are achieved after applying the techniques in combination. We have presented the summarized version in Table 1, with the advantages and disadvantages of all the techniques used for segmentation so far. On the basis of this review some new or combination of these techniques can help in obtaining improvised results for segmentation of instructional video. Segmentation techniques based on color feature extraction works well for a particular color chalk board for which it is designed for but, it fails to extract other color board, i.e. if it is designed for segmenting white chalk-board it fails to segment green or black chalk-board. Thus color feature extraction based segmentation technique should be robust to segment any color chalk-board. The other limitation is that if an instructor/tutor is wearing the same color outfit as that of the chalk-board then this technique fails in segmenting board region.

Edge based segmentation techniques are used for extracting shape feature of the board region. This technique is not self-sufficient enough to segment the board area i.e. some other image processing technique has to be done along with it to achieve good results.

Text based segmentation technique perform well and provide good results as the main aim in segmentation of instructional videos is to remove irrelevant content from the video frames thus extracting only the text from the frame and leaving all the regions from the frame helps in achieving the objective. The limitation of text based segmentation is that they require a learning model which is to be trained. The text which is in a different language or may be sometimes the handwriting is not very clear then the model fails to segment. Moreover, a text also contains symbols and equations also which are often unrecognized by these techniques.

In foreground / background subtraction technique again background has to be updated in such a way that it provides good result; as well as if the illumination or lighting conditions are poor in a classroom then also the performance of the technique is affected. Temporal frame differencing techniques are efficient but they are computationally very expensive and moreover as the tutor and the visual content in the video frames keeps on changing therefore motion based technique are not self-sufficient enough to segment board region or visual content from the frame. Thus based on the above literature review we came to a conclusion that for segmentation of visual content from the non-visual content in an instructional (e-learning) video using only one technique is not sufficient but, applying other image processing technique along with them will help in achieving segmentation.

## References

Amir, A., Ashour, G., & Srinivasan, S. (2001, January). Towards automatic real time preparation of on-line video proceedings for conference talks and presentations. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences (pp. 8-pp). IEEE.

Asghar, M. N., Hussain, F., & Manton, R. (2014). Video indexing: a survey. International Journal of Computer and Information Technology, 3(01), 148-169.

Baidya, E., & Goel, S. (2014, August). LectureKhoj: automatic tagging and semantic segmentation of online lecture videos. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 37-43). IEEE.

Banerjee, P., Bhattacharya, U., & Chaudhuri, B. B. (2014, September). Automatic detection of handwritten texts from video frames of lectures. In 2014 14th International Conference on Frontiers in Handwriting Recognition (pp. 627-632). IEEE.

Bianchi, M. (1998, July). Auto auditorium: a fully automatic, multi-camera system to televise auditorium presentations. In Proc. of Joint DARPA/NIST Smart Spaces Technology Workshop.

Brejl, M., & Sonka, M. (2000). Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples. IEEE Transactions on Medical imaging, 19(10), 973-985.

Chen, W. T., Liu, W. C., & Chen, M. S. (2010). Adaptive color feature extraction based on image color distributions. IEEE Transactions on Image Processing, 19(8), 2005-2016,

Cheng, H. D., Jiang, X. H., Sun, Y., & Wang, J. (2001). Color image segmentation: advances and prospects. Pattern Recognition, 34(12), 2259-2281.

Choudary, C., & Liu, T. (2007). Extracting content from instructional videos by statistical modelling and classification. Pattern Analysis and Applications, 10(2), 69-81.

Davila, K., & Zanibbi, R. (2017, November). Whiteboard video summarization via spatio-temporal conflict minimization. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 355-362). IEEE.

Dickson, P. E., Adrion, W. R., & Hanson, A. R. (2008, December). Whiteboard content extraction and analysis for the classroom environment. In 2008 Tenth IEEE International Symposium on Multimedia (pp. 702-707). IEEE.

Dickson, P., Adrion, W. R., & Hanson, A. (2006, December). Automatic capture of significant points in a computer based presentation. In Eighth IEEE International Symposium on Multimedia (ISM'06) (pp. 921-926). IEEE.

Dong, A., & Li, H. (2005, December). Educational documentary video segmentation and access through combination of visual, audio and text understanding. In Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005. (pp. 652-657). IEEE.

Dorai, C., Oria, V., & Neelavalli, V. (2003, September). Structuralizing educational videos based on presentation content. In Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429) (Vol. 2, pp. II-1029). IEEE.

Ekinci, M., & Gedikli, E. (2003, November). Background estimation based people detection and tracking for video surveillance. In International Symposium on Computer and Information Sciences (pp. 421-429). Springer, Berlin, Heidelberg.

Franklin, D., & Hammond, K. (2001, May). The intelligent classroom: providing competent assistance. In Proceedings of the Fifth International Conference on Autonomous Agents (pp. 161-168). ACM.

Fu, K. S., & Mui, J. K. (1981). A survey on image segmentation. Pattern Recognition, 13(1), 3-16.

Haubold, A., & Kender, J. R. (2005, November). Augmented segmentation and visualization for presentation videos. In Proceedings of the 13th Annual ACM International Conference on Multimedia (pp. 51-60). ACM.

He, L. W., & Zhang, Z. (2006). Real-time whiteboard capture and processing using a video camera for remote collaboration. IEEE Transactions on Multimedia, 9(1), 198-206.

Imran, A. S., Chanda, S., Cheikh, F. A., Franke, K., & Pal, U. (2012, November). Cursive handwritten segmentation and recognition for instructional videos. In 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems (pp. 155-160). IEEE.

Javed, O., Shafique, K., & Shah, M. (2002, December). A hierarchical approach to robust background subtraction using color and gradient information. In Workshop on Motion and Video Computing, 2002. Proceedings. (pp. 22-27). IEEE.

Jeong, H. J., Kim, T. E., & Kim, M. H. (2012, December). An accurate lecture video segmentation method by using sift and adaptive threshold. In Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia (pp. 285-288). ACM.

Ju, S. X., Black, M. J., Minneman, S., & Kimber, D. (1998). Summarization of videotaped presentations: automatic analysis of motion and gesture. IEEE Transactions on Circuits and Systems for Video Technology, 8(5), 686-696.

Lee, G. C., Yeh, F. H., Chen, Y. J., & Chang, T. K. (2017). Robust handwriting extraction and lecture video summarization. Multimedia Tools and Applications, 76(5), 7067-7085.

Li, H., & Dong, A. (2006, August). Hierarchical segmentation of presentation videos through visual and text analysis. In 2006 IEEE International Symposium on Signal Processing and Information Technology (pp. 314-319). IEEE.

Lin, M., Nunamaker, J. F., Chau, M., & Chen, H. (2004, January). Segmentation of lecture videos based on text: a method combining multiple linguistic features. In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the (pp. 9-pp). IEEE.

Lin, Y. T., Tsai, H. Y., Chang, C. H., & Lee, G. C. (2010, September). Learning-focused structuring for blackboard lecture videos. In 2010 IEEE Fourth International Conference on Semantic Computing (pp. 149-155). IEEE.

Liu, Q., Rui, Y., Gupta, A., & Cadiz, J. J. (2001, March). Automating camera management for lecture room environments. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 442-449). ACM.

Liu, T., & Choudary, C. (2006). Content-adaptive wireless streaming of instructional videos. Multimedia Tools and Applications, 28(2), 157-171.

Liu, T., & Kender, J. R. (2002). Rule-based semantic summarization of instructional videos. In Proceedings. International Conference on Image Processing (Vol. 1, pp. I-I). IEEE.

Liu, T., Hjelsvold, R., & Kender, J. R. (2002). Analysis and enhancement of videos of electronic slide presentations. In Proceedings. IEEE International Conference on Multimedia and Expo (Vol. 1, pp. 77-80). IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant key points. International Journal of Computer Vision, 60(2), 91-110.

Ma, D., & Agam, G. (2012, January). Lecture video segmentation and indexing. In Document Recognition and Retrieval XIX (Vol. 8297, p. 82970V). International Society for Optics and Photonics.

Masneri, S., & Schreer, O. (2014, January). SVM-based video segmentation and annotation of lectures and conferences. In 2014 International Conference on Computer Vision Theory and Applications (VISAPP) (Vol. 2, pp. 425-432). IEEE.

Mittal, A., Gupta, S., Jain, S., & Jain, A. (2006). Content-based adaptive compression of educational videos using phase correlation techniques. Multimedia Systems, 11(3), 249-259.

Mukhopadhyay, S., & Smith, B. (1999, October). Passive capture and structuring of lectures. In ACM Multimedia (1) (pp. 477-487).

Ngo, C. W., Wang, F., & Pong, T. C. (2003, December). Structuring lecture videos for distance learning applications. In Fifth International Symposium on Multimedia Software Engineering, 2003. Proceedings. (pp. 215-222). IEEE.

Onishi, M., Izumi, M., & Fukunaga, K. (2000). Blackboard segmentation using video image of lecture and its applications. In Proceedings 15th International Conference on Pattern Recognition. ICPR-2000 (Vol. 4, pp. 615-618). IEEE.

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. Pattern Recognition, 26(9), 1277-1294.

Prabhu, N., Kumar, R. P., Punitha, T., & Srinivasan, R. (2008, October). Whiteboard documentation through foreground object detection and stroke classification. In 2008 IEEE International Conference on Systems, Man and Cybernetics (pp. 336-340). IEEE.

Ram, A. R., & Chaudhuri, S. (2009, August). Automatic capsule preparation for lecture video. In 2009 International Workshop on Technology for Education (pp. 10-16). IEEE.

Subudhi, B. N., Veerakumar, T., Yadav, D., Suryavanshi, A. P., & Disha, S. N. (2017, January). Video skimming for lecture video sequences using histogram based low level features. In 2017 IEEE 7th International Advance Computing Conference (IACC) (pp. 684-689). IEEE.

Tang, L., & Kender, J. R. (2005, July). Semantic indexing for instructional video via combination of handwriting recognition and information retrieval. In 2005 IEEE International Conference on Multimedia and Expo (pp. 920-923). IEEE.

Tuna, T., Joshi, M., Varghese, V., Deshpande, R., Subhlok, J., & Verma, R. (2015, October). Topic based segmentation of classroom videos. In 2015 IEEE Frontiers in Education Conference (FIE) (pp. 1-9). IEEE.

Wallick, M. N., Heck, R. M., & Gleicher, M. L. (2005, March). Marker and chalkboard regions. In Proceedings of Mirage (pp. 223-228).

Wang, F., Ngo, C. W., & Pong, T. C. (2007). Lecture video enhancement and editing by integrating posture, gesture, and text. IEEE Transactions on Multimedia, 9(2), 397-409.

Yadid, S., & Yahav, E. (2016, October). Extracting code from programming tutorial videos. In Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (pp. 98-111). ACM.

Yang, H., Siebert, M., Luhne, P., Sack, H., & Meinel, C. (2011, December). Automatic lecture video indexing using video OCR technology. In 2011 IEEE International Symposium on Multimedia (pp. 111-116). IEEE.

Yang, H., Siebert, M., Luhne, P., Sack, H., & Meinel, C. (2011, November). Lecture video indexing and analysis using video ocr technology. In 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems (pp. 54-61). IEEE.

Yokoi, T., & Fujiyoshi, H. (2006, July). Generating a time shrunk lecture video by event detection. In 2006 IEEE International Conference on Multimedia and Expo (pp. 641-644). IEEE.