Towards partial fulfilment for Undergraduate Degree Level Programme Bachelor of Technology in Computer Engineering

A Project Report on:

Keyword Indexing and Retrieval from Video

Prepared by :

Admission No.		Student Name
U16CO061	-	Paras Purwar
U16CO062	-	Mayank Kumar
U16CO076	-	Manish Kumar Tanti
U16CO078	-	Vinod Kumar Sonkar

Class : B.TECH. IV (Computer Engineering) 8th Semester

Year : 2019-2020

Guided by **:** Dr. Krupa N. Jariwala

DEPARTMENT OF COMPUTER ENGINEERING SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY, SURAT - 395 007 (GUJARAT, INDIA)

Student Declaration

This is to certify that the work described in this project report has been actually carried out and implemented by our project team consisting of-

Sr.	Admission No.	Student Name
1	U16CO061	Paras Purwar
2	U16CO062	Mayank Kumar
3	U16CO076	Manish Kumar Tanti
4	U16CO078	Vinod Kumar Sonkar

Neither the source code there in, nor the content of the project report have been copied or downloaded from any other source. We understand that our result grades would be revoked if later it is found to be so.

Signature of the Students:

Sr.	Student Name	Signature of the Student
1	Paras Purwar	
2	Mayank Kumar	
3	Manish Kumar Tanti	
4	Vinod Kumar Sonkar	

Certificate

This is to certify that the third stage project report entitled **Keyword Indexing and Retrieval from Video** is prepared and presented by-

Sr.	Admission No.	Student Name
1	U16CO061	Paras Purwar
2	U16CO062	Mayank Kumar
3	U16CO076	Manish Kumar Tanti
4	U16CO078	Vinod Kumar Sonkar

Final Year of Computer Engineering and their work is Satisfactory.

SIGNATURE:

GUIDE JURY HEAD OF DEPARTMENT

Abstract

This project aims to implement "**Keyword Indexing and Retrieval from Video**".Text Extraction of information from videos is an important research field of information indexing and retrieval because this technology is emerging.Automatic annotation required important text data which is required for indexing from the videos.In indexed frames there are different types of text such as scene text and caption text.Now a days different types of videos are there such as e-learning videos, news video and TV program videos etc.All of these videos are created by different types of image frames based on its different purposes.Reason behind variation in text because of difference in font size,font style,font color,font orientation, complexity and background of frames,brightness and contrast of background.Extracting,evaluating, analyzing and indexing frames from videos is a process which is called Keyword Extraction and Indexing.In this report we are discussing some of the techniques of key frame selection, text extraction, keyword indexing and comparison between them.It suggests that there is a path to extract video features through text extraction from indexed frames and indexed content which is present on frames.Now a days Online and Digital Videos are used widely professionally and domestically because of the easy availability of internet and palm devices such as mobile phones.Now indexed frames and Online videos are used for professional and personal use such as online study material,TV series.While we make,transmit,store,edit these videos text in these videos are not changeable.

Keywords - *Key Frame, Frame Selection, Video Indexing, Keyword Selection, Indexing, Content Retrieval, Text Extraction, Detection, Binarization, edge, connected component, Frame Extraction,Text Recognition, Keyword Indexing.*

Contents

List of Figures

List of Tables

List of Abbreviations

CC	Connected Component
DR	Detection Rate
ECR	Edge Change Ratio
FAR	False Alarm Rate
FD	Frame Difference
OCR	Optimal Character Recognition
PDE	Partial Differential Equation
PR	Precision Rate
RR	Recall Rate
SSD	Sum of Squared Difference
SVM	Support Vector Machine
TIE	Text Information Extraction

Chapter 1

Introduction

In videos there are different types of text objects.These objects contain information about videos such as logo of a university which tells university name and various texts which provide the contents about the video.That's why extraction of text is important for video indexing and information retrieval.In this report we have done the exactly the same thing and returned the text present in the indexed video in the order of their appearance.

1.1 Application

In recent years the availability of videos are growing rapidly over internet specially on Youtube.The text extraction is used for searching important information from video data sets.Using this extracted text anybody can get an idea about the videos.For categorizing the extracted text play important role as a key sign.It is also used to determine the content of the video.Video text extraction is identified as one of the key components of the video analysis and retrieval system.Video text extraction can be used in many applications,like multilingual video information access,semantic video indexing, video security and surveillance etc.In every video which contain text usually persists for at least some seconds,because of human viewers so that they read it and understand easily.This temporal nature of video is very valuable and can be well utilized for text extraction in videos.There are two types of text in video are that is-caption text and scene text.At the time of editing video caption text artificially superimposed.During capturing of image and video field of the view of the camera that is scene text.Text extraction from scene text is more complex than caption text because of different types of lighting and complexity.Text present in video are important because of following reasons:

1. To describing the contents of video.

2. It can be easily extracted from videos compared to other semantic contents.

3. It enables different types of applications such as automatic video logging, and text-based image indexing.

Texts in videos are usually different because of appearance of that due to changes in font size, color,style,orientation,alignment,texture,colour,contrast and background.Some of the characteristics are as follows:

1. Size: In the video archives there are different types of font.

2. Alignment: Alignment of text in any direction and also may contain geometric distortion.

3. Colour: Color of text which is present in video is different.

4. Edge: Contents in the videos have boundary edges which is sometimes strong or weak.

5. Compression: When any videos is recorded or transferred sometimes it compressed automatically.

1.2 Motivation

There are different types of methods to extract the text from videos.These methods are for specific applications including page segmentation,license plate location and content-based video indexing. After studying such types of text extraction method it is not easy task to design a general text information extraction (TIE) system.In videos there are different types of variations such as complexity of background,font size,color,style,alignment,brightness that's why design of a TIE system is tough.These variations play a important role to not working properly a automatic TIE system.After studying different methods of text extraction,analyzing their evaluation results,performance evaluation approaches not only search for answers to many questions such as: Which text extraction method is better? Why does performance of different methods is varying in different types of dataset ? Which types of error comes at the time of indexing ? These questions actually help to develop new ideas to improve the extraction technology and specific algorithms.

1.2.1 Stages of Text Extraction

There are different stages of text extraction from videos which are given below-

1. **Text detection-** In a video frame finding that regions which contain text.
2. **Text localization-** Combine different text regions into text instances and generating a set of tight boundary areas around all text instances.
3. **Text tracking-** Continue to follow a text event as it moves or changes continuously or not over time and determining the different(temporal and spatial) locations.
4. **Text recognition-** Performing OCR on the indexed text frame.Occasionally recognition step is deleted in favour of applying OCR on colour/grey level images.

For extraction of text different techniques are used by many researchers and which can be classified later.According to different programs and title of that program text is abundant in videos.Generally caption text is used in news videos and many times in sports and videos game and player statistics or we can say player position are often superimposed on the frames in textual form.Sometimes commercials want that information of the product is presented as readable text.When any person extract the video text it is automatically extracted,it not only provides keywords for annotation and search of text and video libraries but also aids in highlighting events which can then be used for summarizing a video.For video categorization,cataloguing of commercials,logging of key events and efficient video digest construction text extraction is used.

Fig. 1.1: **Text in videos appears in different contexts, backgrounds, and font sizes [5]**

While different types of content extraction techniques are reasonably developed for text,data which is presented in video still is essentially opaque.We don't use video generally as a communication medium, due to unavailability of processing because high speed processing of video

and frame generation method and communication supporting platforms has delayed its introduction in a generalized way.Now GPUs are used for that this issue is continuously changing and new video-based applications are being developed.

1.3 Objectives

In this project, methods of how to extract proper text from videos are discussed and also which types of tools are used which method gives how much accuracy shown we are currently developing tools for indexing video archives for later reuse,a system for content analysis of videos in which text appearance is different.These all things are also dependent on their efficient computational support,combining indexed image and video analysis and processing tools. Now a days in text extraction rapid developments are shown hundreds of researcher try to do this in proper way and any research paper is published.Text extraction approaches for videos proposed respectively.In this project,we mainly concentrate on the approaches proposed for text extraction in videos in the most recent 5 years and how to get proper text from videos.To summarize and discuss the recent progress in this research area.

1.4 Contribution

As from our side we have tried our best to create a system which extracts text from videos then after it retrieves relevant information from the extracted text.The project will be completed in three phases-

1. Operation done on video.
2. Text Extraction from Videos.
3. Use Relevant information from Extracted Text.

To document the progress of the system we have created a detailed report and concise presentation.

1.5 Organisation of Project Report

Chapter 1 of the report provides a brief introduction, the reason behind choosing the project, the motivation behind the work the application of this system in real world.

Chapter 2 contains Literary Survey and Theory related to project during the course of project completion.In this chapter we have also explained the various techniques which is used in key frame selection, video content indexing and text extraction and tabulated the comparison of performance among the techniques.In chapter 3 proposed model and methodologies are discussed employed for our model. In chapter 4 simulation of our model and different results are shown, accuracy of model for different data and comparison between them. In chapter 5 conclusion of our project and what we will do further in this are described.

Chapter 2

Literature Survey and Theory

Relevant Information from frames of indexed video is something which has become a new phenomenon upon which many research papers are being published and still the searching continues to go on.Although it's tedious and complex subject but due to its tremendous use it's a hot potato for many years.The research papers which has been published regarding the same is thoroughly analysed and referred for further understanding.The techniques which are mentioned the papers are explained in subsequent parts of the project research.As we move in ahead we discuss different phases of project.

2.1 Methods for Key Frame Selection

2.1.1 Key Frame for Video Copyright Protection

There are some distinct features about the key frame for video copyright protection.So,the key frame for video copyright protection is defined firstly before video pre-processing and key frame extracting. The key frames should meet the following three conditions-

1. The key frame is within a certain range to allow viewers to have subjective perception about the video content.Images with low gray value in Fig.2.1 are extracted from a single video, which is difficult for almost viewers to recognise the content.

2. The final key frame sequence must be arranged in chronological order consistent with original video sequence,in order to satisfy text extraction features and to be different from the short promotion trailer.

3. Appropriate redundancy of some key frames is allowed to ensure the periods or intervals along the processing of video content.Many Images in a video, which are with similar content, that is to say, one judge in the show every once in a while.

Fig. 2.1: Four frames with low gray value extracted from a video. [14]

2.1.2 Two-Stage Method for Key Frame Extraction

In a key frame extraction for digital video copyright protection. First, a digital video is decomposed into video frames.The downloaded video from the network includes several video formats, such as f4v, flv and mp4.In order to improve the universality of video key extraction algorithm,the present method does not consider the specific format and video stream structure, and the video is decoded before the processed video frame decomposition.It is seen that the program to extract key frame is divided into two steps.Firstly, alternative key frame sequence based on the color characteristics of the original difference between video frames is obtained; then key frame sequence is got according to the structure characteristic differences between alternative key frames sequence and finally it is determined by the number of key frames in order to ensure the effectiveness of key frames.

2.1.3 Performance Analysis of Key Frame Extraction Methods

A key frame extraction method based on frame difference with low level features is proposed for video copyright protection.Exactly, a two-stage method is used to extract accurate key frames to cover the content for the whole video sequence.Firstly,an alternative sequence is obtained based on color characteristic difference between adjacent frames from original sequence.

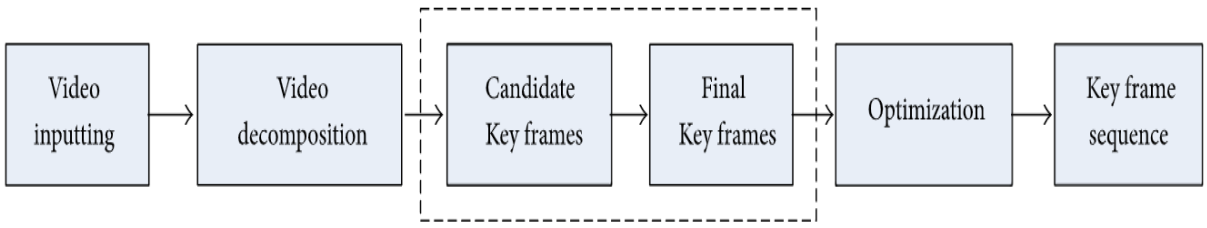


Fig. 2.2: Flowchart of the key frame extraction method [14]

Secondly,the final key frame sequence is obtained by analyzing structural characteristic difference between adjacent frames from the alternative sequence.Two stage method is used mostly because of frame difference value.This method calculate frame difference value is more accurate than video copyright method.

Table 2.1: Comparison of Key Frame Selection Methods[17]

Author	Method	Accuracy in %
Liu[17]	Video Copyright	72
Fan[17]	Two Stage	81

2.2 Methods for Video Indexing

Generally, videos are structured according to a descending hierarchy of video clips, scenes, shots, and frames. Video structure analysis aims at segmenting a video into a number of structural elements that have semantic contents, including shot boundary detection, key frame extraction, and scene segmentation.

- **Shot detection-** A shot is a consecutive sequence of frames captured by a camera action that takes place between start and stop operations, which mark the shot boundaries. There are strong content correlations between frames in a shot. Therefore, shots are considered to be the fundamental units to organize the contents of video sequences and the primitives for higher level semantic annotation and retrieval tasks.

- **Key Frame Extraction-** There are great redundancies among the frames in the same shot; therefore, certain frames that best reflect the shot contents are selected as key frames

to succinctly represent the shot. The extracted key frames should contain as much salient content of the shot as possible and avoid as much redundancy as possible.

- **Scene Segmentation-** Scene segmentation is also known as story unit segmentation. In general, a scene is a group of contiguous shots that are coherent with a certain subject or theme. Scenes have higher level semantics than shots. Scenes are identified or segmented out by grouping successive shots with similar content into a meaningful semantic unit. The grouping may be based on information from texts, images, or the audio track in the video.

Generally there are 2 types of video indexing method-

1. Audio Based Indexing
2. Content Based Indexing

2.2.1 Audio Based Indexing

In this type of indexing defining an architecture that follows the audio track and convert audio signal into text and index according to audio speed. Audio data query can be classified into two different approaches: a whole audio and video search and in object search. Each approach generates a different type of query result. A-whole-object search approach searches for data that are globally similar to the query input; on the other hand, an in-object search approach searches for a large piece of data that contains a fragment that are in similar to their query based on audio and video. An example of a-whole-sample search is to find a song in a database using the song as a query as well as frame query. An example of in-object search is to find a song that contains parts that are similar to the query, where the query is a melody for audio, and frame for video. There are many features that can be used to characterize audio and video signals. Usually audio and video features are extracted in two levels: frame level and sample level. For a feature to reveal the semantic meaning of an audio and video signal, analysis over a much longer period is necessary, usually from one second to several minutes. Such an interval is known as an audio and video sample. A sample consists of a sequence of frames, and sample-level features usually characterize how frame-level features change over a sample. The sample boundaries may be the result of audio segmentation and classification such that the frame features within each sample are similar. Alternatively, fixed length samples, usually 2

to 6 seconds, may be used. In this work, two hierarchical indexing systems are constructed as shown in Fig.

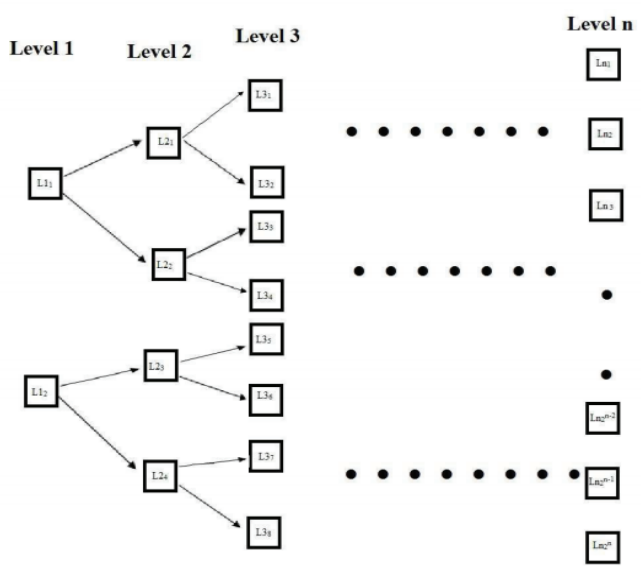


Fig. 2.3: Hierarchical Based Indexing [23]

2.2.2 Content Based Indexing

In this type of indexing defining an architecture that characterizes the tasks of managing video content which is presented on screen. In content Based Indexing there are 4 parameters-

- Content : talking head vs car crash
- Uniformity : smoothness as a function of time
- Panning : horizontal camera movement
- Tilting : vertical camera movement

2.2.3 Performance Analysis of Video Indexing Methods

Basically Content based Indexing applied on 2 algorithm which is discussed below-

Method-A: In this method differences between these DC frames are then computed. Results are presented for two metrics: the sum of the absolute DC frame pixel-to-pixel difference, and the bin to bin difference between histograms of the DC frame pixel luminance.

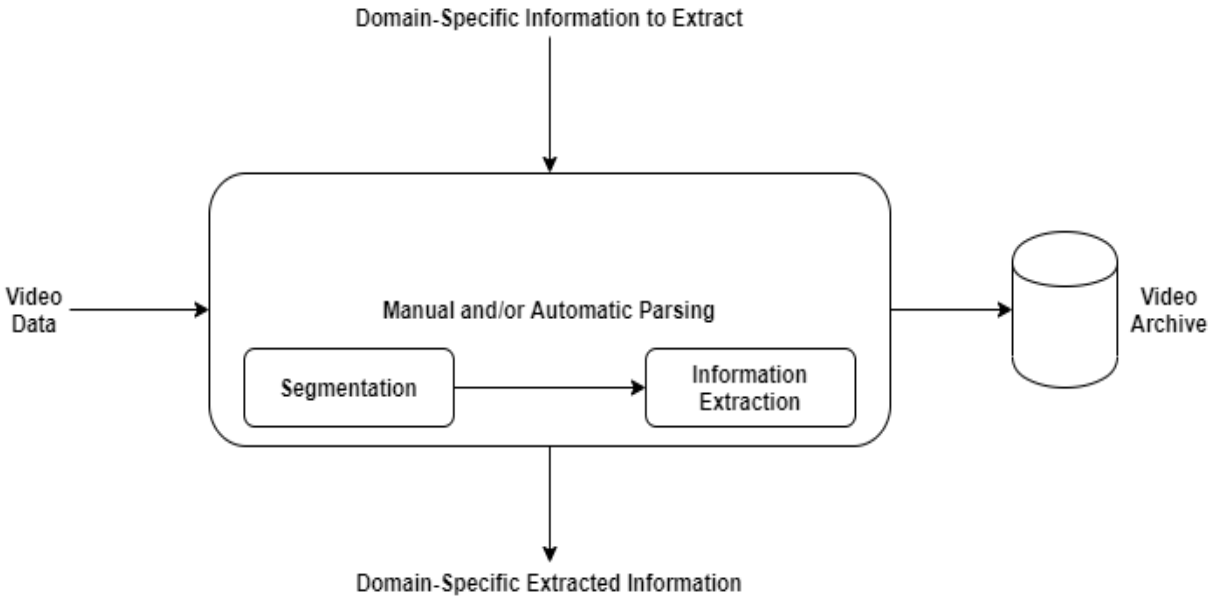


Fig. 2.4: **Video Data Adaptation and information extraction [22]**

Method-B: DC coefficient histograms. 1-D histograms of macroblock DC DCT coefficients are used. This method uses the color information present in the bitstream. Calculate color difference and intensity and then show the result.

When these methods applied on a video then accuracy measured with Recall and Precision parameter where-

Recall = detects/(detects+missed detects)

Precision = detects/(detects+false alarms)

Table 2.2: **Comparison of Video Indexing Methods[21]**

Author	Method	Recall in %	Precision in %
U. Gargi[21]	Method A	69	94
R. Kasturi[21]	Method B	68	50

2.3 Methods for Text Extraction

2.3.1 Region Based Approach

In recent years there is huge increase in multimedia libraries.The size of multimedia data is growing exponentially.Main reason for growing multimedia data is increasing in numbers of television channels that are broadcasting every day.Also due to advancement in technology cameras became affordable,memory device is inexpensive,multimedia data is increasing every second.Surveillance cameras to broadcast videos from phone's camera and various social networking application are adding enormous multimedia data.

With the huge increase in multimedia data,rapid use of internet and tremendous use of audio/video capturing device,content based indexing and text extracting is gaining more and more popularity in research community.

Text embedded videos is classified into two following groups-

1. **Caption text :-** It is laid over video during editing stage,for example score of match and name of the speaker. It is also called artificial text or superimposed text. It usually highlights the multimedia's contents.This provide caption text principally positive for construction of keyword index. Example Fig. 2.5.

2. **Scene text :-** It is actual part of the scene,for example brands of the products,street signs,name plates and text appearing on t-shirts etc.Scene text physically present in the scope of camera view during frame or video capture. Example Fig. 2.5.

Fig. 2.5: Architecture of Text Extraction Process [7]

Architecture of Text Extraction Process

Text extraction and recognition process consists of four following steps-

1. **Text detection**–In this phase,video frame is taken as input and decides whether it contains text or not. It also identifies the text regions in image.

2. **Text localization**– In this phase,Text localization merges the text regions to formulate the text objects and define the tight bounds around the text objects.

3. **Text tracking** - This phase is not for images,applied to video data only.For the readability purpose,characters embedded in the video appears in more than thirty consecutive frames per second(fps).Text tracking phase checks this temporal occurrences of the same text object in multiple consecutive frames.It can be used to correct the results of text detection and localization stage.It is also used to enhance the text extraction process by not applying the binarization and recognition step to every detected object.

4. **Character recognition** –This phase is last module of text extraction process is the character recognition. In this phase binary text object is converted into the ASCII text. Text detection, localization and tracking modules are closely related to each other and is the most challenging and difficult part of extraction process. Fig. 2.2 represents the output of different phases of the text extraction process.

Architecture of text extraction process is visualized in Fig. 2.2. Region based approach

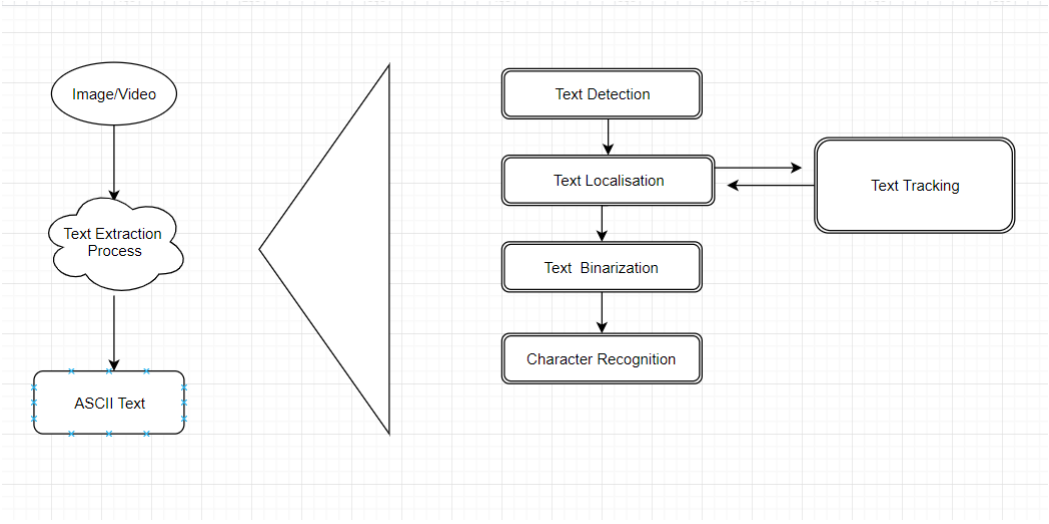


Fig. 2.6: Architecture of Text Extraction Process [3]

uses different region properties to extract characters.To distinguish text from background there

should be sufficient difference between text colour and its immediate background. Edge features,connected features and colour features are commonly used in this approach.This method follows bottom up approach by first segmenting the small regions and then grouping the potential text regions.

Modules of Region Based Approach

1. Segmenting the image into small regions which segregate the character regions from its background.
2. Grouping and merging of small regions to form words and sentences.
3. Differentiating between non-text and text objects.

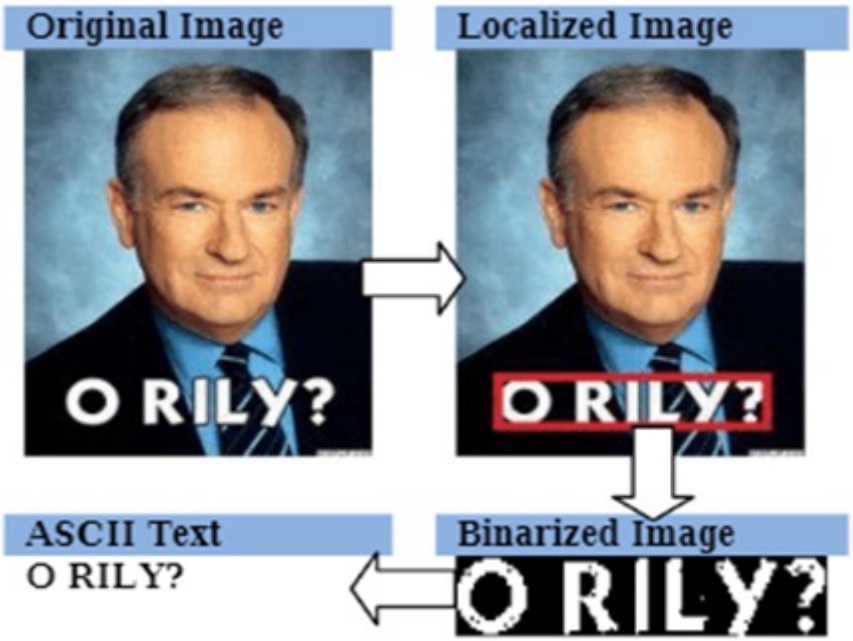


Fig. 2.7: **Modular Results of Text Extraction Process [9]**

2.3.2 Texture Based Approach

Texture based technique use the assumption that text in indexed frames carries distinct textural properties,which may be used to differentiate it from the background.Generally to extract the textural properties of a text region in an image. The usual approach is to use a classifier trained to divide regions to textual/non-textual based on texture features. These methods use machine learning and are less heuristic-based, but they are more computational expensive.

Here we are trying to combine region-based and texture-based method to explore the advantages of both groups.Our result is to perform a cascade filtering of image regions.Firstly we will apply heuristic-based region filtering in order to eliminate certain non-textual regions. It is followed by a more expensive and precise texture based filtering with Support Vector Machines (SVM).

Steps in Texture Based Approach

More precisely, our solution includes the following steps-

1. Pre-processing.
2. Divide an image into a number of segments of size (100*100 pixels) and histogram-based threshold filtering for segments.
3. A texture feature extraction mainly based on the Haar Wavelet Decomposition.
4. Process of filtering of segments with SVM.
5. Merging of segments to form the text blocks.
6. Extraction of text line.
7. Separation of words to detect single words.
8. Finally filtering of word regions by SVM.

Research on Texture Based Approach

1. Zhong[8] used local spatial variations in a grey-scale image to spot text regions with a high variance. They used a horizontal window of size 121 to calculate the spatial variance for pixels in a regional neighbourhood.Then horizontal edges in the image are identified using a Canny edge detector, and after that the small edge components are grouped into longer lines. From this edge image, edges with reversed directions are coupled into the upper and lower boundaries of a text line.However,this method can only detect horizontal part with a large amount of variation compared to the background.A 6.6 second processing time is accomplished with a 256*256 image on a SPARC station 20.
2. Wu[2] sliced an input frame by using a multi-scale texture segmentation scheme.Potential text regions are observed based on nine second-order Gaussian derivatives. A non-linear

transformation is tested to each drained image.The local energy believes, calculated at each pixel using the output of the nonlinear transformation,are then clustered using the K-means algorithm.This process is termed as texture segmentation. Next, the chip generation stage is initialized, which consists of 5 steps-

(a) Stroke Generation.

(b) Stroke Filtering.

(c) Stroke Aggregation.

(d) Chip Filtering.

(e) Chip Extension.

These texture segmentation and chip generation stages are implemented at multiple scales to detect text with a wide range of sizes.and then mapped back onto the original image. The complication with conventional texture-based methods is their computational complexity in the texture classification stage,which accounts for most of the processing time.In general, texture-based filtering approach requires an extensive scan of the indexed image to detect and pin point text regions.This causes the operation computationally expensive.To deal with this problem, Li[2] classified pixels at regular gaps and interpolated the pixels located between the classified pixels.However,this still does not detached the useless texture analysis of non-text regions and merely trades precision for speed.Jung[2] adopt a mean shift algorithm as a mechanism for accordingly selecting regions of interest (ROIs),thereby escaping a time-consuming texture analysis of the entire image. By enclosing a texture analyser into the mean shift,ROIs related to possible text regions are initially selected based on a coarse level of classification.Only the pixels within the ROIs are then restricted at a finer level, which naturally reduces the processing time when the text size does not control the image size.

2.3.3 Edge Based Approach

Text embedded in document in complex coloured and textured backgrounds are increasingly common today, for example,in web pages,in magazines and advertisements.Efficiently detection and extracting of text from these documents is a challenging problem.The procedure generated for ordinary documents, such as binarization by adaptive thresholding are not applicable in general,because it is almost impossible to find an optimal threshold or thresholds to preserve

meaningful information and to discard unnecessary one.

Steps of Text Detection from Complex Images

1. Edge detection using a 3x3 Sobel operator, thresholding and non-maximum suppression.
2. Edge image partitioning into small non-overlapping blocks and computing an edge-based property for each block.
3. Block classification either as text or as non-text based on the value of the edge-based feature.

2.3.4 Morphological Based Approach

In the previous sections we have discussed various approaches towards the text extraction.Now in this is yet another technique which is discussed below.The word "morphology" basically denotes a branch of science which deals with the form and structure of animals and plants.In digital image processing, mathematical morphology is considered as a tool for extracting components of image which are vital in the representation and description of region shape,such as boundaries and the convex hull.Here,we use some fundamental morphological operations such as dilation,erosion,opening,closing etc.The mechanism of morphology-based text line extraction algorithm used for extracting text regions from cluttered images.Firstly,this method defines a set of morphological operations for extracting important contrast regions as possible text line candidates.The most contrasting feature is robust to lighting changes and invariant against to different image transformations like image scaling,image translation and image skewing.In order to detect the skewed text lines a moment-based method is then used for estimating their orientations.Here the morphological based text line extraction method comes in picture.The extracting text lines from images or videos is a potential problem in many applications like document processing,image indexing,video content summary,video retrieval,video understanding and so on.Usually texts embedded in an image or a frame capture important media contexts such as player's name,title,date,story introduction and so on.Therefore, this task can provide us various advantages for annotating an image or a video and thus improves the accuracy of a content-based indexing system to search desired media contents. Moreover,when analyzing video audios, the recognition result of text line can provide extra refinements for correcting the errors of speech recognition.Extracting and recognizing text in images has become a potential

application in many fields like robotics,intelligent transport systems etc.In the field of multimedia,data extraction and recognition has become necessary.For example capturing number plate information through a video camera and extracting license number in traffic signals.The modules of such applications are-

i. Object Localization

ii. Object Extraction

iii. Text Recognition.

As explained in this technique,we have presented a robust approach for text extraction and recognition in images.In which first,the input image is filtered by the Median filter to eliminate any noises.After that the morphological dilation operation is applied for object localization.All the connected components are then extracted and all non-text character components are discarded by a two step process.Features are then extracted from the extracted components.These features form the feature vector for (Support Vector Machine)SVM.These features are tested with SVM for recognizing individual characters are merged to form texts.

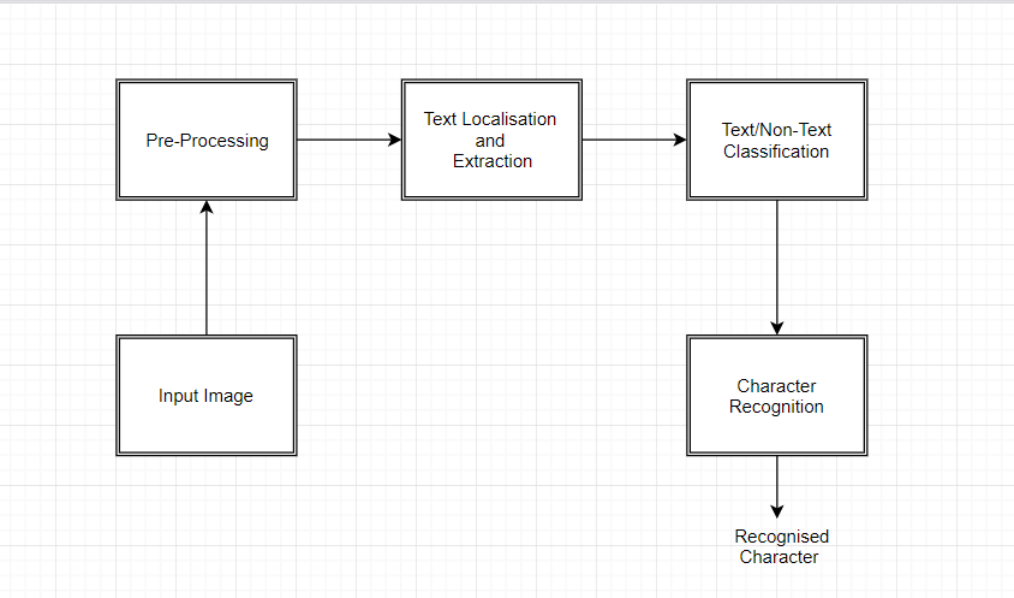


Fig. 2.8: Process of Text Extraction [4]

1. **Pre-Processing:** In this step,firstly the input RGB text is converted to gray-scale image.This conversion is done in order to diminish the processing overload.Median filtering is then applied to the delete and remove any noises present in that.Next edges are extracted from the resultant Image using LOG edge detection algorithm. The choice of

using LOG edge detector is for the reason that it finds the correct places of edges and testing broader area surrounding the pixel.

2. **Text localization and Extraction:** In this phase, the edge image obtained from the last step is binarized. Then the Morphological dilation operation is performed at this edge map. Since texts are normally aligned in the horizontal direction we have used a 2*4 rectangular structuring element.

3. **Character Recognition:** Support Vector Machine: SVM mainly performs the classification tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels. Given some set of data points, SVM tries to find an optimal hyper plane that correctly classifies these data points to two or more classes.
   Here, the main reason of using Support Vector Machine (SVM) is-
   i. To classify components as text and non-text components.
   ii.To recognize characters.

2.3.5 Connected Component Approach

Connected Component-based approach uses a bottom-up approach by grouping smaller components into successively larger components as far as all regions are identified in the image. A geometrical analysis is needed to group the text components by using the spatial adjustment of the components so as to filter out non-text components and imprint the boundaries of the text regions.

Research on Connected Based Approach

1. Ohya[8] given a four-stage method for Text Extraction-
   i. Binarization based on regional thresholding.
   ii. Tentative character component detection by using grey-level difference.
   iii.Character recognition for evaluating the similarities between the character candidates and the standard patterns that were stored in a database.
   iv. Relaxation operation to restores the similarities.
   They were able to notice characters, including multi-segment characters, under varying illuminating conditions, positions, and fonts when handling with scene text images, such as signboard. nonetheless, binary segmentation is not proper for video documents, having many objects with varying grey levels and high levels of fluctuation and noise in

illumination. Moreover, this approach creates several constraints related to text alignment, such as upright and not connected, as well as the colour of the text. Based on several analysis involving 100 images, their recall rate of text localization was 85.4% and the character recognition rate was 66.5%.

2. Lee and Kankan hallien[12] forced a Connected Component-based method to the disclosure and recognition of text on cargo containers, which may have irregular lighting conditions and characters with varying sizes and shapes. Edge information is used for a coarse search before the Connected Component generation. The difference between adjacent pixels is used to resolve the boundaries of probable characters after quantizing the input image. regional threshold values are then selected for every single text candidate, based on the pixels on the boundaries. These probable characters are used to generate Connected Components with the same grey-level. After that, several heuristics are used to filter out non-text components based on contrast histogram, aspect ratio and run-length measurement. although their claims that the approach could be effectively used in other domains, empirical results were only presented for cargo container images.

3. Lienhart[10] regard text areas as Connected Components with the same or similar colour and size, and apply motion study to augment the text extraction results for a video sequence. The input image is disjointed based on the monochromatic nature of the text components by using a split and merge algorithm. Segments which are too small and too large are filtered out. After expansion, motion information and contrast analysis are used to augment the extracted results. A block-matching algorithm which uses the mean absolute difference criterion is engaged to assess the motion. Blocks that are missed during tracking are abandoned. Their main focus is on caption text, such as pre-heading sequences, credit title, and closed sequences, which display a higher contrast with the background. This makes it easier to use the contrast difference between the boundary of the noticed components and their background in the filtering stage. At last, a geometric study, including the width and height and aspect ratio, is used to filter out if there is any non-text components. Based on analysis using 2247 frames, this algorithm extracted 86% to 100% of all the caption text. Given below figure shows an example of their text extraction process-

In today's era, there is rapidly increasing demand of text information extraction from frame.

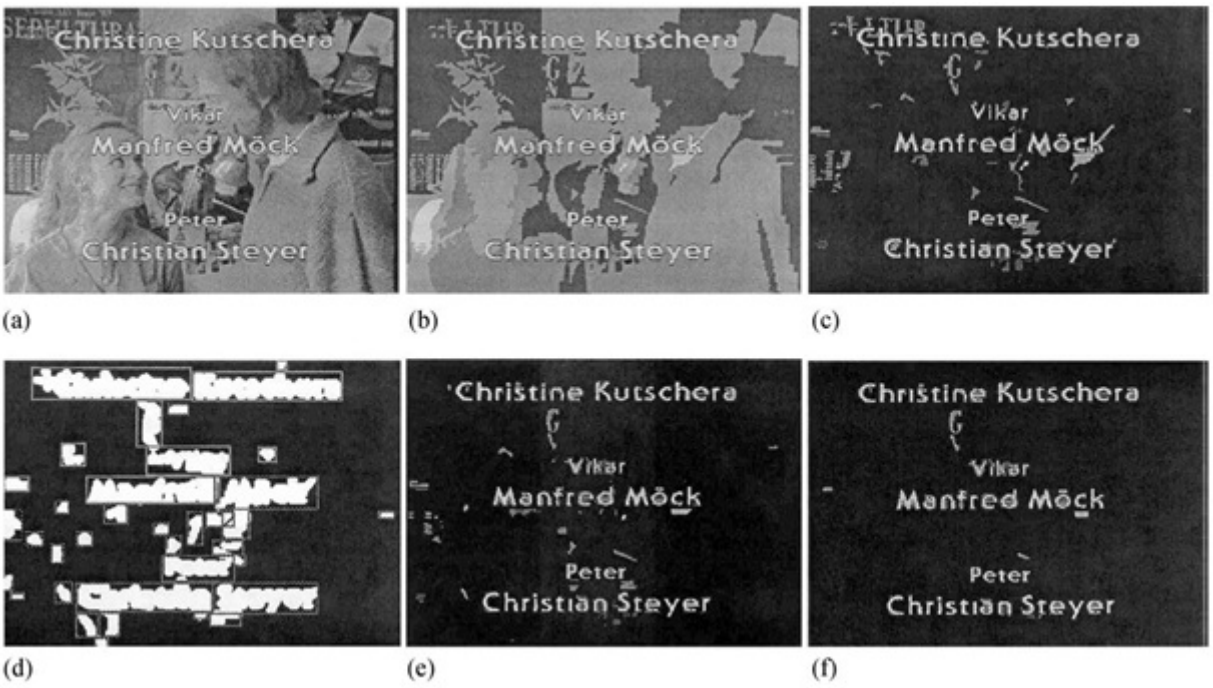


Fig. 2.9: Intermediate stages of processing in the method by Lienhart [10]
(a) original video frame; (b) image segmentation using split-and-merge algorithm;
(c) after size restriction; (d) after binarization and dilation;
(e) after motion analysis; and (f) after contrast analysis and aspect ratio restriction.

So,there are many extracting techniques for retrieving relevant information have been developed and some are in process.Furthermore,extracting text from the colour image takes more time that leads to user dissatisfaction.Here we are exploring a method to extract the text from image which extracts text more accurately.This method is tested with various types of images, both the images with caption text and scene text.Using our method it is easier to extract information within short time. Due to their relatively simple and easier implementation, Connected Component-based methods are widely used.

Processing Stage of CC based Approach

1. Pre-processing, such as colour clustering and noise reduction.
2. CC generation.
3. Altering out non-text components.
4. Component grouping.

Despite,our connected component based approach for text extraction from colour image method has several features than existing method but it becomes very less effective when the text is too

small.In this above case, the text region is not clearly visible or the colour of the text is not visible in proper way.

2.3.6 Performance Analysis of Text Extraction Methods

The performance of Text extraction methods are analyzed by various parameters using different types of data set and their performance are measured as mentioned below in the Fig. 2.6.

i. Detection Rate(DR) = Correct Detected Text / Ground Truth Text

ii.Precision Rate(PR) = Correct Detected Text / (Correct Detected Text+False Positive Text)

iii.Recall Rate(RR) = Correct Detected Text / (Correct Detected Text+False Negative Text)

iv.False Alarm Rate(FAR) = No. of Text Blocks falsely detected / Total no of Text block

Table 2.3: Comparison of Text Extraction Methods[12]

Author,Year	Technique Used	Dataset(Images)	Parameters	Remarks
Yao,2017[12]	C.C.and S.V.M.	Complex Background	PR=64%,RR=60%	Same Color
Liu,2016[2]	Edge Based	Signboard Images	PR=70%	Uneven Illumination
Zhong,2015[8]	Morphology	Different Languages	PR=94.5%	Color Independent
Lien.,2015[10]	Texture	Complex Background	PR=95%	Robust and Effective
W.H.,2015[1]	Edge Based	Logo Detection	PR=95.6%	Efficient Detection

Chapter 3

Proposed Methodology

The main goal of this methodology is to approach for automated video indexing and video search from video lecture archives. The methodology further aims to apply automatic video segmentation and key-frame detection to offer a visual guideline for the video content extraction in the order of their appearance in the video. Extract textual metadata by applying video Optical Character Recognition (OCR) technology on key-frames.

3.1 Proposed Model

In recent chapter there are some methods discussed which is used for text extraction from a video and indexing of that content. Video is a collection of different images. Text extraction from a video is not a easy task because in a video many types of data. Suppose if there is video where teacher teaches to students using projector then there are different slides showing on projector. For indexing of content which is present on slides in video. So for indexing the video content and respective time of user wanted data we follow the flow which is shown in Fig. 3.1

3.2 Frame Generation from Video

First we take a video as a input and generate frame using opencv with fps value=30. Frames as a part of video at a particular instance.Even for a small video many frames are generated.

Number of frames in a video = Time duration of video * 30

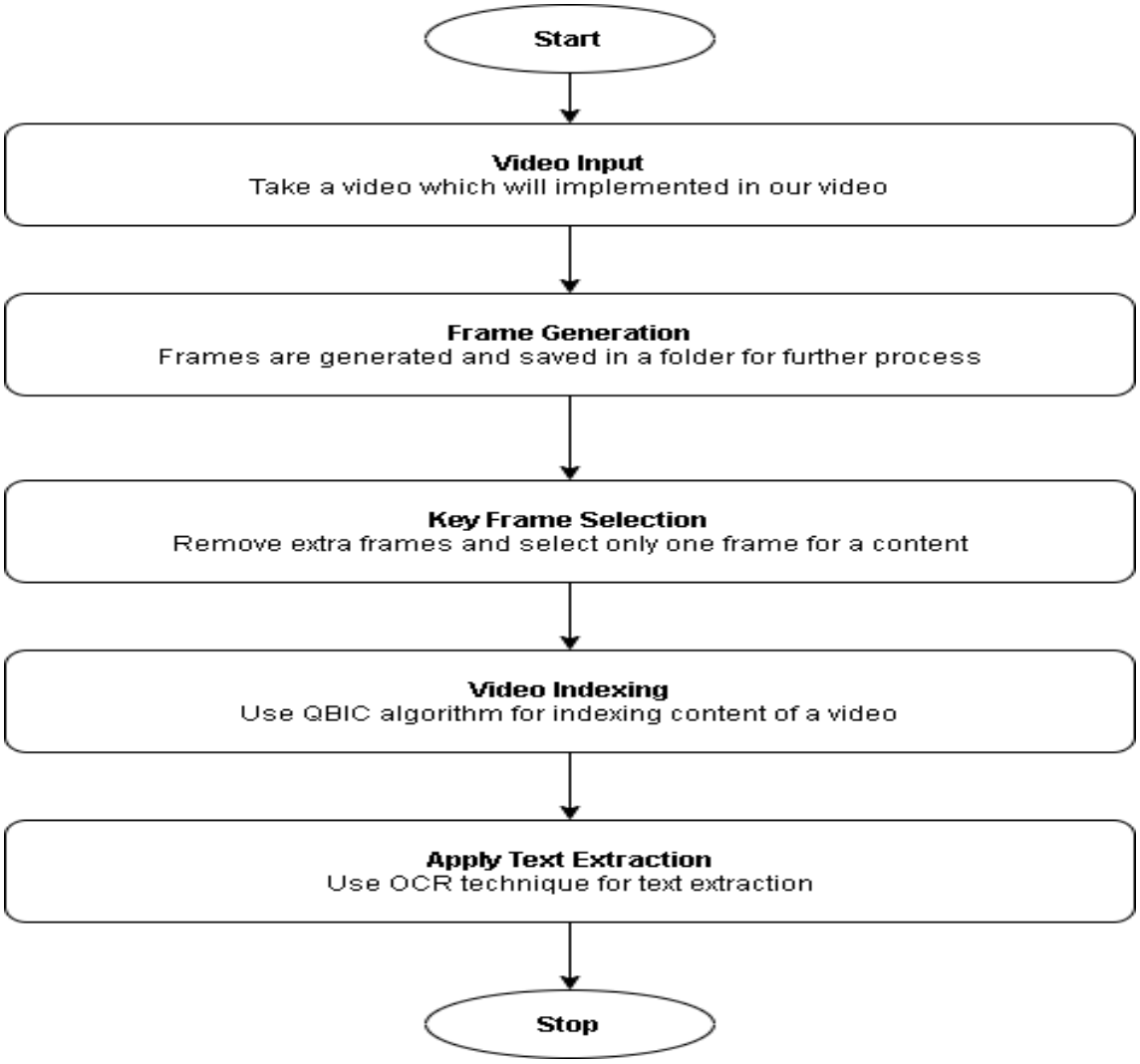


Fig. 3.1: Flow of Model

Process of Frame Genenration-

1. Open the Video file or camera using cv2.VideoCapture()
2. Read frame by frame
3. Save each frame using cv2.imwrite()
4. Release the Video Capture and destroy all windows

3.3 Subset Algorithm for Key Frame Selection

In a video there are many clips. In one clip usually contains more than one keyframe. The number density of keyframes in a clip can be problematic if video is long.

Here key frame is basically that frame having text which should not repeat in multiple frames. So after generating all the frames key frames are extracted using Subset Method by using given algorithm -

1. 2 Image frame is given as input.
2. Then we calculate text difference using concept of subset, if text of frame 1 is subset of text of frame 2 then frame 1 is not taken.
3. Step 2 is repeated until all frames are not checked.

3.4 QBIC Algorithm for Video Indexing

Video has become an important element of multimedia computing and communication environments with applications as varied as broadcasting, education, publishing, and military intelligence. However,video will only become an effective part of everyday computing environments when we can use it with the same facility that we currently use text. Our architecture is based on the assumption that video information will be maintained in a folder of computer.

The process of automatically assigning content-based labels to video documents is called video index. Video indexing is comparable to text indexing or bookmarking. Raw video streams converted into Structured and indexed database-driven information entities. In a video document, three information channels are considered-

- Visual modality
- Auditory modality
- Textual modality

In this model we are focusing on textual modality feature of video because now a days many study videos which is only based on this feature. So using our model anyone can easily find the proper index of their content or keyword and save lot of time. Each keyframe is indexed by given QBIC(Query By Image Content) algorithm-

1. Indexing method based on text which is present on slide in video.
2. Still video is converted into video clip is segmented into small units (frames) (shots) then one or more keyframes in each unit are selected.

3. Each keyframe is automatically indexed using its text characteristics and we calculate timestamp of each keyframe and save it.

4. we apply text extraction on that frame and search wanted text in those keyframes.

5. After getting text in keyframe show respective time of that keyframe.

3.5 Edge Based Connected Component Method(OCR Engine)

Edge-based extraction algorithm is able to detect and extract text in complex images. This method is able to extract text both printed document and also scene text images. It is insensitive to color and intensity of image. It is robust to text size and alignment of text. This method analyze only text blocks and this method is computationally very efficient and useful for real time applications.

It has three stages

1. Candidate text region detection

2. Text region Localization

3. Character extraction

The binary output after text region localization is used as input to our OCR Engine for text extraction for further processing. It is difficult to extract text from a complex and highly colorful images.

For overcoming this problem connected-component method is used. This method follow bottom up approach. Initially the color image is converted to grey scale image and then this grey scale is converted into a binary image. This method extract the text from the image more accurately but it is not effective for small text segment.

3.5.1 Working of OCR

- Image as a input has to be given from which we have to extract text.This image is either to be a picture or scanned one.it is stored as bitmap format.

- To make the image properly aligned we need to apply de-skewing i.e tilting in clockwise and anti-clockwise direction.and also remove noise to improve the quality of image.

- Binarization is done to convert an image to black and white.It is used to separate the text from the background.This is curious because inaccurate binarization will cause lot of issues.
- Detection and removal of lines.
- Combined and broken character analysis.
- Isolation of characters,multiple characters that are connected must be separated and single characters that are into multiple pieces must be connected.
- Classification of characters,the text are divided into lines and then into characters and after that character is recognized using various algorithm.algorithm such as Matrix matching and feature extraction is used to produce a ranked list of candidate character.
- Dictionary support,It helps to improve the recognition quality.characters like "C" and "G" can look similar,so dictionary can help to make decisions.
- At last the result is saved in the selected output format such as PDF,DOC etc.

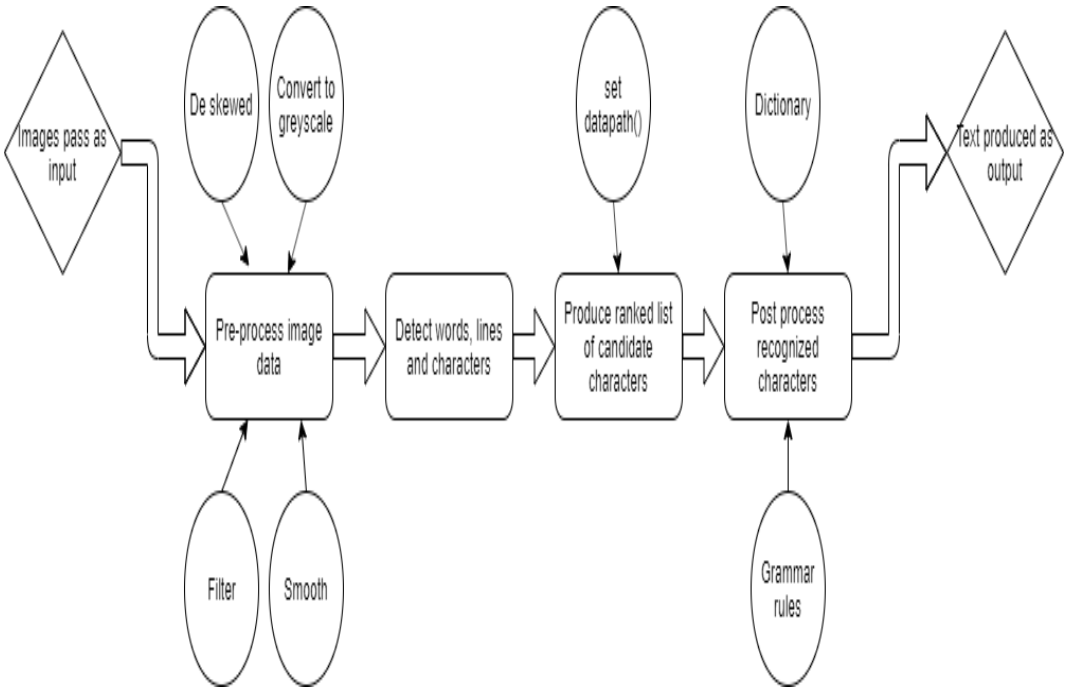


Fig. 3.2: Working of OCR Tool [12]

OCR scans every line of the frame and then check it to find if the combination of 0 and 1 represents any particular number letter or symbols.

Fig. 3.3: **Sample Image [12]**

Fig. 3.4: **Reads a Image [12]**

Chapter 4

Simulation and Results

We design this model in frontend and backend.In frontend we create a GUI where title bar is there, a canvas window where video is running continuously.Right side of this canvas window result window is there where accuracy of model for that particular text and indexing of text is there.At the below of that window video controller option is there like start button, stop button, restart video button, volume controller, find button. In backend all other process like frame generation, key frame selection, video frame indexing, text extraction process is executed and show result on our GUI window Please refer Fig. 4.1 and Fig. 4.2.

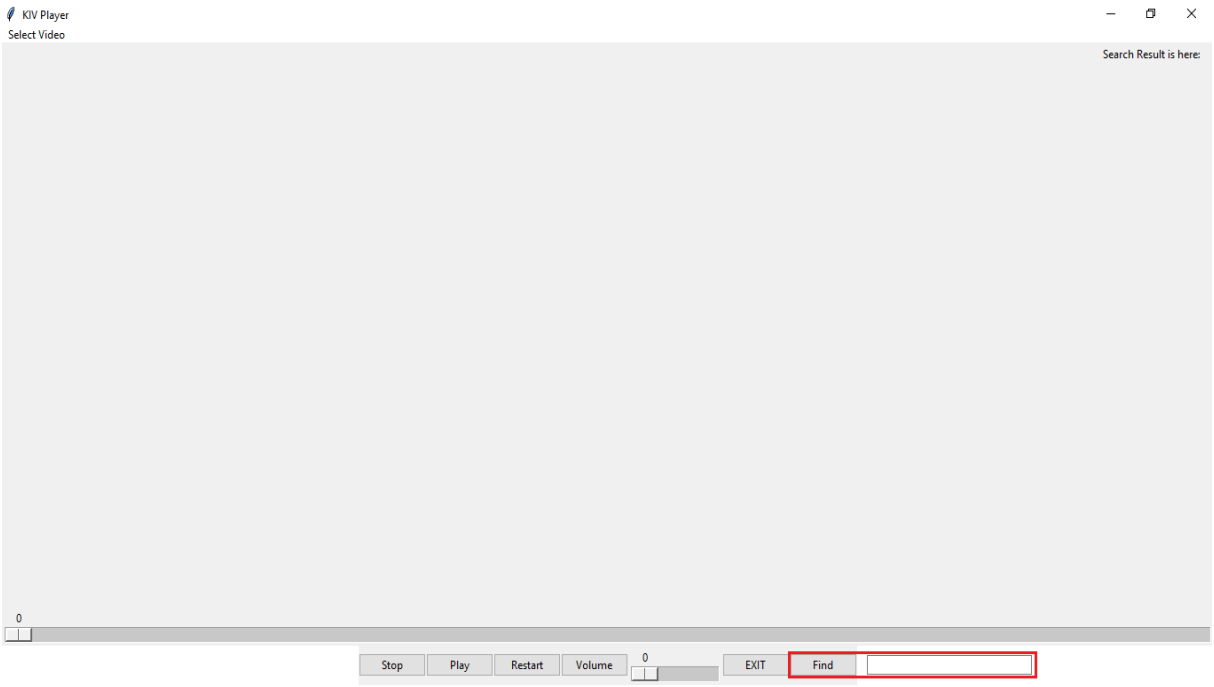


Fig. 4.1: Opening window of GUI

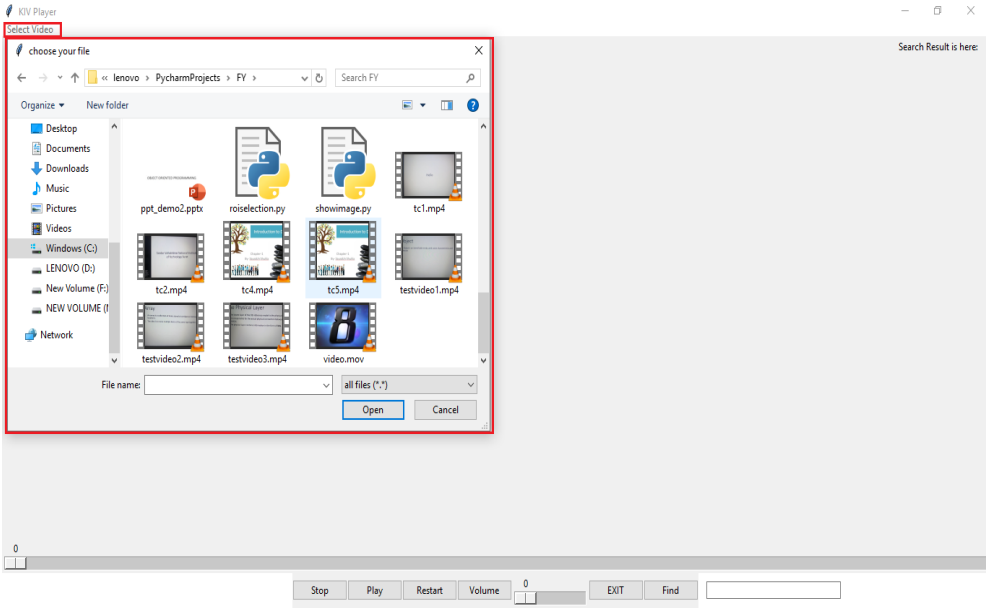


Fig. 4.2: Select Video in GUI

4.1 Frame Generation and Key Frame Selection

First we take a video of 17 second as a input and generate key frames using algorithm which is discussed in chapter 3. In our model fps(frame per second) is 30 so according to this 510 frames should be generated but only 18 frames are generated.In given Fig. 4.3 key frame is shown-

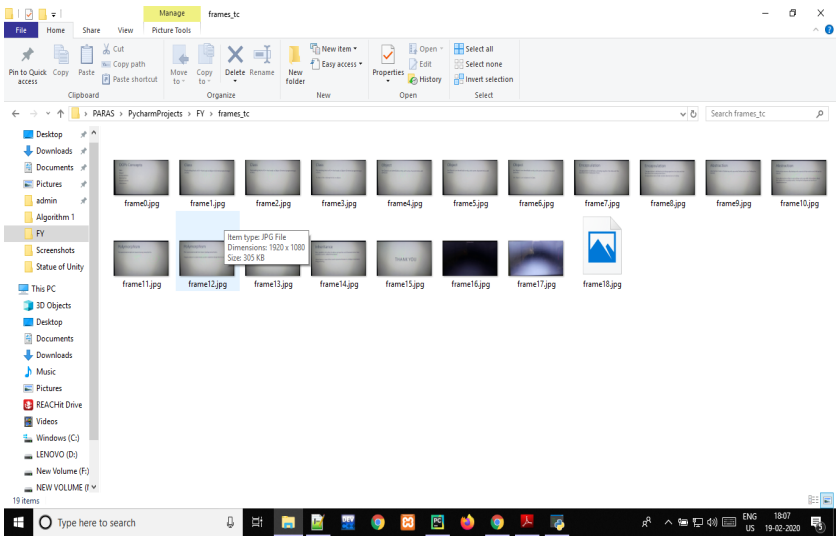


Fig. 4.3: Generated Key Frames

4.2 Video Frame Indexing and Apply Text Extraction

For video frame indexing we save time of each frame.User wants which data to find, we extract data from each key frame and compare with user data.If data is found in any frame then respective time and accuracy shows else a message shown.Please refer the below figure (Fig. 4.4 and Fig. 4.5) for better understanding-

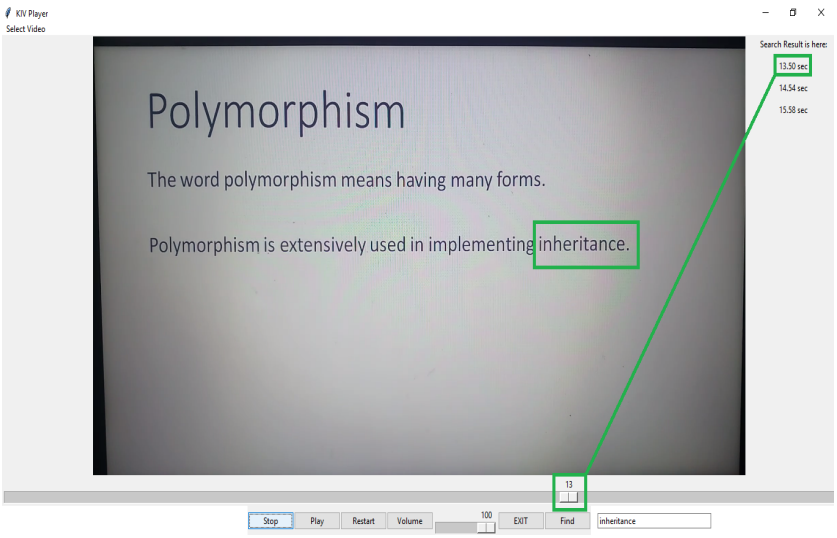


Fig. 4.4: Indexing of Searched data

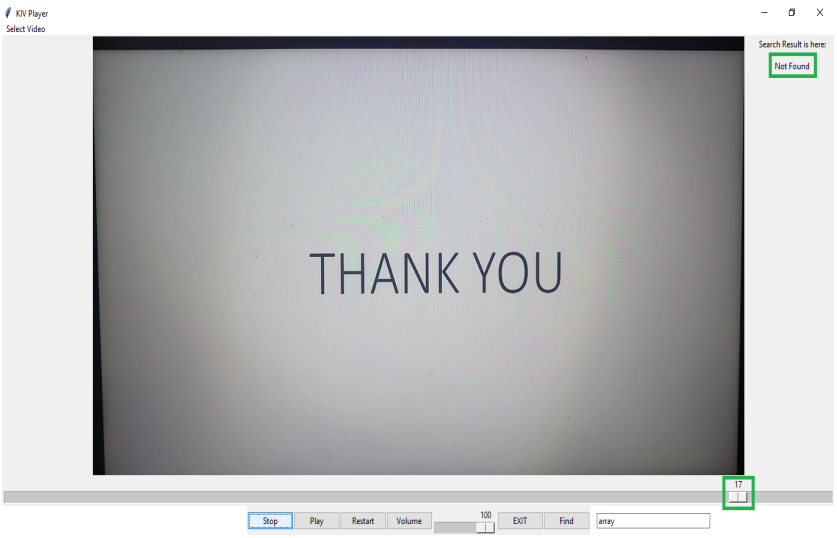


Fig. 4.5: Data Not found message

4.3 Results

We implement this model on study related e-learning videos which is generally seen by students.In this dataset[22] 500 videos are there.These videos are in different font styles with accuracy which are given below-

1. Arial Rounded MT Bold : 93.57%
2. Arial : 93.74%
3. Berlin Sans Fb : 94.66%
4. Calibri Light : 95.95%
5. Cambria Math : 95.23%
6. Elephant : 95.22%
7. Engravers MT : 93.00%
8. Freestyle Script : 73.02%
9. Gadugi : 95.54%
10. Lucida Sans Unicode : 94.83%
11. Microsoft JhengHei : 93.97%
12. Myanmar Text : 95.55%
13. Nirmala UI : 96.17%
14. OCR A Extended : 91.59%
15. Perpetua : 96.23%
16. Script MT Bold : 65.50%
17. Segoe UI : 86.65%
18. Times New Roman : 94.74%
19. Verdana : 95.89%

20. Yu Gothic : 94.80%

These are styles which is generally used in e-learning contents.So we take these style as a standard measure and according to that accuracy of model is 91.79% .

Table 4.1: Comparison with Standard Data

Dataset	Standard Accuracy	Our Model Accuracy
Standard Dataset [22]	82%	91.79%

Chapter 5

Conclusion and Future Work

Now a days everyone wants to save time in any manner.Indexing of video with respect to audio is already done but in this project we have implemented indexing of video with respect to content and whatever keyword user wants this model display presence of that keyword.As per the discussion throughout the project we have presented the recent developments in visual content - based video indexing and keyword retrieval.The various important task involved in the project was divided in various modules such as frame generation, key frame selection and then proper indexing was also done which returns not only keyword but also their time of appearance.We tested our model with the dataset using different videos and the accuracy is 91.79%.Finally the result was displayed on the GUI.

This project basically works on only those video in which content should be written in English language but according to use in future this model can implement in different languages also.After some modification this model also can use for different purposes like news duplication, audio and content indexing of videos.Although the model was working with better accuracy still there has space left for improvement which has to be done such as the performance enhancement. Apart from that the feasibility of the commercial use of the model is another issue which needs to be done for the much better application.

References

[1] Gongqing, W., Jun, H., Li, L.L., et al.: Online content extraction based on label path feature fusion. J. Softw. 27(3), 714–735 (2018).

[2] Wu, Jung G.Q., Hu, J., Li, L., Xu, Z.H., Liu, P.C., Hu, X.G., Wu, X.D. :Web news extraction via tag path feature fusion. Ruan Jian XueBao/J.Softw. 27(3), 714–735 (2018).

[3] Jiazhen, C., Yan, G., Qiang, L., et al.: An automatic text extraction method for short text web pages. Chin. J. Inf. Sci. 30(1), 8–15 (2016).

[4] Q. Ye, D. S. Doermann, “Text Detection and Recognition in Imagery: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37(7), pp. 1480-1500, 2015.

[5] V. Khare, P. Shivakumara, P. Raveendran, M. Blumenstein, “A blind deconvolution model for scene text detection and recognition in video”, Pattern Recognition, Vol. 54, pp.128-148, 2016.

[6] A. Gonzalez, L. M. Bergasa, J. J. Yebes. "Text detection and recognition on traffic panels from street-level imagery using visual appearance", IEEE Transactions on Intelligent Transportation Systems, Vol. 16(3), pp. 228-238, 2015.

[7] A. K. Bhunia, A. Das, P. P. Roy, U. Pal, “A Comparative Study of Features of Handwritten Bangla Text Recognition”, In Proceedings of International Conference on Document Analysis and Recognition, pp.636-640, 2015.

[8] Zhong, A., X. Peng, X. Zhuang, P. Natarajan, H. Cao,Ohya “Text detection and recognition in natural scenes and consumer videos”. In Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp. 1245-1249, 2014.

[9] H. Yang, B. Quehl, H. Sack, “A framework for improved video text detection and recognition”, Multimedia Tools and Applications, Vol. 69(1), pp.217-245, 2014.

[10] Lienhart, S. Roy, P. Shivakumara, P. P. Roy, U. Pal, C.L. Tan, T. Lu, “Bayesian classifier for multi-oriented video text recognition system”, Expert Systems with Applications, Vol. 42(13), pp.5554-5566, 2015.

[11] Chitrakala Gopalan,Lee, and kankan hallien D. Manjula, “Contourlet Based Approach for Text Identification and Extraction from Heterogeneous Textual Images”, International Journal of Computer Science and Engineering , 2-4, 2013.

[12] P. Shivakumara,Yao, R. Raghavendra, R., L. Qin, K. B. Raja, T. Lu, U. Pal, “A new multimodal approach to bib number/text detection and recognition in Marathon images”, Pattern Recognition, Vol. 61, pp.479-491, 2017.

[13] Tinne Tuytelaars1 and Krystian Mikolajczyk2, “Local Invariant Feature Detectors: A Survey,” Foundations and Trends Rin Computer Graphics and Vision, May (2008).

[14] Mahesh and Dr.M.V.Subramanyam “invariant corner detection using steerable filters and harris algorithm” signal Image Processing : An International Journal (SIPIJ) Vol.3, No.5, October (2012).

[15] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Simultaneous structure and texture image inpainting”, Proc. Conf. Comp. Vision Pattern Rec., Madison, WI, 2015.

[16] T. Chan, J. Shen, Non-texture inpainting by Curvature-Driven Diffusions (CCD), J. Vis. Commun. Image Represent. 12 (2016).

[17] L. Liu and G. Fan, “Combined key-frame extraction and object-based video segmentation,” IEEE Transactions on Circuits Systems for Video Technology, vol. 15, no. 7, pp. 869–884, 2016. View at Publisher · View at Google Scholar · View at Scopus

[18] S. Lei, G. Xie, and G. Yan, “A novel key-frame extraction approach for both video summary and video index,” The Scientific World Journal, vol. 2014, Article ID 695168, 9 pages, 2014. View at Publisher · View at Google Scholar · View at Scopus

[19] Choudhary, R., Raina, N., Chaudhary, N., Chauhan, R., Goudar, R.H.: An Integrated Approach to Content Based Image Retrieval. 2404–2410 (2014).

[20] M. Ravinder and T. Venugopal. Content-based video indexing and retrieval using key frames texture, edge and motion features. 2016.

[21] U. Gargi, R. Kasturi, S. Antani Performance Characterization and Comparison of Video Indexing. 2017

[22] G. Ahanger, Thomas D. C. Little, A Survey of Technology for Parsing and Indexing Video. 2014

[23] K. Shubhashini Audio Video Indexing and Retrieval, https://shodhganga.inflibnet.ac.in/ 2013

Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them. We are highly indebted to Dr. Krupa N. Jariwala for her guidance and constant supervision as well as for providing necessary information regarding the project. We would like to express our gratitude towards our department for their kind co-operation and encouragement which helped us during this project. We would like to express our special gratitude and thanks to our mentors Mr. Gaurav and Mrs. Rasika for giving us such attention and time. Special thanks and appreciations to our colleague in developing the project and people who have willingly helped us out with their abilities.

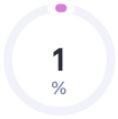
U16CO061.edited

by Rocean Wang

General metrics

51,088	2,079	859	8 min 18 sec	15 min 59 sec
characters	words	sentences	reading time	speaking time

Plagiarism



1 source

1% of your text matches 1 sources on the web or in archives of academic publications