Text Recognition and Extraction from Video

Kiran Agre
B.E. Student of Computer branch
Atharva College of Engineering
Mumbai, India

Ankur Chheda
B.E. Student of Computer branch
Atharva College of Engineering
Mumbai, India

Sairaj Gaonkar
B.E. Student of Computer branch
Atharva College of Engineering
Mumbai, India

Prof. Mahendra Patil
Head of Computer Department
Atharva College of Engineering
Mumbai, India

***Abstract*— Videos have become a great source of information. The text in the video contains huge amount of information and data. But this information is not in editable form. If this text is converted to an editable form it becomes simpler and efficient to store useful information. The paper describes the technique that aims at extraction of the text which occurs is video. The main focus of the proposed system is on educational and news video. The user will have to provide the video as input from which he wants to extract text. The system will process the video and generate the text output in editable text file.**

***Keywords*— Frames, Text Recognition, MESR, OCR, Gray Scale, MPEG.**

I. INTRODUCTION

With the rapid advancement in technology and the increasing speed of internet, the focus of the people is shifting from Television to YouTube. The main advantage of YouTube over television is that YouTube provides shows at user's preference irrespective of time. Television has programs that are shown at a particular fixed time which creates time constrain for users. As the focus is shifting towards YouTube the paper has proposed system which makes it easy for user to access information contained by the text in these video in efficient and quicker way. The proposed system will convert the text in the video into editable form which is stored in a text file.

YouTube is used widely used for news and educational videos. These videos contain text which adds information to videos and makes it more meaningful. If the text from the videos is converted to editable form, it can be stored efficiently and it will be easier to access it next time. Once the user has watched the educational video, next time he may not want to go through the entire video as he has already watched it and reading the main points may be sufficient for him to revise the topic from that video. In such case the proposed system helps user to get access to the information by converting to text in video to editable form. The editable from is a text file format. The main advantage of text file format is that it requires very small size as compared to the size of a video. Also the information in text can be edited if it changes in future or if user wants to add any additional information in it which is not possible in case of video.

The working of the proposed system is very simple. User downloads the video form the YouTube or any other website from which he wants to extract text. This video is provided as input to the proposed system. Proposed system converts video into series of frames and applies text detection and extraction on each frame. The detected text from each frame is stored in text file.

II. LITERATURE REVIEW

Datong Chen, Jean-Marc Odobez [1] have proposed the system that minimizes character error rates and also removes noise from the character that greatly disturb the optical character recognition.

Mati Pietikainem , Oleg Okun [2] have proposed combined edge based text detection that minimizes degradation in extracted text and can work with images having complex background.

C. P. Sumati , N. Priya [3] have proposed combined edge based method [2].This method is sensitive to skew and text orientation.

Z. Cennekove, C. Nikou, I. Pitas [4] have proposed the system that uses entropy based metrics. It involves checking color histogram for each frame against the histogram of the next consecutive frame. This method fails when two different images having exactly same color histogram values.

Priti Rege, Chanchal Chandrakar [5] has explained text image separation in document images using boundary/perimeter. Text detection is performed using sobel operator and thresholding. As text enhancement is not been used the extracted text can be noisy.

Arvind, Mohamed Rafi [6] have explained text extraction using connected component based method. The prerequisite for this method is that, text should have more contrast compare to its background.

Lifang Gu [7] explained text detection in MPEG (Moving Picture Experts Group) video frames. It reduces spatial and temporal data redundancies. This method is only applicable to MPEG videos.

Baseem Bouaziz, Tarek Zlitni, Walid Mahdi [8] explained automatic video text extraction. It performs content based video indexing. This method can detects only static superimposed text.

Punit Kumar, P. S. Puttaswamy [9] have proposed the system that performs area based filtering to eliminate noise blobs present in the image. This method fails when background has greater intensity transitions.

III. PROBLEM STATEMENT

There are two types of text occurring in a video.

- Natural text.
- Superimposed text.

a) Natural/scene text: Natural text is the text which occurs in the video when it is being recorded. These texts are part of scene where video is recorded. Example: House number, Car plate number.

Figure 1 Natural Text

b) Superimposed text: Superimposed text is the text which is not part of video when it is recorded but is superimposed to give extra information about that particular scene. Example: Text occurring in News Video.

Figure 2 Superimposed text

Natural text is of not great use as it contains information of less significance but superimposed text contains information which is of great importance. Hence the main aim of proposed system is to detect superimposed text occurring in the video.

IV. EXISTING SYSTEMS AND THEIR GAPS

There are many systems developed to detect the text in video. Each system is based on a particular method and has a drawback associated with it.

Some of the commonly used methods to detect text are

- Sliding Window based method
- Connected Component based method

3.1 Sliding Window based method:

This method uses sliding window to search for a specific text. It starts by taking small rectangular patch of the given image. This rectangular patch is of specific dimension. This rectangular patch is slide over the entire area cover by the image to check whether or not there is text in that image patch. Different sliding window classifiers are used to decide if there is text in the patch. The window is initially placed at the leftmost top corner of the image and slides over the different locations of the image starting with the first row and then going in the further rows of the image. This method is slow as image has to process in multiple scales. Even if the text is present at the bottom of the image window has to start from the top of the image. Also the accuracy of the detected text is depends on the dimensions of the window.

3.2 Connected Component based method:

In connected component based approach first we extract pixel regions which have similar color, edge strength or texture and evaluate each one of them for being text or non-text using machine learning techniques.

Connected component based method is efficient for caption text with plain background images but it doesn't works well for images with clustered background.

V. COMPARISION WITH PROPOSED SYSTEM

Unlike previous systems which showed the detected text in the video frame the proposed system will store the text output in a separate text file. The advantage of this feature is that the algorithm need not run every time the video is played. The proposed system does not compares consecutive frames for detection of text region which the system proposed by Z. Cennekove, C. Nikou[4] as it may assume any new object introduced in successive frame as text. The proposed system is able to detect text even if there are two sentences with different font size.

VI. SYSTEM OVERVIEW

The proposed system has three main components:

- Frame Generation
- Text Recognition and Extraction
- Text File Generation

4.1 Frame Generation: In this step, the video is converted into frames. Frames are the images of a particular time of a video. At regular time interval the frames are generated so that text in the successive frame is not repeated very often. These frames can be saved in any image format.

The user will have two options while converting video to frame. The first is convert entire video and second is converting a selected portion of video. When users want text from entire video they will select first option. If users want text from only from a particular time frame, they will select the second option where user will be able to select start time and end time for text extraction. The selected portion of video will converted to images which will be stored in a separate folder for easy access while applying text extraction algorithm on it.

4.2 Text Recognition and Extraction: This step is applied on every frame. In this step the text region is detected using algorithm described in the next part. The detected text regions are then refined to increase the efficiency of extracting text. Text Extraction algorithm is the applied to the detected regions. The efficiency of detecting text depends on font color, text size, background color and resolution of the video [9].

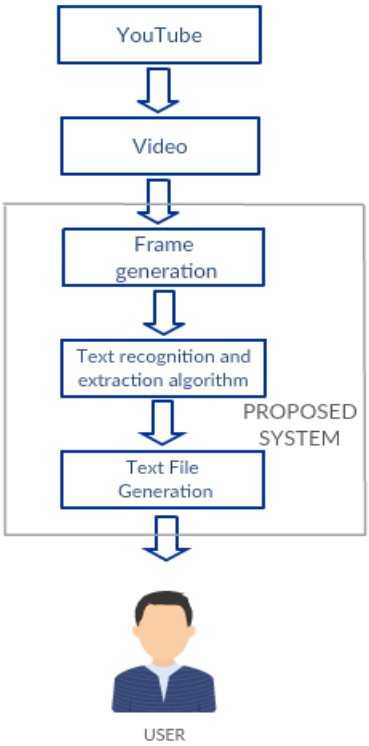

Figure 3 System Overview

4.3 Text File Generation: The extracted text is stored in a text file. For every frame the generated text is appended to the previous text in the text file and stored. At the end of extracting text from all the images the path of the output file will be given to user. Size of the text file is very less as compared to the size of the video. This saves memory and also makes quicker access to information possible.

VII. TEXT RECOGNITION AND EXTRACTION ALGORITHM

Step-1 Text Region Recognition: The MSER (Maximally Stable External Region) algorithm is used to detect candidate text region from the given image. MSER first converts the color image into gray scale image. It selects the regions which stay in the range of given threshold. All the pixels above or equal to a given threshold are black and all the pixels below given threshold are white.

Step-2 Removal of Non Text Region: The MSER may also detect non text regions. Stroke width is used to discriminate between text and non-text regions. Stroke width is a measure of the width of the curves and lines in the characters. Text region will have little stroke width variations whereas not text region will have larger variations.

Step-3 Merge Text Regions for Final Detection: All the detection results are composed of individual text characters. To use this result for recognition task, the individual text characters must be merged into words. This enables the recognition of actual words in an image.

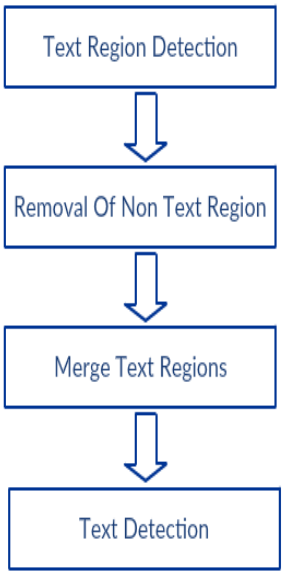

Figure 4 Text Recognition and Extraction Algorithm

Step-4 Recognize Detected Text Using OCR: After detecting the text regions, the OCR method such as edge based method is used to recognize the text. The detected text regions are then refined to increase the efficiency of extracting text.

VIII. FUTURE SCOPE

The proposed system can be enhanced by allowing the user to select a particular portion of the screen and then only the text occurring in that part will be extracted. This will be useful in situation where user wants text from only a particular region. Another enhancement that can be made is that instead of providing the video as input the user can directly provide the URL of the video and the system will auto download video and extract text from it.

IX. CONCLUSION

In this paper we discussed our proposed method of detecting and extracting text from the video. The system automates the manual process of extracting text from videos and hence is economical in terms of time and human efforts. The system will be implemented in MATLAB language. The system can be mainly used for educational and news videos which contain information in form of text.

X. REFERENCES

[1] Datong Chen, Jean-Marc Odobez. "Text detection and recognition in images and video frames" The Journal of The Pattern Recognition Society, 2004 pages 595-608.
[2] Matti Pietikainen and Oleg Okun. "Text extraction from grey scale page images by simple edge detectors" Machine Vision and Intelligent Systems Group.
[3] C.P.Sumathi, N.Priya "A Combined Edge-Based Text Region Extraction from Document Images" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, August 2013 ISSN: 2277 128X.
[4] Cerenkov, Z Greece Nikou, C. Pitas, I."Shot detection in video sequences using entropy based metrics" Proceedings. International Conference on Volume: 3 2002.
[5] Priti P. Rege Chanchal A. Chandrakar "Text-Image Separation in Document Images Using Boundary Perimeter Detection" ACEEE Int. 1. On Signal & Image Processing, Vol. 03, No. 01, Jan 2012.
[6] Arvind, Mohamed Rafi "Text Extraction from Images Using Connected Component Method" Journal of Artificial Intelligence Research & Advances Volume 1, Issue 2, 2014.
[7] Lifang Gu "Text Detection and Extraction in MPEG Video Sequences" In Proceedings of the International Workshop on Content-Based Multimedia Indexing, 2001 pages 233-240.
[8] Punit Kumar, P. S. Puttaswamy "VIDEO TO FRAME CONVERSION OF TV NEWS VIDEO BY USING MATLAB". IJARSE, Vol. No.3, Issue No.3, March 2014.
[9] Punit Kumar, P. S. Puttaswamy "Moving text line detection and extraction in TV video frames".IEEE International Advance Computing Conference (IACC) 2015.