Text Recognition and Extraction from Videos

[1]Mr. P. Ramkishor, [2]V. Lahari, [3]R.V. Divyesh,, [4]T. Nadiya Kumari, [5]S. Likitha
[1]Assistant Professor, Department of Computer Science and Engineering
[2,3,4,5] Under Graduate, Department of Computer Science and Engineering,
Aditya Institute of Technology and Management, Srikakulam, Andhra Pradesh-India

ABSTRACT:

In recent years, increasing impact of deep learning on Text recognition and extraction from Images and videos has attracted lots of things. In deep learning, extracting text is the primary step which contains more information for analyzing the content, indexing and retrieval of videos. In many works, text features are extracted based on morphological features such as color histograms and aspect ratios. However, under those features, similar text objects are not sufficiently distinguishable to make a distinction between them. To address this issue, we will use deep learning methodologies. The main motto behind this extraction is to recognize the text from videos. In our work, we propose a method for a variety of experiments on variety of datasets to verify that our proposed approach largely improves the performance of text detection and extraction from videos.

Index Terms- Text Detection, Deep Learning, Indexing, Morphological features

I. INTRODUCTION

Reading is fundamental in everyday life. In real world printed content is available wherever as records prefer reports, receipts, proclamations, eatery menus, item bundles, directions, billing and so on. Text information plays a major role in numerous applications for giving a considerable measure of descriptive and dynamic information. Now-a-days major requirements in real world applications like object recognition, facilitative navigation, scene understanding, etc., makes the detection of text to be important task in content-based image analysis.

The existing methods for text extraction were implemented in many ways and one of the best methods is CRM i.e., Corner Response Feature Map [1]. There is a Graph Matching Approach [2] which utilizes the relationship between two objects or trajectories and also Bayesian-based framework of Tracking based Text Detection and Recognition [4] which is composed of three major components, i.e., text tracking, tracking based text detection, and tracking based text recognition These methods are very popular meanwhile they has few disadvantages such as low contrast, low resolution etc. And these don't work well for complex background images and handwritten documents. They can't able to work efficiently on images with low resolution and have multiple orientations of text. Also, the detection accuracy was low with it. The results of these methods were not enough for recognizing the text candidates. Contrast is the difference between the blacks and the whites. Typically, something that has low contrast looks like it has a gray wash to it. The blacks are more like a dark gray, and the whites a light gray. These images with low contrast are unfit for detection of edges of objects or texts within them. Our proposed system deals with some of such issues.

The proposed approach is composed of four steps: video decoding, text detection, Frame conversion and text extraction. Main advantage in our project is it works well for complex backgrounds and it can extract from handwritten documents also. First, the video will be captured using web cam and it will be saved in a directory. Now, the text in the video will be detected using EAST library and then video will

be converted to frames. After converting into Frames, the text in the frames will be extracted using OCR. In this project, Text Detection and Extraction from videos is performed. Firstly, text regions detection is done by few sequential Image Processing steps to the resulting image. Finally the result is fed into an OCR for character recognition and extraction. This project can help out the visually impaired people with new kind of devices fixed with device cameras that can capture the scene images.

II. LITERATURE SURVEY

In 2018 Wei Lu, Hongbo Sun, Jinghui Chu, Xiangdong Huang, and Jiexiao Yu proposed a novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network for character detection. The text presented in videos contains important information for content analysis, indexing, and retrieval of videos. Here their main motto is to extract the information for verifying the text in various languages in different backgrounds. Their approach composed of four steps: Video decoding, text detection, candidate text line localization and false text line elimination. First, they use OpenCV library to decode the video into frames and then use a corner response feature map detector to obtain candidate text regions. Candidate text lines are partitioned through projection analysis on to the contours of candidate text regions. If projection analysis fails, then they proceed with FCM-based separation which is bit complicated. In FCM method they extracted the candidate text layer and convert it into grayscale image. And the final step is to remove false text lines using transferred deep CNN classifiers and true text lines undergo FCM separation and morphological process to obtain binary text. [1].

In 2018 Jianfeng Dong, Xirong Li proposed a method for Predicting Visual Features from Text for Image and Video Caption Retrieval which attacks the problem of image and video caption retrieval, i.e., finding amidst a set of possible sentences the one best describing the content of a given image or video. In this paper they proposed to perform image and video caption retrieval in a visual feature space exclusively. For this they contributed a deep neural network architecture Word2VisualVec which predicts the visual features from the textual input. They further generalized Word2VisualVec for video caption retrieval, by predicting from text both 3-D convolutional neural network features as well as a visual-audio representation. [2].

In 2018 WEI-YI PEI, CHUN YANG proposed Scene Video Text Tracking with Graph Matching. In this paper, they proposed a text tracking method with graph matching. The method performs detecting by tracking multiple text blocks frame by frame using template matching, object prediction, trajectory initialization and trajectory elimination. In general, the adjacent matrices are modelled to represent the extracted tracking object features. Then, often, the Hungarian algorithm is applied to find the Correspondence pairs between consecutive frames.

In many works, text features are extracted based on morphological features such as color histograms and aspect ratios. However, under those features, similar text objects are not sufficiently distinguishable to make a distinction between them. To address this issue, we regard the template matching task as a graph matching problem. [3].

In 2018 Shu Tian, proposed A Unified Framework for Tracking Based Text Detection and Recognition from Web Videos. In this paper, they proposed a generic Bayesian-based framework of

Tracking based Text Detection and Recognition from web videos for embedded captions, which is composed of three major components, i.e., text tracking, tracking based text detection, and tracking based text recognition. There are four major contributions of this paper. The first contribution is a unified Bayesian-based framework for both tracking based text detection and tracking based text recognition from complex (web) videos. The second contribution is a novel tracking-by-detection approach for text tracking, where the appearance model for region matching and the motion model for text tracking are adaptively designed and utilized to link the detections into trajectories, different from conventional methods only focusing on region matching. The third contribution is well-designed tracking based text detection and tracking based text recognition approaches. The fourth contribution is a practical dataset for text detection and recognition from web videos. . A variety of experiments on this dataset verify that their proposed approach largely improves the performance of text detection and recognition from web videos. [4].

In 2016 Xu-Cheng Yin, Ze-Yu Zuo proposed Text Detection, Tracking and Recognition In Video: A Comprehensive Survey. Here, this paper presents a comprehensive survey of text detection, tracking, and recognition in video with three major contributions. First, a generic framework is proposed for video text extraction that uniformly describes detection, tracking, recognition, and their relations and interactions. Second, within this framework, a variety of methods, systems, and evaluation protocols of video text extraction are summarized, compared, and analyzed. Existing text tracking techniques, tracking-based detection and recognition techniques are specially highlighted. Third, related applications, prominent challenges, and future directions for video text extraction (especially from scene videos and web videos) are also thoroughly discussed. [5].

In 2015, Liang Wu, Palaiahnakote Shivakumara et.al. Proposed an A New Technique for Multi-Oriented Scene Text Line Detection and Tracking in Video. Detecting and tracking text from videos is really a challenging task because of its background, contrast, resolution and orientation. It is much more difficult to detect if both caption and scene text collaborate. This paper proposed a new technique for detecting and tracking video texts of any orientation by using spatial and temporal information, respectively. The technique explores gradient directional symmetry at component level for smoothing edge components before text detection. Spatial information is preserved by forming Delaunay triangulation in a novel way at this level, which results in text candidates. Text characteristics are then proposed in a different way for eliminating false text candidates, which results in potential text candidates. Then grouping is proposed for combining potential text candidates regardless of orientation based on the nearest neighbor criterion. To tackle the problems of multi-font and multi-sized texts, we propose multi-scale integration by a pyramid structure, which helps in extracting full text lines. Then, the detected text lines are tracked in video by matching the sub graphs of triangulation. Experimental results for text detection and tracking on our video dataset, the benchmark video datasets. [6]

In 2012, Xiaoqian Liu et. al., proposed Robustly Extracting Captions in Videos Based on Stroke-Like Edges and Spatio-Temporal Analysis. In this paper, they proposed an efficient and effective way to extract the captions from the videos. .First, we propose a novel stroke- like edge detection method based on contours, which can effectively remove the interference of non-stroke edges in complex background so as to make the detection and localization of captions much more accurate. Second, our approach highlights the importance of temporal feature, i.e., inter-frame feature, in the task of caption extraction (detection, localization, segmentation). Instead of regarding each video frame as an

independent image, through fully utilizing the temporal feature of video together with spatial analysis in the computation of caption localization, segmentation and post-processing, we demonstrate that the use of inter- frame information can effectively improve the accuracy of caption localization and caption segmentation [7].

In 2011 Xu Zhao, Kai-Hsiang Lin et. al., proposed Text from Corners: A Novel Approach to Detect Text and Caption in Videos. In this paper, they presented a corner based approach to detect text and caption from videos. This approach is inspired by the observation that there exist dense and orderly presences of corner points in characters, especially in text and caption. They used several discriminative features to describe the text regions formed by the corner points. The usage of these features is in a flexible manner, thus, can be adapted to different applications. Language Independence is an important advantage of the proposed method. Moreover, based upon the text features, we further develop an ovel algorithm to detect moving captions in videos. In the algorithm, the motion features, extracted by optical flow, are combined with text features to detect the moving caption patterns. The decision tree is adopted to learn the classification criteria. Experiments conducted on a large volume of real video shots demonstrate the efficiency and robustness of our proposed approaches and the real-world system. [8]

III. METHODOLOGY

In this project, we address the problem of recognizing text in images with low resolution, complex background and having text in multiple orientations.

Our proposed approach is composed of four steps: video decoding, text detection, Frame conversion and text extraction. Main advantage in our project is it works well for complex backgrounds and it can extract from handwritten documents also. First, the video will be captured using web cam and it will be saved in a directory. Now, the text in the video will be detected using EAST library and then video will be converted to frames. After converting into Frames, the text in the frames will be extracted using OCR. In this project, Text Detection and Extraction from videos is performed. Firstly, text regions detection is done by few sequential Image Processing steps to the resulting image. Finally the result is fed into an OCR for character recognition and extraction. This project can help out the visually impaired people with new kind of devices fixed with device cameras that can capture the scene images.

IV. IMPLEMENTATION

Our work mainly composed of four phases i.e., Video Capturing, Text Detection, Splitting the Video and Text Extraction. The system can take a real time video or a video from stored database as input. Each stage is explained by presenting experimental results for a set of static images.

Step 1: The first step is the video should be captured using web cam and it should be saved in a specific directory or the video which was already existed can also be given as an input.
Step 2: After the video was captured the text in the video will be detected using EAST library. EAST comes both as a CLI application and as a python library (which can be imported and used

in python).

Step 3: To extract the text from the videos, it should be in either JPG or PNG format. So, we have to split the video into frames and we should select the best frame among them. The selected frame should be given to pre-processing to eliminate the false values and to separate the text region from the background.

Step 4: The final step is the text will be extracted from the videos with an accuracy of 60-80%. Accuracy of text depends on the clarity of image and background style of images.

Here, text extraction from images is done mainly in 2

- stages: Pre-Processing
- Segmentation

Tesseract OCR

After Pre-Processing and Segmentation, text should be extracted and it will be done using Pytesseract. Tesseract is an optical character recognition (OCR) system. It is used to convert image documents into editable/searchable PDF or Word documents. It is free, open source software run through a Command-Line Interface (CLI). Tesseract is considered one of the most accurate open source OCR engines currently available and its development has been sponsored by Google since 2006.That being said, its capabilities can be more limited than commercial software like Adobe Acrobat Pro and ABBYY Fine Reader. However, because it is open source software, anyone with programming knowledge can edit the code behind Tesseract and help it learn what you need to do. It can be used on Mac, Windows and Linux machines. Running it through command prompt. Basic OCR Operations in Tesseract:

- Image format (JPG, TIF, PNG, etc.) to PDF, Microsoft Word
- New document appears in the same directory as initial document
- Run through your Command-Line Interface.

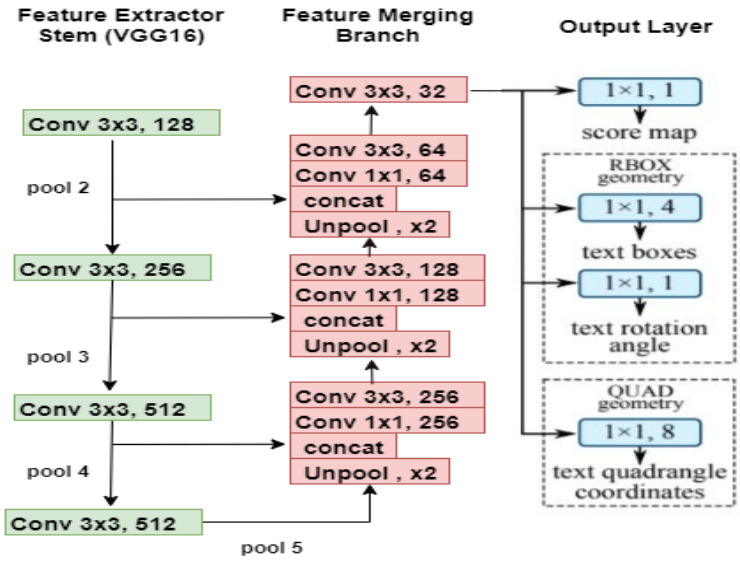
V. ALGORITHMS USED

EAST (Efficient and Accurate Scene Text)

The EAST algorithm is used to detect text. This algorithm uses a single neural network to predict a word or line-level text. It can detect text in arbitrary orientation with quadrilateral shapes. In 2017 this algorithm outperformed state of the art methods. This algorithm consists of a fully convolutional network with a non-max suppression (NMS) merging state. The fully convolution network is used to localize text in the image and this NMS stage is basically used to merge many imprecise detected text boxes into a single bounding box for every text region (word or line text).

The EAST architecture was created while taking different sizes of word regions into account. The idea was to detect large word regions that require features from the later stage of the neural network

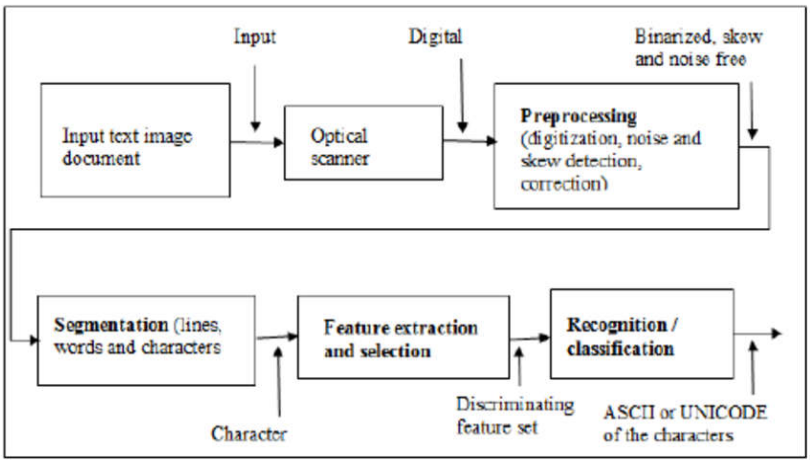
while detecting small word regions that require low-level features from initial stages. To create this network, authors have used three branches combining into a single neural network.



EAST

OCR (Optical Character Recognition)

Optical character recognition (OCR) algorithms allow computers to analyze printed or handwritten documents automatically and prepare text data into editable formats for computers to efficiently process them. Human eyes naturally recognize various patterns, fonts or styles. For computers, it is hard work to do. Any scanned document is a graphics file, i.e., a pattern of pixels. A computer localizes, detects and recognizes characters on an image and turns the image of paper documents into a text file.



OCR

VI. RESULTS

These are the output screens of the project.

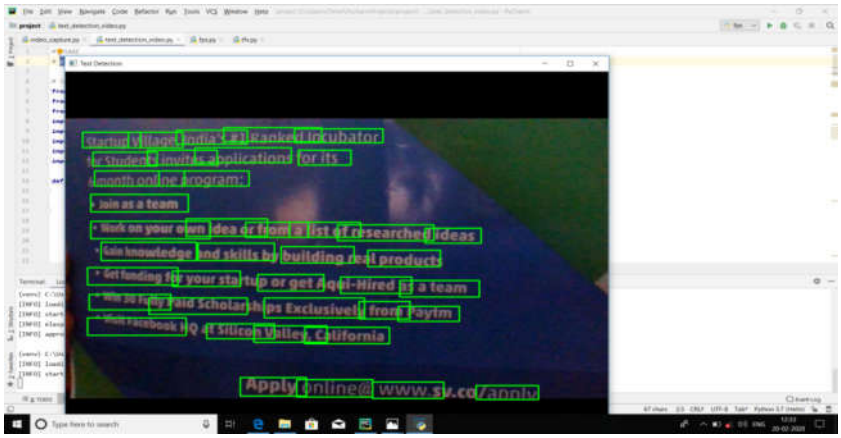


Figure 1 Text Detection from Video

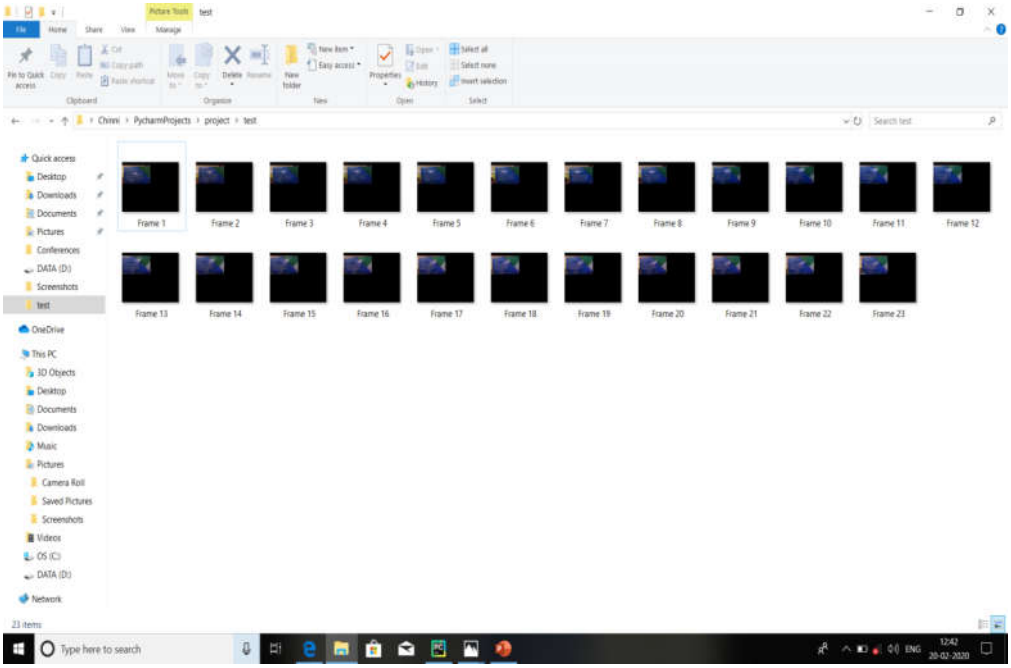


Figure 2 Decoding into frames

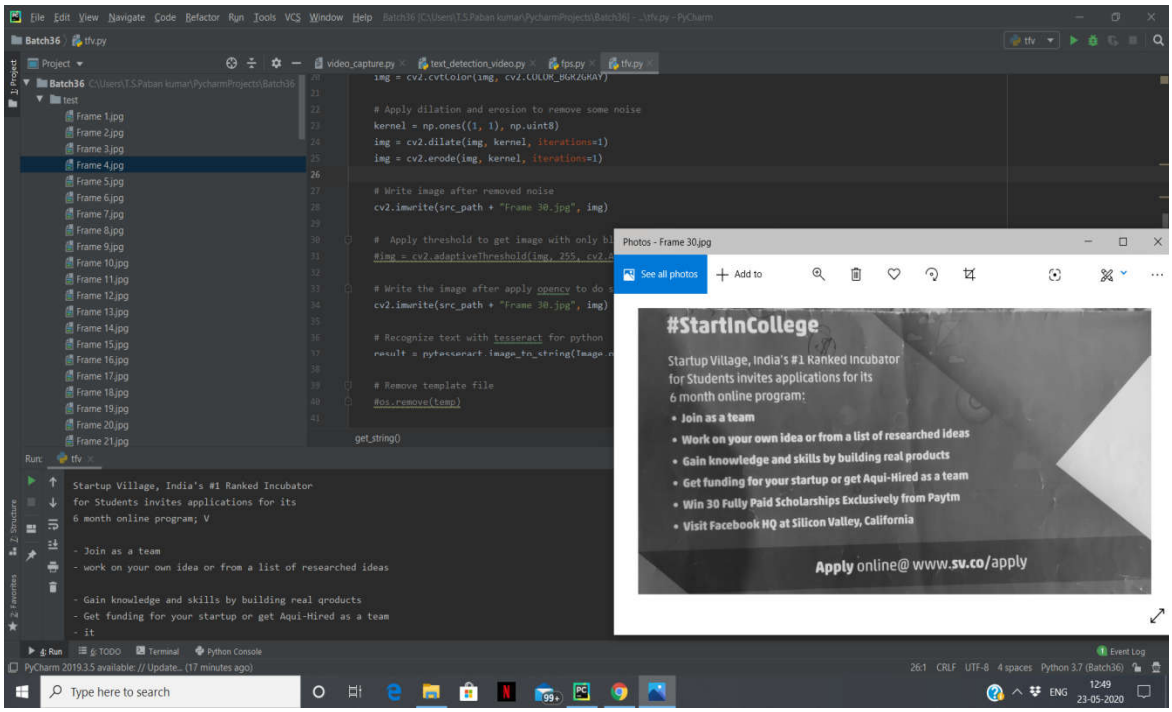


Figure 3 Text Extraction from video

CONCLUSION

Text detection, localization, and extraction are often used interchangeably in the literature. We can conclude that the text appearing in a video can be extracted using this software with the only restriction i.e. the video should not be much distorted. As, the characters are recognised on run-time basis, there may be a few cases in which one or two characters may get misrecognised i.e. a character may get recognised as some other character. But from the experiments performed on the software, most of the text gets recognized successfully.

FUTURE SCOPE

In extension to this, the work can be further extended to speech from text. Whatever the text we extracted from the videos will be converted to an audio file so that it can be helpful for the blind people and we can also convert the audio to any language. And it can also be extended to extract the text from complex backgrounds with max accuracy.

- Load the video
- Divide it into individual frames
- Detect the text from each of those individual frames
- Implement an algorithm to each frame to convert the text to audio clip.
- Implement the algorithm by eliminating external disturbances and some false values to extract the maximum accuracy.

REFERENCES:

*[1] Wei Lu, Hongbo Sun, Jinghui Chu, Xiangdong Huang, and Jiexiao Yu "**A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network**", IEEE Access, Vol. 06, no.1, pp. 40198-40211, August 2018.*

*[2] Jianfeng Dong, Xirong Li, and Cees G. M. Snoek **"Predicting Visual Features from Text for Image and Video Caption Retrieval",** IEEE Transactions on Multimedia, Vol. 20, no.12, pp.3377-3388, December 2018.*

*[3] WEI-YI PEI, CHUN YANG, LI-YU MENG, JIE-BO HOU, SHU TIAN1, and XU- CHENG YIN "**Scene Video Text Tracking With Graph Matching**", IEEE Access, Vol. 06, pp. 19419-19426, April 2018.*

*[4] Shu Tian, Xu-Cheng Yin, Ya Su, and Hong-Wei Hao "**A Unified Framework for Tracking Based Text Detection and Recognition from Web Videos**", IEEE Transctions on Pattern Analysis and Machine Intelligence, Vol. 40, no.3, pp.542-554, March 2018.*

*[5] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu "**Text Detection, Tracking and Recognition in Video: A Comprehensive Survey**" , IEEE Transactions on Image Processing, Vol. 25, no.06, pp.2752-2773, June 2018.*

*[6] Liang Wu, Palaiahnakote Shivakumara, Tong Lu, and ChewLim Tan **"A New Technique for Multi-Oriented Scene Text LineDetectionandTrackinginVideo"**, IEEE Transctions on Multimedia, Vol. 17, no.8, pp.1137-1152, August 2015.*

*[7] Xiaoqian Liu, and Weiqiang Wang **"Robustly Extracting Captions in Videos Based on Stroke-LikeEdgesandSpatio-TemporalAnalysis"**, IEEE Transctions on Multimedia, Vol. 14, no.2, pp.482- 489, April 2012.*

*[8] Xu Zhao, Kai-Hsiang Lin, Yun Fu, Member, IEEE, Yuxiao Hu, Member, IEEE, Yuncai Liu, and Thomas S. Huang, **"Text From Corners: A Novel Approach to Detect Text and Caption in Videos"** IEEE Transactions on Image Processing, Vol. 20, no.03, pp.790-799, March 2011.*

*[9]Wen Wu, Xilin Chen, and Jie Yang, "**Detection of Text on Road Signs From Video**", IEEE Transactions on Intelligent Transportation Systems, Vol. 06, no.04, pp. 378-390, December 2005.*