IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Special Issue 2, April 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

Enhanced Extraction of Textual Characters from Multimedia using NLP

Pathi Sai Jahnavi [1], Ramya.V.S [1], Mr.Dhakshunhaamoorthiy [2]

U.G. Student, Department of Computer Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India[1]

Assistant Professor, Department of Computer Engineering, Velammal Engineering College, Chennai, Tamil Nadu, India[2]

ABSTRACT: Enhanced extraction of textual characters represents a comprehensive framework for the detection and recognition of textual content in video frames as well as from images.Detection of text from document images enables Natural Language Processing algorithms to decipher the text and make sense of what the document conveys.Furthermore, the text can be easily translated into multiple languages and summarized the text into meaningful phrases using the transformers and pipelines in NLP text summarization. Then converts the summarized text to audio using the Text-To-Speech library.NLP also recognizes text from supermarket product names, traffic signs, and even from billboards, making them effective translators and interpreters.The development of this application will be a great help to people with visual impairment and those who don't know another language

KEYWORDS: Text Detection, Machine Learning, UnSupervised Learning, easyocr, matplotlib, spacy, numpy

I. INTRODUCTION

In the recent years, there has been a tremendous increase in the amount of digital multimedia data, especially the video content, both in the form of video archives and live streams. According to statistics, 300 h of video is being uploaded every minute on YouTube. A key factor responsible for this enormous increase is the availability of low-cost smart phones equipped with cameras. With such huge collections of data, there is a need to have efficient as well as effective retrieval techniques allowing users retrieve the desired content. The term content may refer to the visual content (for example, objects or persons in the video), audio content (the spoken keywords for instance), or the textual content (News tickers, anchor names, score cards, etc.).
The textual content in video can be categorized into two broad classes, scene text, and caption text. Examples of scene text include advertisement banners, sign boards, and text on a T-shirt. Scene text is commonly employed for applications like robot navigation and assistance systems for the visually impaired. The key components of a textual content based indexing and retrieval system include detection of text regions, extraction of text (segmentation from background), identification of script (for multi-script videos), and finally recognition of text (video OCR).

Detection of text can be carried out using unsupervised, supervised, or hybrid approaches. Unsupervised text detection relies on image analysis techniques to discriminate between text and non-text regions. Supervised methods, on the other hand, involve training a learning algorithm with examples of text and non-text regions to discriminate between the two. In some cases, a combination of the two techniques is also employed where the candidate text regions identified by unsupervised methods are validated through a supervised approach.
This paper presents a comprehensive framework for video text detection and recognition in a multi-script environment. The key highlights of this study are outlined in the following.
• Development of a comprehensive dataset of video frames with ground truth information supporting evaluation of detection and recognition tasks.
• Investigation of state-of-the-art deep learning CRAFT Algorithm based object detectors,
fine-tuning them to detection of textual content.
• Validation of proposed techniques using a comprehensive series of experiments.

This paper is organized as follows. In the next section, we present the database developed in our study along with the ground truth information. In section 3 , we present an overview of the current state-of-the-art from the view point of text detection and Recognition, Summarization , Translation and Speech Representation. While Section 4 defines the realized results and the corresponding discussion. Finally, Section 5 concludes the paper and future work.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

II. RELATED WORK

Li et al. [9] proposed a method for video text tracking based on wavelet and moment features. This method takes advantage of wavelet decomposition, spatial information provided by the moments and with a neural network classifier for identifying text candidates. Huang et al. [4] proposed a method for scrolling text detection in video using temporal frames. This method uses motion vector estimation for detecting text. However, this method is limited to only scrolling text but not arbitrarily-oriented texts. Zhou et al. [38] exploited edge information and geometrical constraints to form a coarse-to-fine methodology to define text regions. However, the method is unable to deal with dynamic text objects. Mi et al. [15] proposed a text extraction approach based on multiple frames. The edge features are explored with a similarity measure for identifying text candidates. Wang et al. [31] used a spatio-temporal wavelet transform to extract text objects in video documents. In the edge detection stage, a three-dimensional wavelet transform with one scaling function and seven wavelet functions is applied on a sequence of video frames. Multimedia Tools Appl Huang [3] detected video scene text based on the video's temporal redundancy. Video scene texts in consecutive frames have arbitrary motion due to camera or object movement. Therefore, this method performs the motion detection in 30 consecutive frames to synthesize a motion image. Zhao et al. [37] proposed an approach for text detection using corners in video. This method proposes to use dense corners for identifying text candidates. Finally, optical flow has been used for moving text detection. Liu et al. [12] proposed a method for video caption text detection using stroke-like edges and spatio-temporal information. The color histogram is used for segmenting text information. Li et al. [10] proposed a method for video text detection using multiple frame integration. This method uses edge information to extract text candidates. Morphological operations and heuristic rules are proposed to extract final text information from video. Recently, Khare et al. [8] proposed method for multi-oriented text detection in video, which uses temporal information. The main objective of the method is to introduce a new descriptor called the Histogram Oriented Moments Descriptor (HOM), which works as Histogram Oriented Gradients (HOG) for text detection in video but does not exploit the temporal coherency for text candidate detection. This is because the method uses optical flow for the text candidates to find moving text with the help of temporal information, but not text candidate detection. In other words, the method considers a frame as an individual image to find text candidates and then it uses temporal information for tracking moving text candidates in temporal frames. Therefore, the performance of the method degrades for static foregrounds and dynamic backgrounds because moving text candidates may overlap with the moving background. Similarly, Wu et al. [32] proposed a new technique for multi-oriented scene text line detection and tracking in video. The method explores gradient directional symmetry and spatial proximity between the text components for identifying text candidates. The text lines are tracked by matching sub-graphs of text components. The ability of the method is not tested on arbitrarily-oriented text and multi-script text lines in video. It is observed from the above literature review that most of the methods focus either on horizontal or non-horizontal text and very few on arbitrarily-oriented text detection in video. Hence, the detection of multi-script text lines in video is not sufficiently addressed in the literature. However, Liu et al. [14] proposed a method for multi-lingual text extraction from scene images using a Gaussian mixture model and learning. This method uses binarization for detecting text lines through connected component analysis. This idea works for high contrast images but not low contrast images such as video frames. Huang et al. [5] also makes an attempt to solve text detection multi-scripts by exploring temporal information based on a motion perception field in consecutive frames. The methods are limited to a few languages, such as Chinese, English and Japanese.

III. METHODOLOGY

We organize our discussion in four main sections. We first discuss the text detection problem followed by an overview of summarization techniques and followed by Translation techniques. At the end, we present the current state-of-the-art and open challenges from the view point of speech representation.

Text Detection and Recognition :

The first step in the proposed framework is the detection of candidate text regions from the extracted video frames. For detection of textual content in a given frame, we have employed state-of-the-art CRAFT Algorithm - based object detectors. Although, many object detectors are trained with thousands of class examples and provide high accuracy in detection and recognition of different objects, these object detectors cannot be directly applied to identify text regions in images.

OpenCV:
OpenCV is an open-source computer vision library written in C/C++. It is mainly focused on image processing

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

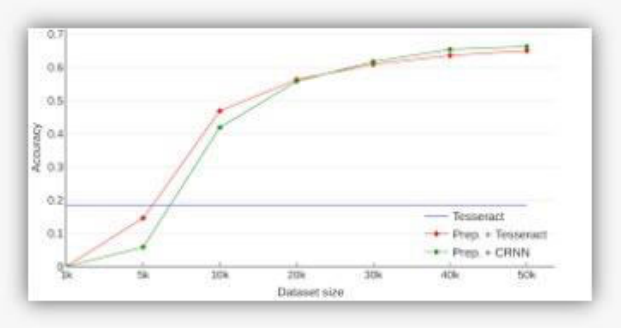
8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

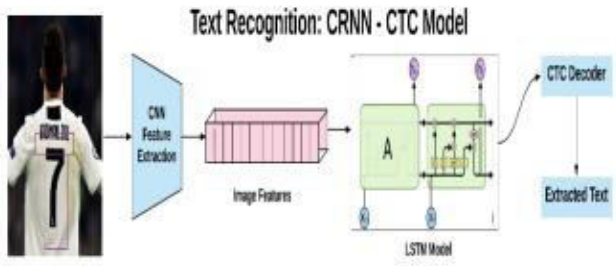
Department of CSE, Velammal Engineering College, Chennnai, India

Easy OCR:
Easy OCR is built with Python and Pytorch deep learning library, having a GPU could speed up the whole process of detection. The detection part is using the CRAFT algorithm and the Recognition model is CRNN.

CRAFT ALGORITHM :
Many deep learning models in computer vision have been used for the scene text detection task over the past few years. However, the performance of these models is not up to the mark when the text in the image is skewed or curved. The CRAFT(Computerized Relative Allocation of Facilities Technique) model performs well on even curved, long and deformed texts in addition to normal text. The CRAFT text detection model uses a convolutional neural network model to calculate region scores and affinity scores.



CRNN ALGORITHM :
After the text detection step, regions, where the text is present, are cropped and sent through convolutional layers to get the features from the image. Later these features are fed to many-to-many LSTM architecture, which outputs soft max probabilities over the vocabulary. These outputs from different time steps are fed to the CTC decoder to finally get the raw text from images.



The final fine-tuned model takes a video frame as input and localizes the text regions. Once the text is localized, the detected regions are fed to the summarization module to summarize the script of the detected text.

Summarization Techniques :

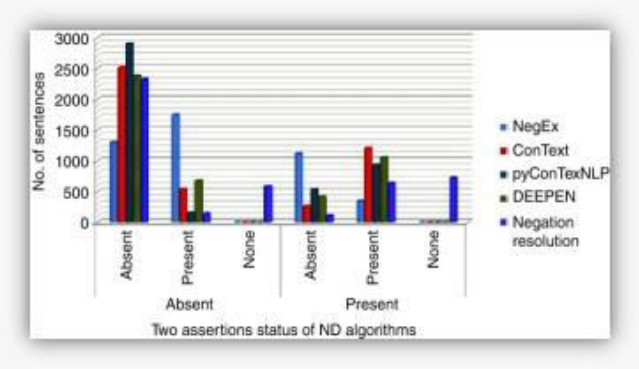
Text summarization is a natural language processing (NLP) task that allows users to summarize large amounts of text for quick consumption without losing any important information.
Spacy :
Spacy pipeline object for negating concepts in text based on the Neg Ex algorithm.
NegEx ALGORITHM :
The NegEx algorithm relies on four types of lexical cues: negation triggers that indicate a negation (e.g., “denies”), pseudo-negation triggers that contain negation triggers but do not negate the clinical condition (e.g., “no increase”), and termination terms that stop the scope of the negation trigger (e.g., “but”)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India



Our approach is to simply:
1.Look at the use frequency of specific words
2.Sum the frequencies within each sentence
3.Rank the sentences based on this sum

Translation Techniques :

Often we are faced with a problem where either we cannot understand a piece of the critical text, or we need to provide information to speakers of another language.

Googletrans:

Since then, Google Translate began using neural machine translation (NMT) in preference to its previous statistical methods (SMT) which had been used since October 2007, with its proprietary, in-house SMT technology. Google Translate's NMT system uses a large artificial neural network capable of deep learning.

GNMT ALGORITHM :

With the power of deep learning, Neural Machine Translation (NMT) has arisen as the most powerful algorithm to perform this task. While Google Translate is the leading industry example of NMT, tech companies all over the globe are going all in on NMT

Speech Representation :

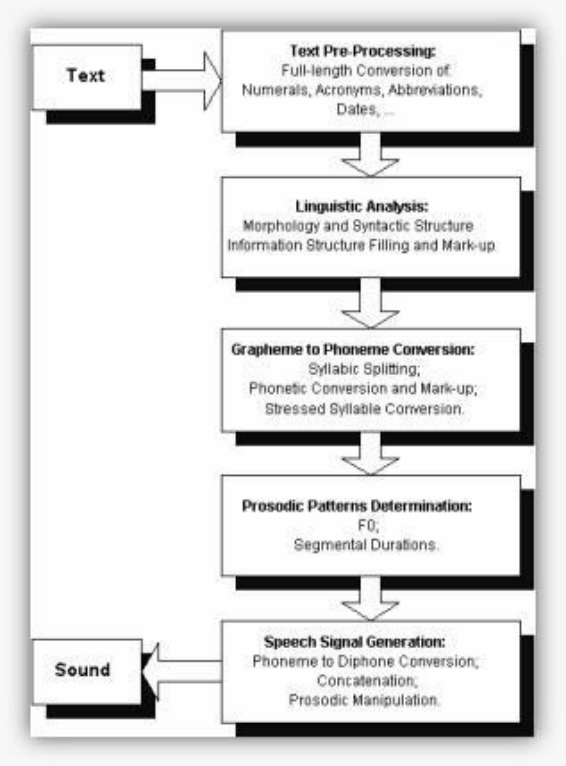
The TTS system gets the text as the input and then a computer algorithm which called TTS engine analyses the text, pre-processes the text and synthesizes the speech with some mathematical models. The TTS engine usually generates sound data in an audio format as the output.

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

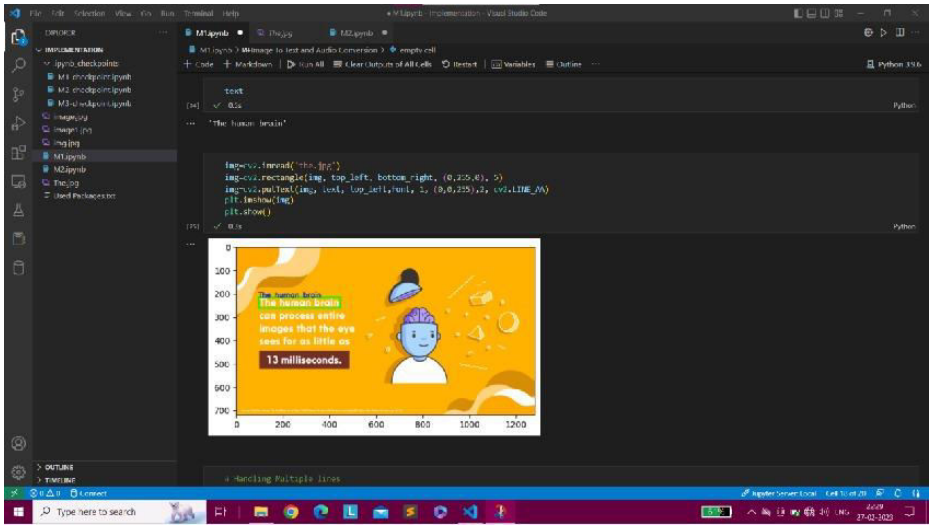
8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India



IV. EXPERIMENTAL RESULTS



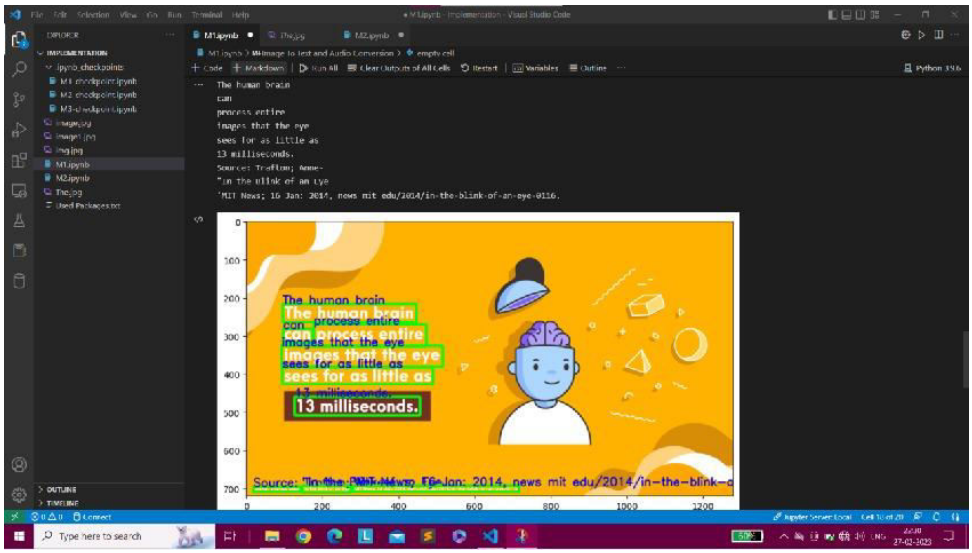
fig(a)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

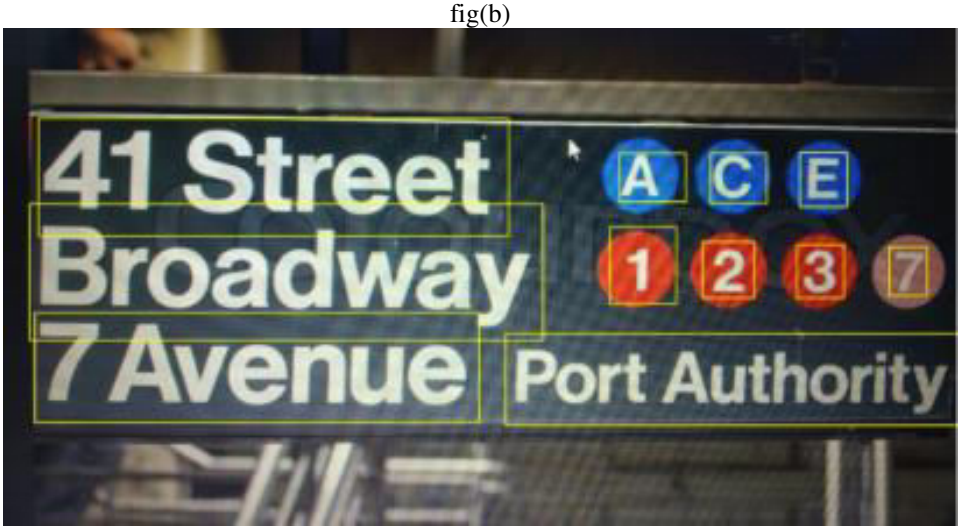
8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India



fig(b)



fig(c)

fig(a). Handling single line , fig(b).Handling Many lines ,fig(c).Expected output in detection

V. CONCLUSION

By this project, we have investigated methods in building an efficient way of detecting and recognizing text in video sequences.We are implementing Language translation while reading the image which is faster than translating the extracted text. Implementation of summarising the text is helpful in understand the whole context of the text or paragraph. The development of this application which will be great help to people with visual impairment and those who don't know other language.

REFERENCES

[1] Jamil, A. Batool, Z. Malik, A. Mirza, I. Siddiqi, Multilingual artificial text extraction and script identification from video images. Int. J. Adv. Comput. Sci. Appl. 1(7), 529–539 (2016)

[2] J. Hochberg, L. Kerns, P. Kelly, T. Thomas, in Document Analysis and Recognition, 1995., Proceedings of the Third International Conference On, vol. 1. Automatic script identification from images using cluster-based templates (IEEE, Montreal, 1995), pp. 378–381

[3] A. L. Spitz, Determination of the script and language content of document images. IEEE Trans. Patt. Anal. Mach. Intell. 19(3), 235–245 (1997)

|| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | SJIF: 8.423|| Volume 12, Special Issue 2, April 2023 ||

8th National Conference on Advanced Computing Technologies (NCACT'23)

Organized by

Department of CSE, Velammal Engineering College, Chennnai, India

[4] U. Pal, B. Chaudhuri, in Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference On. Automatic identification of English, Chinese, Arabic, Devnagari and Bangla script line (IEEE, Washington, 2001), pp. 790–794

[5] Z. Malik, A. Mirza, A. Bennour, I. Siddiqi, C. Djeddi, in 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). Video script identification using a combination of textural features (IEEE, Bangkok, 2015), pp. 61–67

[6] C. Zhu, W. Wang, Q. Ning, in Advances in Multimedia Information Processing-PCM 2006. Text detection in images using texture feature from strokes (Springer, Hangzhou, 2006), pp. 295–301

[7] Z. Li, G. Liu, X. Qian, D. Guo, H. Jiang, Effective and efficient video text extraction using key text points. IET Image Process. 5(8), 671–683 (2011)

[8] N. Sharma, S. Chanda, U. Pal, M. Blumenstein, in Document Analysis and Recognition (ICDAR), 2013 12th International Conference On. Word-wise script identification from video frames (IEEE, Washington, 2013), pp. 867–871

[9] G. Peake, T. Tan, in Document Image Analysis, 1997.(DIA'97) Proceedings., Workshop On. Script and language identification from document images (IEEE, San Juan, 1997), pp. 10–17

[10] P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein, C. L. Tan, in Pattern Recognition (ICPR), 2014 22nd International Conference On. Gradient-angular-features for word-wise video script identification (IEEE, Stockholm, 2014), pp. 3098–3103

[11] N. Sharma, U. Pal, M. Blumenstein, in Neural Networks (IJCNN), 2014 International Joint Conference On. A study on word-level multi-script identification from video frames (IEEE, Beijing, 2014), pp. 1827–1833

[12] N. Sharma, R. Mandal, R. Sharma, U. Pal, M. Blumenstein, in Document Analysis and Recognition (ICDAR), 2015 13th International Conference On. Icdar2015 competition on video script identification (CVSI 2015) (IEEE, Nancy, 2015), pp. 1196–1200

[13] J. Mei, L. Dai, B. Shi, X. Bai, in 2016 23rd International Conference on Pattern Recognition (ICPR). Scene text script identification with convolutional recurrent neural networks (IEEE, Cancún, 2016), pp. 4053–4058

INNO SPACE
SJIF Scientific Journal Impact Factor

SJIF: 8.423

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INTERNATIONAL CENTRE

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details