A Review on Natural Scene Text Understanding for Computer Vision using Machine Learning

Mr. T. Gnana Prakash[1], B. Tejaswini[2], Ch. Varshini[3], Ch. Harathi[4], G. Bhavani[5]

[1]Assistant Professor, Dept. of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
[2,3,4,5]Student, Dept. of Computer Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

---***---

Abstract - *Textual depictions of landscapes and other natural settings give us insightful information about the location and can reveal some crucial details. Reading characters in unconstrained scene images is a challenging problem of considerable practical interest. This problem can be observed in medical fields. Scene text identification in a general condition is still a very open and difficult research subject, despite the fact that various text identification techniques have been established. Usually, the characters in one text are of similar size and font in simple images. However, there are texts in different scene images which can vary in size or fonts. So, End to End Scene Text Understanding using computer vision is applicable to detect text from complex images with complex backgrounds and fonts which will give accurate results. The text we extracted from the images can be stored in a document and we will be able to refer to the data in the future.*

***Key Words*: Image processing, Text recognition, machine learning, Convolutional neural network.**

1.INTRODUCTION

Text in images which are naturally scene gives more information about the scene. Detection of text is useful in many areas like self-driving cars, Hospitals, product recognition, and for visually impaired people who cannot read text in images from far. As far now, different methodologies have been introduced for scene text detection from natural images. But text recognition from natural scene images with complex backgrounds, variations in fonts, highly illuminated, and different layouts is still a challenging task in computer vision.

Fig -1

Fig-1 depicts these text graphics. They contain a complicated background, uneven lighting, poor contrast, blurred typefaces, font distortion, and a variety of colors. Traditional OCR methods are ineffective in recognition. As a result, there is focus of current research in the field is the investigation of scene text recognition techniques.

Deep learning techniques are more accurate and produce better results than image processing techniques, according to a survey on text recognition from natural photos [1]. For text detection from photos of natural scenes, a variety of deep learning-based techniques are employed. Examples include recurrent neural networks (RNN), feature pyramid networks (FPN), and convolutional neural networks (CNN). For the detection and extraction of text from photos of natural scenes, image processing techniques such as image binarization, morphological operations like erosion and dilation, and MSER (Maximally Stable Extremal Regions) detector are utilized.

And after text recognition, converting the test into speech takes place by reading alphabets that are present in the image using some libraries and changing it to voices. This conversion has a social impact in helping the visually impaired people to understand the text in scene images.

2. LITERATURE SURVEY

Canjie Luo et al., [2] introduced Moran (Multi-Object Rectified Attention Network), and different CNN methods are applied. It consists of two parts. One is MORN (Multi-Object Rectification Network) which is used to obtain rectified images by applying various sampling methods according to the predicted offsets of each part of the input image. The results show that a rectified image is easier for the purpose of further recognition. Second part is ASRN (Attention based Sequence Recognition Network) where rectified image is thresholder using a CNN-BLSTM framework containing a decoder in it to extract the text information.

Xinyu Zhou et al., [3] suggested a pipeline, EAST (Efficient and Accurate Scene Text Detector), which utilizes FCN for generating per-pixel predictions on text information by avoiding intermediate steps like candidate proposal, character partition etc.The maximum number of text instances the detector can process is inversely proportional to the network's receptive field. The generated predictions are thresholded and delivered to NMS to obtain the pipeline's final output.

Ruijie Yan et al,. [4] went through a process where they retrieved geometric structures as elements from local level to global level in the feature maps from a text image. Here they focused on specific features and encoded using MEA. They used CNN for feature extraction, MEA method for encoding and decoder for final text output.

Weilin Huang et al.,[5] have proposed a model which has four main steps: component detection, components filtering, text-line constructing, and text-line filtering. The connected component method separates textual and non-textual information at the pixel level. Stroke width transform (SWT) is a connected component method used for component detection. SWT provides stroke width transform as output. SWT will consider inter-component connections due to which it makes errors while connecting multiple characters, to overcome this problem stroke feature transform (SFT) is used. SFT provides a stroke width map and width color as output. SFT mitigates the inter-component connections and enhances the intra-component connections. Text covariance Descriptors (TCD) are used for filtering. the model used two types of TCDs for filtering TCD-C (TCD for components) and TCD-T (TCD for text line).

Pengwen Dai et al., [6] adapted an attention network that is able to recognize text with different scale orientations. In this method, first the network they adapted will rotate the text according to their scales then dynamic scaling is done. After that, encoder-decoders do the task of encoding the text and decoding that text into readable form.

Asghar Ali Chandio et al., [7] constructed a model that, without pre-segmenting the input into individual letters, modifies the order of the relevant properties from a whole word picture. The three main components of this model are the deep convolutional neural network (CNN) with shortcut connections used for feature extraction and encoding, the recurrent neural network (RNN) used for feature decoding of the convolutional features, and the connectionist temporal classification (CTC) used to map the predicted sequences into the target labels.

Hamam Mokayed et al.,[8] chose a method which has two modules that feature extraction and defects component detection. In the feature extraction in order to separate character components from the detected text binarization approach is used. To detect defect components the segmented components are given as input to the SVM classifier to generate the confidence score. Calculate the weight-based Gaussian distribution based on pixel values then the weights are multiplied with the extracted features finally two clusters are given as a result text component cluster and the defect component cluster.

Mandana Fasounaki et al., [9] by combining several image processing techniques obtained an approach where initially, MSER features are applied in order to identify ROI and then some geometric elimination and SWT elimination takes place and followed by the connection of characters where non-character regions are eliminated without OCR assistance. The results show that this model can effectively give promised results by combining various text detection methods.

Wenhao He et al., [10] Convolutional feature extraction is the first step in the proposed method, which then combines multi-level feature fusion and multi-level learning in the network part to detect text from complex images. A separate module performs pre-processing to demand a word-level text. The most effective quadrilateral boundary regression method is direct regression. These tactics have achieved cutting-edge performance and have been proven effective in experiments.

Jeff Donahue et al.,[11] gone through an algorithm, which utilizes neural network model and is fully differentiable and feed forward design where the generator contains the aligner which is used align the random input to aligned input, and the decoder which is used to convert the aligned input to the audio as output.

Randheer Bagi et al., [12] A portable scene text detector that can deal with scene photos' crowded surroundings has been proposed. It is a fully trainable deep neural network which leverages contextual cues from oriented area suggestions, global structural traits, and local component information to identify text occurrences.

Angia Venkatesan Karpagam et al.,[13] suggested a methodology to differentiate the background with text information by creating a method to eliminate the background and extract the semantic information from the natural image. Firstly, when the image is fed, Color space conversion and reduction takes place. Renyi entropy-based thresholding and edge map generation is done where Renyi entropy acts as a good background filter where edge pixels are considered for determining the text from thresholder image. And the next step is removing background noise and then location of the text area and extracting text from it. The results show that they work well for only natural images where background color and foreground color are completely contrast to each other.

R. Saini et al.,[14] Proposed a method to detect anomalies in detection results generated by various text detection methods. The model works in two stages segmenting components and anomaly detection. In the segmenting components, saliency maps and rough set theory are used. And in the second stage anomaly detection fuzzy logic is used as classifier by giving extracted features as the input.

Zihao Liu et al., [15] focused on the core part of Faster R-CNN to detect text proposal regions is Region Proposal Network (RPN). And the second module is the text line constructor Non-Maximum Suppression (NMS) is used to choose the most appropriate delimitation area for the object. To make the predictions more accurate and reliable very deep VGG 16 network is used to extract deep features of natural scene images.

Chae Young Lee, Youngmin Baek, Hwalsuk Lee,[16]Text detector evaluation (TedEval) evaluates the detected text via the matching policy and scoring policy. The model is compared with two other models (IoU, and DetEval) and their limitations granularity and incompleteness are resolved by TedEval. The future work for this model is to create a polygon around the curved text.

Youngmin Baek et al.,[17] proposed a procedure where character-level images are taken out from the original image, the trained neural network model predicts the character region (region score), then the watershed algorithm is applied to get the character bounding boxes, the coordinates of character bounding boxes are combined and then transformed to the original image. To deal with curved or aligned text polygon is generated around the entire text.

Ranjith.P et al., [18]have suggested a procedure that can be broken down into four main steps: input image, image processing methods to localize the text, information extraction from the obtained text, and categorization of them into useful groups.

Liu, Juhua, et al. [19] introduced a semi-supervised scene text recognition system (SemiText), which uses unannotated data and a model which is pre-trained supervised to train reliable and accurate scene text detectors.

Nagdewani, Shivangi, and Ashika Jain. [20] after researching on different methodologies for converting Speech-to-Text and Text-to-Speech , recommended that THE HIDDEN MARKOV MODEL works best for both STT and TTS conversions.

Zuo, Ling-Qun, et al[21]In this method, a CNN network is trained to output a text sequence when an input text picture is presented.A feature decoder is then applied to the created sequences in order to distinguish the significant features in the sequence's form.The text sequence is produced using Bi-LSTM after feature coder and CTC have been applied.When compared to other approaches, this strategy has a high accuracy and performance level.

Q. Wu, P. Chen and Y. Zhou, [22] In this Paper, there is a study of entire synthetic system overflow.In the first step,original text image as input is fed and then,depth map estimation is done ,where depth image will be produced and in this step,estimation of depth and text segmentation takes place.First, the contour of the picture region is changed to a frontal parallel perspective using the computed plane normal; The fronto parallel area is then placed with a suitable rectangle; Last but not least, align the text such that it is the width (that is, the longer side) of the rectangle.and displaying of text takes place.

Pise, Amruta, and S. D. Ruikar [23] In this study, text region detector is created utilizing the histogram of directed gradients, a popular feature descriptor (HOG). Because they are less sensitive to visual noise and can capture the properties of text regions, histograms of oriented gradients (HOG) are a good choice for text detection challenges. Components that are related are segmented using local binarization. In order to separate text from non-text components during text extraction, factors like normalized height width ratio and compactness are taken into consideration. A distance metric feature extraction approach based on zone centroid and picture centroid is used to accomplish text recognition.

G. Vaidya, K. Vaidya and K. Bhosale, [24] Text Recognition System for Visually Impaired using Portable Camera.The three essential components of this system are audio output, text detection, and image capture. Scenes containing objects of interest are collected by the image capture component. Webcam is used here for taking pictures. Text detection software converts the text area from photos into usable codes. In this setup, the processing device is a laptop. For blind users, the audio output component turns the detected text into speech.

Karanje, Uma B., and Rahul Dagade.[25] Mentioned various MSER approaches along with their benefits and drawbacks. To extract text from photos, MSER++ is employed. It generates several MSERs and employs exhaustive searches for trimming. A two-stage approach that uses learned classifiers and tuning parameters is used to quickly extract text from photos. MSER is employed as a candidate for characters to detect low-quality texts, low-contrast characters, and texts with loud noises. To eliminate the non-text MSERs from the image, the graph cut technique is combined with MSER.

F. Wang et al.,[26] created an end-to-end ITN, Instance Transformation Network, to detect text from natural scene images which are in the form of complicated geometric layouts.It encodes the unique geometric configurations of each text instances from the scene.It works for texts lines or words with multi-scale, multilingual and multi-oriented form within one pass.

Cunzhao Shi et al.,[27] have proposed a method which has 3 steps: in the first step, they used a TSM, part based models which are tree structured to detect the characters based on their locations.Then in the second step they build a CRF model on the potential locations of the characters. Based on the potential locations they build a CRF model. By using character detection scores,language model and spatial constraints to define a cost function which may be unary and pairwise. Lastly, by minimizing cost function word recognition would be done.

J. Zhou, H. Gao, J. Dai, D. Liu and J. Han,[28] The main structure of the suggested text recognition network is First, the encoder, which creates multi-granularity feature maps, encodes the input image. The residual stage feature map's cell relationships are recorded by the MSR layer, which is also used to integrate cells. The MSR layer feature map is then translated into a series of characters using the MHA blocks. A multi-head self-relation layer following the i-th residual stage is referred to as MSRi.The i-th multi-head attention block is denoted by MHAi. Additionally, L stands for elemental addition and N for point multiplication. This method is applicable not only to scene text recognition, but also to image recognition, image captioning, and other tasks.

Venkateswarlu, S. & Duvvuri, Duvvuri B K Kamesh [29] Proposed a framework that has two main steps. In the primary step, the image is converted into text using Tesseract optical character recognition (OCR). Then in the next step, the text is converted into speech using festival software. This model requires future work as it cannot work for text which has less than 12pt size and also for a reading distance greater than 42cm.

Tong He et al.,[30] did research on component based text extraction. They trained a model which has character label, text region segmentation, and binary text/non-text information. They help to extract specific features. The system they introduced has two parts: for component classification, a Text-CNN and for generating components, a CE-MSERs.

Zhen Zeng et al.[31], created an AlignTTS which is Align Text to Speech, it includes: Feed-Forward Transformer, Duration Predictor and Mix density network. They have designed a Mix density network for the model to learn alignment.it is based on a feed forward transformer. This AlignTTs generates a mel-spectrum for sequences of characters. They trained the model in a way that it is able to differentiate different text alignments. They used LJSpeech dataset for the experiment and the results proved that the model they have proposed, gives better performance and also 50 times faster than the real time in case of efficiency. This model has a duration predictor which predicts the duration of each character. It gives better performance.

S. Roy, P. P. Roy, P. Shivakumara,[32]Proposed a model which deals with the curved text using Hidden Markov Model(HMM). A sliding window's path is estimated, and the HMM system is given features from the window for recognition. examined the performance of MartiBunk and local gradient histogram, two frame-wise extracting features techniques. Used boundary growth to extract curved text lines and gradient directional characteristics to obtain text candidates. There are four main steps in the proposed model. First step, Wavelet-Gradient-Fusion method is used for the binarization of the text lines. Second step, the foreground text lines pixels are selected from the binary text lines pictures and then supplied into a curved fitting algorithm. Third step, feature extraction is achieved by the Marti-Bunke feature and the LGH feature. Final step the recognition of text is performed by HMM.

Lokkondra, Chaitra Yuvaraj, et al., [33] Text detection can be done in the first stage using approaches based on connected components or the sliding window method. While the linked component based method extracts the text and non-textual portions at the pixel level, the sliding window method scans a picture by sliding sub windows through the image. The best methods are linked component based since the sliding window method is computationally expensive. To avoid the drawbacks of existing techniques, text recognition is accomplished through three groups: bounding box regression, segmentation, and hybrid approaches. For text detection, various neural network techniques exist. Dynamic log polar transform and sequence recognition network are combined to handle the aligned text.

Xue, Minglong, et al., [34] CNN was used to create a technique for identifying texts in blurred and unblurred photos. This method determines the degree of blurriness in the input blurred image after applying a low pass filter based on gradient values. When pixels containing text areas are blurry to that extent, it is possible to detect the presence of text in the image. To distinguish between pixels that are text and those that are not, K-means clustering is used. To extract symmetry features, we employ the Bhattacharyya distance measure, a well-known distance metric for assessing similarity and dissimilarity across histograms of sub-regions of text candidate photos. The SVM Classifier's output of text components is aggregated for text detection.

3. METHODOLOGY

The proposed methodology can be addressed in three main steps:

3.1 Localize text by bounding boxes

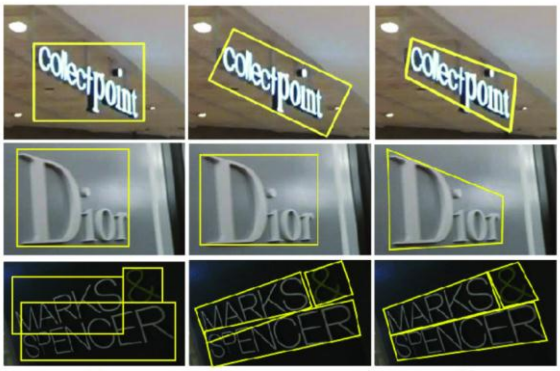

Fig -2: Localizing text by bounding boxes

This phase involves breaking down the entire word into smaller, more easily recognizable components. Elements on an image can be located by using localization information. In order to effectively execute text detection and bounding-box regression at all places and multiple scales in an image, we train the convolutional neural network using pictures and datasets like SVT and MSRA that contain fonts with varying sizes, colors, and orientations.

3.2 Crop the images and recognize text

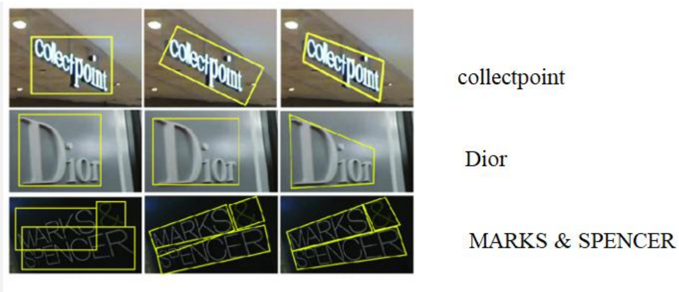

Fig -3: Segmenting text and non-text part from image

In this study, according to the predicted elements and geometrics, some threshold measures like K-means clustering are applied in order to filter the unnecessary background information and produce the character stanzas from the picture pixels. We also add a spatial transformer to this network to more effectively handle text with irregular shapes. The text extraction stage is represented by the first two steps.

3.3 Converting the text into speech

Fig -4: Converting detected text to speech

This is the last step where text extracted from image is converted into audio using a convolution network where the text input is aligned properly and then a decoder is used for converting aligned content to audio file.

4. CONCLUSION

The research mainly aims at understanding the text from natural scene images and converting it into audio file has challenging problems, due to multiple deteriorations like text variability of size, font, color, orientation, alignment. So, our proposed methodology introduced in this survey would solve issues that are existing in present applications.

5. ACKNOWLEDGEMENT

We would like to express our special thanks to our mentor, Mr. T. Gnana Prakash, for his time and efforts he provided throughout the year. His useful advice and suggestions were really helpful to us during the project's literature survey phase. In this aspect, we are eternally grateful to him.

6. REFERENCES

[1] Revathy A S, Anitha Abraham, Jyothis Joseph: A Survey on Text Recognition from Natural Scene Images. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.

[2] Luo, Canjie, Lianwen Jin, and Zenghui Sun. "Moran: A multi-object rectified attention network for scene text recognition." Pattern Recognition 90 (2019): 109-118.

[3] Zhou, Xinyu, et al. "East: an efficient and accurate scene text detector." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.

[4] Yan, Ruijie, et al. "MEAN: multi-element attention network for scene text recognition." 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.

[5] Huang, Weilin, et al. "Text localization in natural images using stroke feature transform and text covariance descriptors." Proceedings of the IEEE international conference on computer vision. 2013.

[6] Dai, Pengwen, Hua Zhang, and Xiaochun Cao. "SLOAN: Scale-adaptive orientation attention network for scene text recognition." IEEE Transactions on Image Processing 30 (2020): 1687-1701.

[7] Chandio, Asghar Ali, et al. "Cursive Text Recognition in Natural Scene Images Using Deep Convolutional Recurrent Neural Network." IEEE Access 10 (2022): 10062-10078.

[8] Mokayed, Hamam, et al. "A New Defect Detection Method for Improving Text Detection and Recognition Performances in Natural Scene Images." 2020 Swedish Workshop on Data Science (SweDS). IEEE, 2020.

[9] Özgen, Azmi Can, Mandana Fasounaki, and Hazim Kemal Ekenel. "Text detection in natural and computer-generated images." 2018 26th signal processing and communications applications conference (SIU). IEEE, 2018.

[10] He, Wenhao, et al. "Multi-oriented and multi-lingual scene text detection with direct regression." IEEE Transactions on Image Processing 27.11 (2018): 5406-5419.

[11] Donahue, Jeff, et al. "End-to-end adversarial text-to-speech." arXiv preprint arXiv:2006.03575 (2020).

[12] Bagi, Randheer, Tanima Dutta, and Hari Prabhat Gupta. "Cluttered textspotter: An end-to-end trainable light-weight scene text spotter for cluttered environment." IEEE Access 8 (2020): 111433-111447.

[13] Karpagam, Angia Venkatesan, and Mohan Manikandan. "Text extraction from natural scene images using Renyi entropy." The Journal of Engineering 2019.8 (2019): 5397-5406.

[14] Mokayed, Hamam, et al. "Anomaly detection in natural scene images based on enhanced fine-grained saliency and fuzzy logic." IEEE Access 9 (2021): 129102-129109.

[15] Liu, Zihao, Qiwei Shen, and Chun Wang. "Text detection in natural scene images with text line construction." 2018 IEEE International Conference on Information Communication and Signal Processing (ICICSP). IEEE, 2018.

[16] Lee, Chae Young, Youngmin Baek, and Hwalsuk Lee. "Tedeval: A fair evaluation metric for scene text detectors." 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). Vol. 7. IEEE, 2019.

[17] Baek, Youngmin, et al. "Character region awareness for text detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[18] Ranjitha, P., and K. Rajashekar. "Multi-Oriented Text Recognition and Classification in Natural Images using MSER." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.

[19] Liu, Juhua, et al. "SemiText: Scene text detection with semi-supervised learning." Neurocomputing 407 (2020): 343-353.

[20] Nagdewani, Shivangi, and Ashika Jain. "A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION." (2020).

[21] Zuo, Ling-Qun, et al. "Natural scene text recognition based on encoder-decoder framework." IEEE Access 7 (2019): 62616-62623.

[22] Q. Wu, P. Chen and Y. Zhou, "A Scalable System to Synthesize Data for Natural Scene Text Localization and Recognition," 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR), 2019, pp. 59-64, doi: 10.1109/RCAR47638.2019.9043965

[23] Pise, Amruta, and S. D. Ruikar. "Text detection and recognition in natural scene images." 2014 International Conference on Communication and Signal Processing. IEEE, 2014.

[24] G. Vaidya, K. Vaidya and K. Bhosale, "Text Recognition System for Visually Impaired using Portable Camera," 2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW), 2020, pp. 1-4, doi: 10.1109/ICCDW45521.2020.9318706.

[25] Karanje, Uma B., and Rahul Dagade. "Survey on text detection, segmentation and recognition from a natural scene images." International Journal of Computer Applications 108.13 (2014): 39-43.

[26] F. Wang, L. Zhao, X. Li, X. Wang and D. Tao, "Geometry-Aware Scene Text Detection with Instance Transformation Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1381-1389.

[27] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao and Zhong Zhang, The State Key Laboratory of Management and Control for Complex Systems, CASIA, Beijing, China, "Scene Text Recognition using Part-based Tree-structured Character Detection." IEEE conference 2013.

[28] J. Zhou, H. Gao, J. Dai, D. Liu and J. Han, "A Multi-head Self-relation Network for Scene Text Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 3969-3976, doi: 10.1109/ICPR48806.2021.9413339.

[29] Venkateswarlu, S. & Duvvuri, Duvvuri B K Kamesh & Jammalamadaka, Sastry & Rani, R.. (2016). Text to Speech Conversion. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i38/102967.

[30] Tong He, Weilin Huang, Yu Qiao, and Jian Yao, "Text-Attentional Convolutional Neural Network for Scene Text Detection", IEEE conference 2016.

[31] Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, Jing Xiao, "AlignTTS: Efficient Feed-Forward Text-to-Speech System without Explicit Alignment", 978-1-5090-6631-5/20/$31.00 ©2020 IEEE

[32] S. Roy, P. P. Roy, P. Shivakumara, G. Louloudis, C. L. Tan and U. Pal, "HMM-Based Multi Oriented Text Recognition in Natural Scene Image," 2013 2nd IAPR Asian Conference on Pattern Recognition, 2013, pp. 288-292, doi: 10.1109/ACPR.2013.60.

[33] Lokkondra, Chaitra Yuvaraj, et al. "ETDR: An Exploratory View of Text Detection and Recognition in Images and Videos." Rev. d 'Intelligence Artif. 35.5 (2021): 383-393.

[34] Xue, Minglong, et al."Curved text detection in blurred/non-blurred video/scene images." Multimedia tools and applications 78.18 (2019): 25629-25653.