Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS 2022)
IEEE Xplore Part Number: CFP22CB5-ART: ISBN: 978-1-6654-6084-2

Text Localization and Recognition from Natural Scene Images using AI

D.Shekar Goud[1], Vigneshwari. M[2], Aparna P[3], Vijayasekaran G[4], Ajay Singh Yadav5, Ashok Kumar[6]

[1]Department of Electronics and Communication Engineering, Ellenki College of Engineering and Technology, Patelguda, Telangana, India. shekar.embedded@gmail.com

[2]Department of Information Technology, Vel Tech Multi Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu, India. vigneshwari@veltechmultitech.org

[3]Department of Computer Science Engineering, R. M. K. College of Engineering and Technology, Thiruvallur, Tamil Nadu, India. aparna.prakasam88@gmail.com

[4]Department of Computer Science and Engineering, Sir Issac Newton College of Engineering and Technology, Nagapattinam, Tamil Nadu, India. gunavijay90@gmail.com

[5]Department of Mathematics, SRM Institute of Science and Technology, Ghaziabad, Utter Pradesh, India, ajay29011984@gmail.com

[6]Department of Computer Science, Banasthali Vidyapith, Rajasthan, India. kuashok@banasthali.in

2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) | 978-1-6654-6084-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACRS55517.2022.10029220

Abstract— **In computer vision systems, text detection and recognition (TDR) in natural scene images can be used for things like license plate recognition, automated street sign interpretation, and assisting blind people. Accordingly, finding text within an image is a time-based challenge in the field of computer vision. Because of factors like cluttered backgrounds, image blurring, partially obscured text, various fonts, noise, and fluctuating lighting, text identification in natural scenes has become a significant task with the increase in the use of actual vision systems. Images and videos with accompanying textual data can be leveraged for automatic annotation. This study provides a system for automatically identifying the text from images, and it discusses the methodology behind locating and recognizing text in images of natural scenes. This article handles the scene text recognition challenge from start to finish, breaking it down into text localization and recognition. The Maximally Stable Extremal Regions (MSER) technique is used to detect text and non-text regions in images for localization purposes. Convolutional Neural Networks (CNNs) and convolutional recurrent neural networks (CRNNs) are utilized for text recognition. By evaluating the accuracy, precision, and F1 score, the CRNN is determined to be the best.**

Keywords— Text, Natural images, Resize, Enhancement, Localization, Accuracy, Deep Learning.

I. INTRODUCTION

Understanding the visual scene is a topic of current research in the computer vision community. It necessitates extensive study in computer vision and related subfields. Understanding a visual image necessitates the processing of both text and image, and comprehending the scene and analyzing the text displayed in the image is typically a challenging task. Because of the value, it provides in a variety of circumstances, like content-based image retrieval, assistive devices for vision impairment, automated automotive navigation, and the digitization of academic materials, this field of study is progressively increasing.Although optical character recognition (OCR) is a tried-and-true method for extracting text from scanned documents, it falls short when applied to scene images[1]. Text identification from a scene requires specialized features since the characters in the picture can vary in terms of dimensions, hues, fonts, orientations, perspectives, illumination, sharpness, depth of field, and overall intricacy of the background [2]. Such problems emerge regularly in the domain of TDR. Text detection, in general, locates the exact pixels on an image where text is present, constructing a bounding box around each letter, word, and line of text to improve the precision of subsequent TR.Text Recognition (TR) software may analyze an image of a scene as input and generate predictions about the text contained within it. TR is the commonly used technique for translating old documents into electronic formats. Because most image processing algorithms follow the same principles, this procedure appears simple and easy to execute.

Based on natural language processing with the help of Artificial Neural Networks (ANN), the research [3] offered a method for identifying written text information. As a result of its capacity to gather information from various scanners, digital iPads, cameras, and documents, the framework can accommodate a wide variety of sources. Utilizing methods from image processing, and artificial intelligence, this study details the creation of a framework capable of identifying both natural and written text. Accurate handwriting recognition is achieved by the suggested technique.The goal of this work [4] is to fill the knowledge gaps in the field of Arabic scene TR. To begin, by introducing a completely new Arabic/English dataset.

Second, while DL techniques have improved quickly and set a new standard for Latin-based languages, they have found minimal application to Arabic. As a result, use the dataset to assess the TR problem from three perceptions: i) exploring the applicability of DL methods for Arabic scene TR, wherein we isolate critical features for a high-performing model; ii) differences between Arabic and Latin texts that need to be addressed differently; iii) exploring a model that can simultaneously process Arabic and Latin text. We analyze the pros and cons of many approaches for handling directed text and compare them.In the paper [5,] we compare various DL approaches used in the Transformer's encoder to recognize scene text. We begin by replacing the encoder's feed-forward network (FFN) module with either a squeeze-and-excitation (SE-FFN) or a cross-stage partial (CSP-FFN). The encoder's general design is then replaced by local dense synthesizer attention (LDSA) or the Conformer structure. When the total number of parameters is considered, the conformer encoder obtains the highest test accuracy, while SE and CSP-FFN also perform comparably. When their attention maps are examined, different combinations of encoders can produce higher performance.

The study [6] creates and develops a DL-based Fusion Neural Network (FNN) for TDR. The goal is to enhancethe accuracy of TR. A fusion NN, in particular, employs many NN, each of which provides a distinct layer. The Convolutional Layer (CL) is in charge of acquiring the feature sequence, whereas the Recurrent Layer (RL) is in charge of building the best training model to increase accuracy. We compare the proposed model to existing approaches and evaluate FNN using the Devanagari MLT-19 dataset to determine its efficacy. More crucially, the FNN model correctly detects a large proportion of script words.Number plates on moving vehicles can contain both text characters and numbers; this study [7] explains how to understand both. A RaspberryPi with an attached cam records footage of oncoming traffic. License plate images are taken from the footage and pre-processed to increase resolution and decrease jitter. Data from the text is input into a NN that makes predictions. To perform feature extraction and character recognition in license plate images, this study utilizes a ResNet-34 architecture.The fundamental aim of thestudy [8] is to design a model that can recognize written Sinhala and English scripts at the same time using character image geometry properties and ANN. The system will be modernized so that it can accommodate all four of Sri Lanka's languages. The main technological foundations on which this research was built were character geometry features and Ann. Using a collection of 800 images, an 85% efficiency was attained. TR is evaluated using bi-lingual texts via isolating specific characters and delivered to the algorithm.The paper [9] describes a detailed approach for uploading a text image and extracting the text verbatim from the entire image. Our thesis contributes significantly in three areas: i) a more accurate text localizer that has been trained on both printed and handwritten texts; ii) new word-classification methods for telling typed text apart from the handwritten text; iii) the application of effective image pre-processing techniques to cropped examples of handwritten text to prepare them for use as inputs to a DL model, thereby increasing the model's accuracy.

According to research [10], machine learning and DL have had a substantial impact on TDRsystems. The primary aim of this research is to offer a method for TDRbased on an AI model in light of the aforementioned difficulties and solutions. Suggesting a scene TR system that comprises three stages: The image is first processed using an image processing approach to reduce noise. We can localize the text area, construct a bounding box around each letter, and considerably speed up the recognition process by applying a localization algorithm to the processed image. The third stage entails the creation of an AI-based framework for character prediction. The classifier is reinforced in this case by combining CNN, RNN, and CTC; this is a valuable technique for TR. The rest of the paper is formatted as follows: Part 1 discussespast works. Part2 confers the suggestedframework. Part3 discusses data collection and analysis methods. Parts4 and 5 go over the processes of text localization and recognition. Part6 presents the experiment results and compares them to standard rules. Part7 contains the concluding ideas and their potential impact.

II. METHODOLOGY

The methodology of the suggested system is given in figure 1. The suggested system is composed of five important steps.

Step 1 - Data Collection: To detect the text present the natural images, the data is very important to start the research. The research uses the CUTE-80 dataset.

Step 2 – Data Processing: The image collected from CUTE-80 are of different sizes and there are some other problems like noise, and color illumination issues. To eliminate all the irregularities, pre-processing is done.

Step 3 – Text Localization: To identify the text from the images. The localization of text is very important because it makes the process simpler. After localization in this stage, character grouping is used to identify the individual character.

Step 4 – Text Recognition: After localization, the outcome of the text localization is given to the DL model for identifying the text. The two DL models are CNN and CRNN.

Step 5 – Validation: The two DL models are compared using the metrics like accuracy, precision, and F1 to identify the best one.

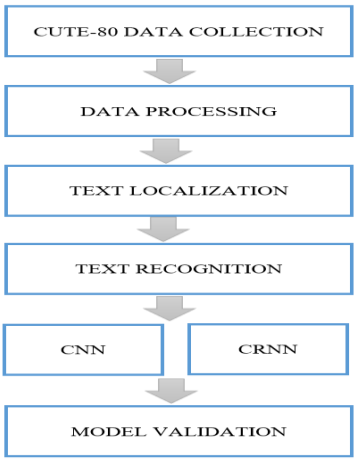

Fig. 1. Proposed System Flowchart

III. DATA AND PROCESSING

It's CUTE80 [11], a collection of eighty images of curving text lines with complicated backgrounds, perspective distortion effects, and low-resolution effects (circle, Z, S shaped text). Curved text demonstration using CUTE80 is required to validate the effectiveness of the suggested approach. The sample images from CUTE-80 data are given in figure 2

Fig. 2. CUTE-80 Sample Data

The ability to process images using advanced technology known as "image processing" is a crucial skill. Processing that examines and enhances a digital image through analysis and manipulation. Many different kinds of image processing, including resizing, enhancing, and so on, are available in Processing.

A. Resize

In recent years, there has been a lot of focus on image scaling as a crucial method for presenting images on various devices. Because of the wide variety of viewing environments, resizing images to ensure they look good in each image is now a challenge. Image resizing, often called image retargeting, involves making adjustments to an image's aspect ratio and dimensions so that it better fits its intended use. Scaling and other simple image resizing operators can often achieve good results, nevertheless. With a deeper comprehension of image semantics, image resizing can be accomplished more successfully [12].

B. Enhancement

To see all of an image's details, image enhancement is necessary. Capturing an image in reduced illumination results in poor contrast and visibility. There are several methods available for enhancing image quality. The process of improving an image includes cleaning it of unwanted images like images and debris, making it brighter and sharper so that important images stand out, and removing unwanted noise [13]. The image is improved using a technique called histogram-based enlargement.

IV. TEXT LOCALIZATION

Text localization and detection will considerably improve any suggestion for scene detection and identification. We used maximal stable extreme regions(MSER) for text localization.

The MSER method evaluates the extracted stable images from each of the image's color channels to decide which text candidates are of the greatest quality. When it comes to coordinating images over large-scale changes the multi-resolution MSER performs better due to the advantages it offers in both cases. To detect small letters in low-resolution images, the suggested MSER combines the complementing capabilities of canny edges and MSER. Figure 3 depicts what MSER createsthe outcome.

Fig. 3. Localization Of Text from Natural Scene

Character level grouping

As a starting point, we employ the region around each identified character in a low-resolution image. Because the criteria for correct word identification are dependent on geometric qualities, the widths (w1, w2) and heights (h1, h2) of bounding boxes will be considered. Eqs. (1), (2), and (3) follow the description of geometric properties and are written as follows:

$$h = \min(h_1, h_2) \quad [1]$$

$$\partial x = \frac{|l_1 + l_2| - (w_1 - w_2)}{2} \quad [2]$$

$$\partial y = |m_1 - m_2| \quad [3]$$

where ∂x is always negative when the boxes are aligned along the x-axis. Thus, they complement one another and can safely be assumed to be from the same source. Character-bounding explanations will be simplified if the below conditions are met:

$$|h_1 - h_2| < k_1 h \quad [4]$$

$$\partial x < k_2 h \quad [5]$$

$$\partial y < k_3 h \quad [6]$$

where k_1; k_2 and $k_3$were character-bounding variables. To classify a set as character or non-character, the third argument, k_3, must be present. All characters that have been

discovered undergo a similar procedure. Figure 4 below displays the final result of character bounding:

Fig. 4. Character Grouping of Text from Natural Images

V. TEXT RECOGNITION

The next step is TR, which can be done once we've found the bounding boxes containing the text. Several methods exist for reading the text. CNN and CRNN are the primary topics of this study.

A. CNN

A CNN's architecture is based on the arrangement of neurons in the human brain's visual cortex. Convolution is most often utilized in image processing, specifically for smoothing, enhancing, and sharpening. Each pixel in the image is multiplied by the corresponding kernel pixel, and the resulting products are added together to create a unified pixel value of the filtered image.Pooling is another important aspect of CNN. To reduce network computation complexity, we shall gradually expand the spatial dimension of the image. The maximum pool maintains the model's efficient training cycle at the expense of less computational power by reducing dimensionality while retaining the major characteristics, which are rotational and translational variants. It removes noisy activations, de-noises them, and reduces their dimensions [14]. A fully connected NN is used at the end of a CNN. The network's weights are calculated during the training phase. The output of the convolution/pooling process is sent to the fully connected neural network, which uses the input data to find the best-fitting label for the image. In this image, we connect the image's feature vector to its classification. Multiplying the convolution/pooling firm's outcomes by the network's channel weights is a common way to improve accuracy.

B. CRNN

The CRNN is a neural network architecture that combines CNN with recurrent neural networks (RNN) with the Connectionist Temporal Classification (CTC) loss for TR tasks [15, 16]. The CRNN network structure consists of the CL, RL, and transcription layer (TL). Automatic feature sequence extraction from each input image is the responsibility of the CLs at the CRNN's foundation. As the CL's output is fed into a recurrent network, predictions may be made for each frame in the feature sequence. The RL makes predictions per frame, and the TL at the CRNN's top translates those predictions into a label sequence. The CRNN model's CL is based on the CL and max-pooling layers of a standard CNN model (fully-connected layers are removed). You can utilize this part to get the image's sequential feature representation. To construct the RL of a deep bidirectional RNN, one starts with the CL. To begin, RNN's capacity to acquire sequence-level context is remarkable. Second, the recurrent and the CL can be trained simultaneously since RNN can back-propagate error differentials to its input, the CL. The third advantage of RNN is that it can process sequences of any length by going through them from start to finish.

VI. RESULTS AND DISCUSSION

The CUTE-80 data is used to identify the text from the natural scene images. The data collected were first processed using rescaling and enhancement techniques. The processed data is used for the localization and recognition process. For localization, MSER is used and for recognition, CNN and CRNN are employed. To identify the best model the three various metrics are used accuracy, precision, and F1. The accuracy of CRNN will be greater and the attained value is 80.98%. For CNN the obtained accuracy is 77.48%. Next, in precision, the value will be good for CRNN (78.76%) and the value will be lesser for CNN (74.85%). Finally, the 72.56% F1 score value is the result of CNN and 76.21% will be the outcome of CRNN. The metrics comparison is given in table 1.

TABLE 1. DL MODEL COMPARISON ON TEXT RECOGNITION

METRICS	CNN	CRNN
Accuracy (%)	77.48	80.98
Precision (%)	74.85	78.76
F1 (%)	72.56	76.2118

The above-mentioned metrics in table 1 are used for numerical comparison. For visual comparison, the line plot is used and it is given in figure 5. The brown line represents the CRNN metrics in % and green for CNN metrics in %. The outcome shows the CRNN will be the better model for TR from natural images.

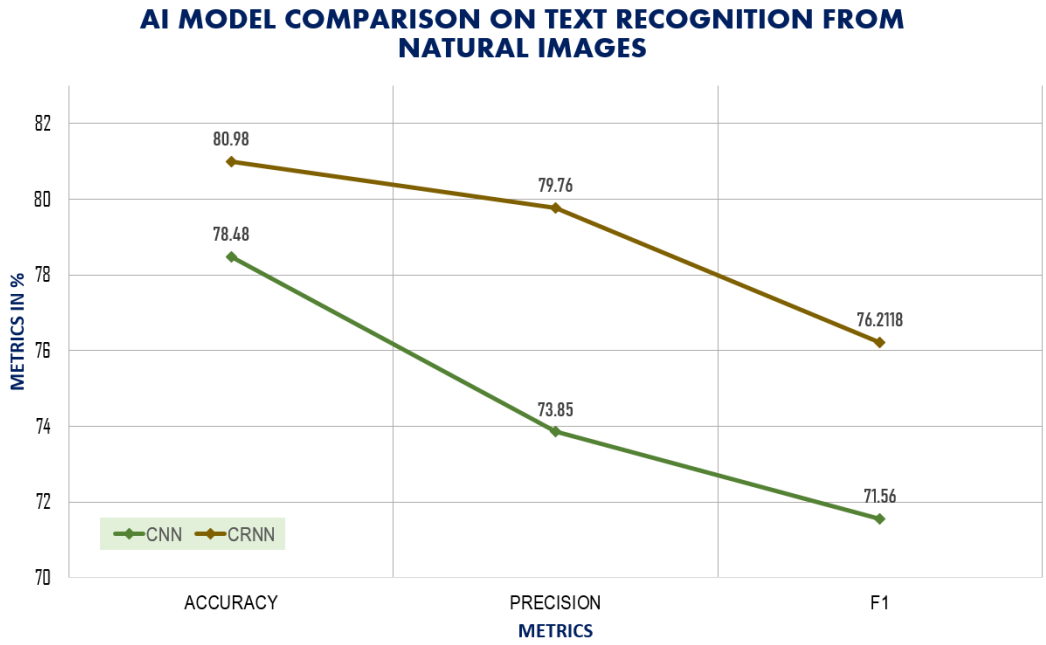

Fig. 5. AI Comparison of TR from Natural Images

The outcome of the CRNN is given in figure 6. Figure 6.a shows how well the CRNN model on the TR task. The CRNN correctly identified the text like "ARSENAL", "WALCOTT", and "14" from the uploaded t-shirt image. Similarly, Figure 6.b shows the CRNN correctly identified the text "WEST HAM UNITED" in the given image.

a)

b)

Fig. 6. CRNN Outcome

VII. CONCLUSION

The recognition of text from natural images is very important and it has great welcome in many domains like medicine, security, agriculture, military, etc. The CUTE-80 data is collected and processed to remove unwanted information and make the data perfect for further processes like localization and recognition. MSER for localization, CNN, and CRNN for recognition of text from the images. The outcome of both DL models is compared using the metrics. The metrics score tells the superiority of CRNN over CNN. Lastly, the CRNN is finalized as a model for deployment. In the future, a mobile app will be created. The app will recognize the text and convert the text to speech format. This will be helpful for illiterate and blind people.

REFERENCES

[1]. K. Hamad and M. Kaya, "Detailed analysis of optical character recognition technology," International Journal of Applied Mathematics, Electronics and Computers, vol. 4, no. Special Issue-1, p. 244, 2016.

[2]. Q. Ye and D. Doermann, "Text detection and recognition in imagery: a survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 7, pp. 1480–1500, 2015.

[3]. S. I. Manzoor and J. Singla, "A Novel System for Image Text Recognition and Classification using Deep Learning," 2021 International Conference on Computing Sciences (ICCS), 2021, pp. 61-64, doi: 10.1109/ICCS54944.2021.00020.

[4]. H. Hassan, A. El-Mahdy and M. E. Hussein, "Arabic Scene Text Recognition in the Deep Learning Era: Analysis on a Novel Dataset," in IEEE Access, vol. 9, pp. 107046-107058, 2021, doi: 10.1109/ACCESS.2021.3100717.

[5]. Y. -G. Kim, H. Kim, M. Kang, H. -J. Lee, R. Lee and G. Park, "Analysis of the Novel Transformer Module Combination for Scene Text Recognition," 2021 IEEE International Conference on Image Processing (ICIP), 2021, pp. 1229-1233, doi: 10.1109/ICIP42928.2021.9506779.

[6]. S. K. Dasari and S. Mehta, "Text Detection and Recognition Using Fusion Neural Network Architecture," 2022 8th International Conference on Advanced Computing and Communication Systems

(ICACCS), 2022, pp. 2067-2071, doi: 10.1109/ICACCS54159.2022.9785137.

[7]. K. T. Ilayarajaa, V. Vijayakumar, M. Sugadev and T. Ravi, "Text Recognition in Moving Vehicles using Deep learning Neural Networks," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 279-283, doi: 10.1109/ICAIS50930.2021.9395980.

[8]. H. W. H. Premachandra, A. Jayakody and H. Kawanaka, "Converting high resolution multi-lingual printed document images in to editable text using image processing and artificial intelligence," 2022 2nd International Conference on Image Processing and Robotics (ICIPRob), 2022, pp. 1-7, doi: 10.1109/ICIPRob54042.2022.9798739.

[9]. Nidhi, D. Ghosh, D. Chaurasia, S. Mondal and A. Mahajan, "Handwritten Documents Text Recognition with Novel Pre-processing and Deep Learning," 2021 Grace Hopper Celebration India (GHCI), 2021, pp. 1-5, doi: 10.1109/GHCI50508.2021.9514054.

[10]. Kantipudi, Mvv Prasad & Kumar, Sandeep & Jha, Er. Ashish,"Scene Text Recognition Based on Bidirectional LSTM and Deep Neural Network", Computational Intelligence and Neuroscience, vol 2021, pp. 1-11, 2021, doi: 10.1155/2021/2676780.

[11]. Risnumawan, A., Shivakumara, P., Seng, C., & Lim, C,"Expert Systems with Applications A robust arbitrary text detection system for natural scene images", Expert Systems With Applications, vol. 41, no. 18, pp. 8027–8048, 2014, doi: 10.1016/j.eswa.2014.07.008, 2014

[12]. Priyanka C. Dighe, Shanthi K. Guru, "Survey on Image Resizing Techniques", International Journal of Science and Research (IJSR), vol. 3, issue. 12, 2014

[13]. Bansal, A., & Singh, N, "Image Enhancement Techniques: A Review", Asian Journal For Convergence In Technology (AJCT), vol. 6, pp. 07-11, 2020 ,doi: 10.33130/AJCT.2020v06i02.002

[14]. S. Ahlawat, A. Choudhary, A. Nayyar, S. Singh, and B. Yoon, "Improved handwritten digit recognition using convolutional neural networks (CNN)," Sensors, vol. 20, no. 12, p. 3344, 2020.

[15]. F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. "Learning precise timing with LSTM recurrent networks", JMLR, vol. 3, pp. 115–143, 2002

[16]. B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, 1 Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.