iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

C-RNN Approach for Text Detection and Recognition

Varsha Bhashyam[1], Krithigha Sukarna Kumar[2], Sasi Karthik. D[3], Deepa. R[4]
[1, 2, 3,]*U.G Scholar,* [4]*Assistant Professor, CSE department, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamil Nadu, India*

Abstract: ***Text detection and recognition in today's world plays a crucial role in examining the context of the images. One way of storing information present in the paper form of documents into system as its digital version of document is by scanning those documents. The texts that are present in the scanned images In this paper, CRNN algorithm is used for text detection and recognition, Googletrans API for language translation and gTTS API for text-to-speech conversion. The proposed system is tested by Synthetic word dataset (MJSynth). The extracted text undergoes language translation and is converted into speech for visually impaired people.***

Keywords: Text extraction, OCR, Googletrans, gTTS, CRNN, Deep learning.

I. INTRODUCTION

The technical term for text information present in the images is called Image text. These image texts can be found in magazines, newspapers, pamphlets, posters and so on. One way of storing information present in these paper records into the system is by scanning those records. By doing so, the documents will be stored as images in the system. The image text present in the images cannot be modified by the users. But for reusing the information, the computer system might find it difficult to read and search the contents from these documents. [13]The main challenges involved are: Quality of the images and font characteristics of the characters in documents which are different compared to that of font of the characters present in the computer system. The image text helps in solving problems and also cultural gaps.

The prevailing system is the OCR (Optical Character recognition) technology that has been used for converting the image text in scanned documents into ASCII symbols. However the limitation of OCR testing is that the existing OCR systems do not go well when the image background is very dark or shaded. These limitations occurs in monetary documents, magazines and manuscripts. In the current OCR system, the scanned images are required to be binarized before text segmentation and recognition can be performed. As noted by many researchers, global thresholding is not possible for blurry, shady and complicated images. Unfortunately, the current OCR systems tends to work poorly in these cases.

The proposed system is an advanced system which detects and recognizes text in images. The input to the system are colored images which contains image text. The text detection process undergoes text localization, tracking and enhancement followed by removal of non-text regions and recognizing the text. After the recognition of text, it is fed into Googletrans API for language translation and gTTS API for text to speech conversion.

The paper is discussed as follows: Sub-division II presents related work, Sub-division III provides overview and architecture of text recognition, Sub-division IV describes the methodology used, Sub-division V discusses the applications, Sub-division VI reviews the conclusion followed by future work present in Sub-division VII.

II. RELATED WORK

Different methodologies for text detection have been developed to effectively understand pattern recognition. Since this is an agile research there is more demand for developers in this field of study. Below listed are some of the findings discovered by some researchers to improve the efficiency of text detection.

Xiaoqing Liu et al. [1] has discovered a Multi-scale edge based algorithm, where the user can detect and extract text from noisy and shady background images. By using this algorithm, the precision rate of the result is 91.8% and recall rate is 96.6%. The future work of this paper entails using suitable prevailing OCR techniques to perceive the text extracted.

Adam Coates et al. [2] has worked-on an algorithm in which the performance measures are high and solutions are adaptable. This research has focused on feature learning algorithm that uses natural scene images. They evaluated the results of unsupervised learning algorithm using ICDAR 2003 training dataset. By using more subtle and advanced algorithm, accuracy of the results might increase.

Neha Gupta et al. [3] has proposed a digital processing Discrete Wavelet Transform algorithm. The size, style, orientation, color of the font and alignment of text and similar factors do not affect the results. This method's results proved to have low processing time and thereby increasing the accuracy. The future work of this paper involves using suitable extant OCR techniques to recognize the text detected.

Xu-cheng Yin et al. [4] has used MSERs pruning algorithm that permits the user to detect characters even if the quality of the image is shallow. The distance weights and threshold can be grasped concurrently using self-training distance metric algorithm. They evaluated the results using ICDAR 2011 dataset. An prevailing OCR technique might improve this paperwork results.

Tong He et al. [5] has suggested Text-Attentional Convolutional Neural Network algorithm for differentiating between regions containing texts from other regions. It supports images that has unclear, non-text and also binary texts. This method uses four databases, namely: the ICDAR 2005[8], ICDAR 2011[9], ICDAR 2013[10] and MSRA-TT500 [11]. Using text line construction, better performance can be expected.

Xiaohang Ren et al. [6] has proposed Multilayer CNN, statistical learning algorithm to address the challenges such as noise, contrast, contrast variations and plentiful complex backgrounds in natural images. This method can detect text independent of perspective and rotation. The limitation of this method is that small letters cannot be detected in case of motion or de-focus blur by applying plain Multilayer CNN algorithm to images of limited resolution.

Sahil Thakare et al.[7] has discussed Tesseract, OCR, Segmentation and Google-Trans so as to separate the document in such a simplest way that it would scale back the complexity to grasp the document and build it easily available in the most understandable form anyone could need. One of the constraint of the proposed system is that the user has to feed the input language manually.

III. TEXT RECOGNITION SYSTEM

An optical recognition problem basically comes under image based sequence recognition problem. The most suitable neural network for the sequence recognition problem is Recurrent neural network and convolutional neural network is used for an image based problem. In order to cope up with the problems related to Optical character recognition, merging both convolutional neural network and Recurrent neural network algorithms seems like an ideal choice.

Fig.1 describes the flow of C-RNN architecture [14]. The architecture is built from bottom to top manner. It contains three important layers:

1) *Convolutional Layers:* The text present in the input images are given to few convolutional layers in order to get recognized. These layers are used for extracting significant features from the input images.
2) *Recurrent Layers:* The final result from the convolutional layers are fed into recurrent layers as its input. Deep bi-directional Long Short Term Memory is one of the important network in the recurrent layer. RNN is capable of capturing contextual information as a sequence. The output from the RNN layer contains probability values for which each label corresponds to its input feature (or time step).
3) *Transcription Layer:* The terminal layer in the architectural diagram is the transcription layer. This layer uses connectionist temporal classification to speculate the output for all input features. The actual output is calculated by using the label with highest probability score for all input feature which then merges the outputs of those features.

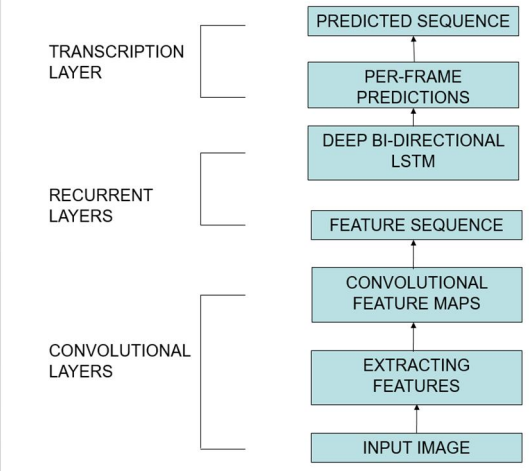


Fig.1 Architecture Diagram

IV. METHODOLOGY

A. Dataset

The data used in the dataset is taken from Visual Geometric Group. This is a very large dataset that consists of 10GB images. Here, only 1,35,000 images are used for training set and 15,000 images for validation dataset. The data contains text image segments as shown in the Fig.2.

Fig.2 Image samples from MJSynth dataset

B. Pre-Processing

After collecting dataset, preprocessing both the input image and output labels is required to be performed for the model to be acceptable. Preprocessing of input image is performed as follows:

1) Image is read and is converted into gray-scale image.
2) Each image is resized by using padding.
3) To make the image accordant with the input shape of architecture, image dimension is expanded.
4) The image pixel is then normalized.

The preprocessing of output labels is performed as follows:

a) Every image that has textual information present in it is read.
b) Each character in the word is encoded into random numerical value for which a function is created.
c) The maximum length of the words are estimated and every output label is padded to ensure it is of same size as the maximum length so that it is suitable with the output shape of RNN architecture.

Two other new lists are also created in this preprocessing step, namely: label length and input length to RNN. The CTC loss depends mainly on these two lists. The length of each output text label is called as label length and the input to LSTM layer is called as input length.

C. Network Architecture

1) The dimension of the input image is height- 32 and width- 128.
2) The number of convolutional layers used are seven. Out of which, six layers are of size 3x3 kernel filters and one layer is of size 2x2 kernel filter. The filters are made to increases gradually[12].
3) The number of max-pooling layers added with greater magnitude of width are four, which aids in feature extraction and estimate texts that are consecutively long. Two of which are of size (2, 2) and other two are of size (2, 1).
4) Batch normalization layers are used mainly after 5^{th} and 6^{th} layers for stimulating the training period.
5) The output of the convolutional layer is extracted in such a way that it is made compatible with the LSTM layer by using a lambda function.
6) The number of bi-directional LSTMs used are two.

D. CTC Loss Function

CTC loss function is abbreviated as Connectionist Temporal Classification. In any neural network, the function of the transcription layer is carried out by the CTC loss function. Main disadvantages of not using CTC is that: a) The processing time is high. B) Also scanning of a single character might be stored multiple times and when decoded, might produce ambiguous results. These problems can be surmounted by using CTC loss function. Fig.3 gives a diagrammatic comparison of decoding characters with and without using CTC.

FIG.3.1 WITHOUT CTC LOSS FUNCTION

FIG.3.2 WITH CTC LOSS FUNCTION

Fig.3.1 depicts the output predicted without using any loss function wherein the similar characters that are adjacent to each other are merged to give one single character producing wrong results. Whereas, Fig.3.2 depicts the accurate output of the encoded word "GOOD". In case of using CTC, the duplicated characters which needs to be removed are separated by a blank space character and hence producing exact results.

E. Post-Processing

1) Training: Adam Optimizer is used for minimizing the cost function and to train the model that consists of 150000 images. Out of which, 135000 images belongs to training set and 15000 images belongs to validation set. Fig.4.1 represents parameters of the trained model.

```
=================================================================
Total params: 6,619,711
Trainable params: 6,617,663
Non-trainable params: 2,048
```

FIG.4.1 Parameters of the model network

2) Testing: As the probability for each class at each input feature is estimated by this model, a transcription Iis recorded for decoding to original texts. CTC decoder helps in decoding the image texts. Fig.4.2 represents the sample results from the trained model.

Fig.4.2 Sample results from the trained model

F. Translation & Text-To-Speech Conversion

After testing the model, the output is passed as a parameter to the Googletrans API. This API sends query, requesting for the access to Google translation web page by using Google Translate Ajax API. The parameter passed gets translated into different language by feeding the output from the model into the translation text box in that web page. The translated text is then returned to the source code. Subsequently, the translated text is given as the input to gTTS (Google Text-to-Speech) API. This API is a Command Line Interface tool that converts the translated text into audio form such as mp3 file format. The final output will be the translated text in the form of audio.

V. APPLICATIONS

The text recognition, translation and text-to-speech conversion modules that are bound together as one whole unit of technology can be a buoyant implementation in real world applications. With an effective increase in accuracy of this technology, upcoming points are some of the possible applications in different sectors:

A. Accessibility

Communication devices for people who are visually impaired or having speech and learning disabilities and also for people who face literacy issues.

B. Banking and Finance Sector

1) In the field of stock market, the system of voice messaging helps the customer to a great extent by keeping a check on their investments and keeping them abeam of market news.

2) The paper cheque can be given to the bank officials who then places it in a machine that scans the cheque using the text recognition concept and transfers the amount accordingly.

3) In banking sector, automated audio messages translated in various languages aids customers to operate on an ATM machine through guidance system.

C. Broadcasting and Media Notifications

In radio broadcasting, emergency alerts, warnings and hot news can be broadcasted through air waves (as it travels more swiftly) by scanning the newspaper using text recognition, translating and converting text-to-speech conversion and thereby providing information for layman listening to the radio.

D. Legal Sector

In the present scenario, the legal profession involves paper documents mostly in digital format. This is more time-saving than manually reading through pile of paper files. Hence if the documents are entered into a common database such as digital library, it would provide easy and efficient access to documents for the legal professionals.

E. Health Care

1) In hospitals, patients have many forms documented such as general health form and insurance forms. To keep track of those records, an image recognition system might come in handy to store them in an electronic database.

2) Sometimes the names of the medicines prescribed by the doctor might be unclear to the pharmacist. In order to avoid the ambiguity in these situations, text recognition and text-to-speech conversion can be an optimal solution that involves low processing time and improved efficiency.

VI. CONCLUSION

In this paper, an optimal solution for text detection and recognition using C-RNN algorithm has been recommended to outperform the existing efficiency. The proposed system is an advanced system that detects and recognizes text in images. To cope up with OCR related complications, rather than applying either CNN or RNN algorithm solely, both the algorithms are integrated to work on it. This algorithm imparts advantages of both local feature extraction by convolutional neural network and transient characterization by recurrent neural network. After the text is extracted, it is translated from English to any preferable language by using Googletrans API. Furthermore, this translated text is fed into gTTS API for text-to-speech conversion. The final output will be in the form of audio.

VII. FUTURE WORK

Although the result of the proposed method proves to be efficient in contrast to other traditional methods, there is still some space left for improvements that can be carried out in future research. To begin with, upgrading the system to recognize handwritten texts rather than using just synthetic dataset can be a good refinement of the concept. For an augmented research, datasets categorized by the challenges of real world is a salient breakthrough. Also, the proposed system can be used as a base for developing this technology into a mobile application.

REFERENCES

[1] Xiaoqing Liu, J. S. (2006). Multiscale edge-based text extraction from complex images. IEEE, 4.

[2] Adam Coates, B. C. (2011). Text detection and character recognition in scene images with unsupervised feature learning. IEEE, 6.

[3] Neha Gupta, V. (April 28-29,2012). Image segmentation for text extraction. 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE'2012) (p. 4). Singapore: Semantic Scholar.

[4] Xu-Cheng Yin, X. Y.-W. (2nd June 2013). Robust text detection in natural scene images. IEEE, 14.

[5] Tong He, W. H. (24th March 2016). Text-Attentional Convolutional Neural Network for scene text detection. IEEE, 13.

[6] Xiaohang Ren, K. C. (7th April 2016). A Novel scene text detection algorithm based on Convolutional Neural Network. IEEE, 5.

[7] Sahil Thakare, A. K. (1-2 December 2018). Document Segmentation and Language Translation Using Tesseract-OCR. IEEE 13th International Conference on Industrial and Information Systems (ICIIS) (p. 4). Rupnagar: IEEE.

[8] S. Lucas, "Icdar 2005 text locating competition results," 2005, in International Conference on Document Analysis and Recognition (ICDAR).

[9] A. Shahab, F. Shafait, and A. Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," 2011, in International Conference on Document Analysis and Recognition (ICDAR).

[10] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. de las Heras., "Icdar 2013 robust reading competition," 2013, in International Conference on Document Analysis and Recognition (ICDAR).

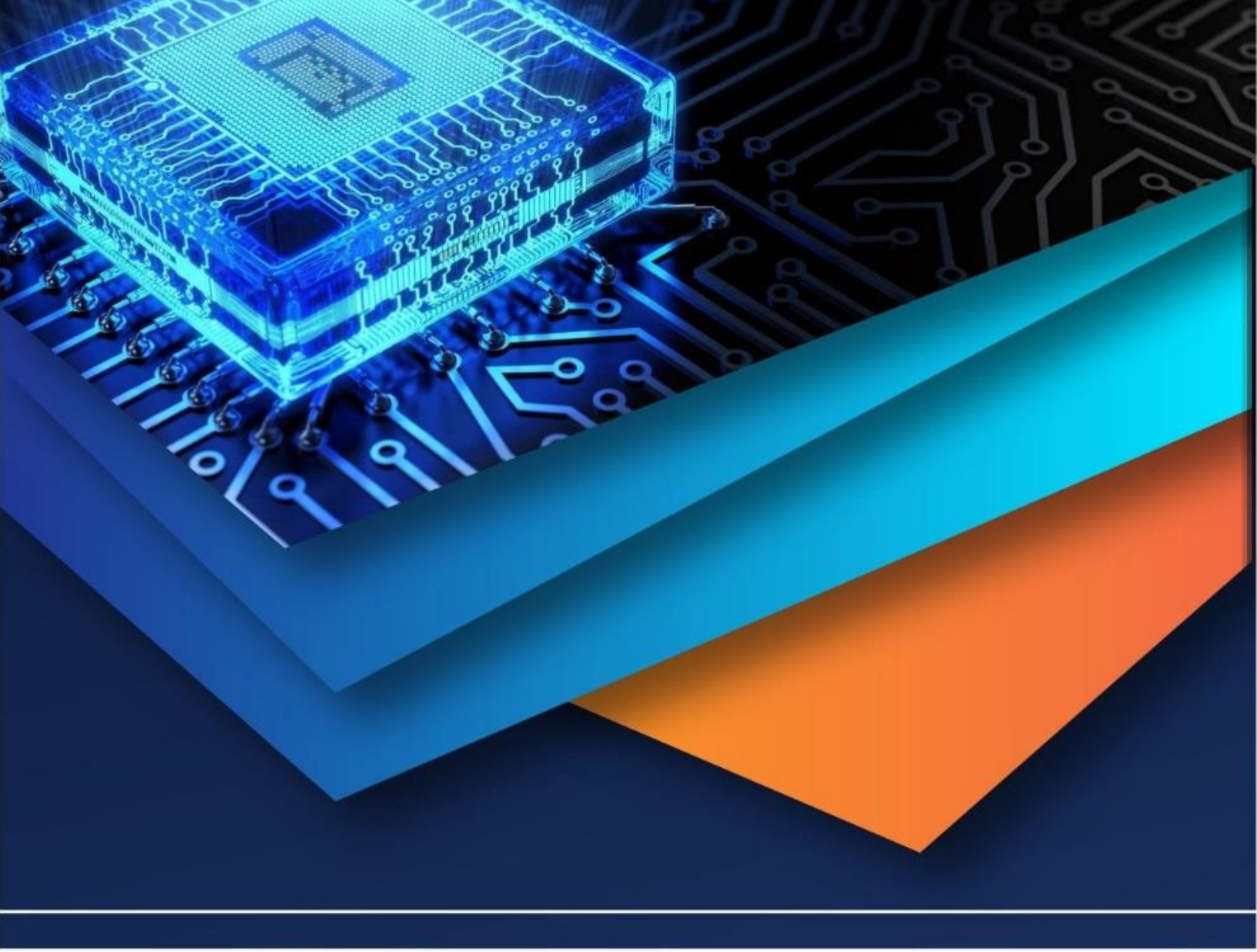
[11] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," 2012, in IEEE Computer Vision and Pattern Recognition (CVPR).

[12] Baoguang Shi, X. B. (29 December 2016). An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. IEEE, 9.

[13] K.N. Natei, J. V. (2018). Extracting Text from Image Document and Displaying Its Related Information. K.N. Natei Journal of Engineering Research and Application, 7.

[14] Pratik Madhukar Manwatkar, D. K. (2015). A Technical Review on Text Recognition from. IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO) (p. 5). Nagpur: IEEE.

[15] Max Jaderberg, K. S. (2016). Reading Text in the Wild with Convolutional Neural Networks. International Journal of Computer Vision, 20.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY