Text Recognition from Images: A Study

Sahana K Adyanthaya

Assistant Professor: Department of ECE

A. J. Institute of Engineering and Technology

Mangaluru, India

Abstract - **Recognition of text from images is an important process in the present scenario. Nowadays there is a great surge in storing the information found in the paper document in digital format. This helps in preserving the information, easy storage and also permits retrieval of information as and when required. The various stages of text recognition are: preprocessing, segmentation, feature extraction, classification and post processing. The preprocessing step involves a number of operations the most significant being converting color image into a binary image thus separating the text from the background. The segmentation step helps in separation of characters. The feature extraction enables to obtain the most pertinent information from the image to aid in text recognition. The process of classification enables to identify the text according to well defined rules. Then the post processing is done to reduce the errors. Text recognition is of utmost importance in several applications. This paper discusses the text recognition module and also presents the various applications of text recognition from images. Related work is also examined in this paper.**

Keywords - Text recognition; preprocessing; classification; post processing

I. INTRODUCTION

Text recognition has gained a lot of prominence in recent years as it has entered into a large arena of applications such as in automatic reading of license plate, signboards. Everything and everyone has gone digital these days. Most of the information transfers these days takes place via images or scanned documents. There is a huge amount of information being stored and also accessed at the same time.

Text recognition is a field which is driven by the need to preserve and have access to the information containing documents in an easier and quicker way. One of the convenient ways to transfer the information from the paper or books is to scan them which convert the information into an image thus preventing reuse of the scanned information in the form of a text. Thus it is necessary to develop tools to convert them into editable form. The purpose of this paper is to study various steps in text recognition process with an objective to convert text images into editable documents. One of the popular techniques used for text recognition is Optical Character Recognition (OCR). It converts scanned images of text into editable format.

Text recognition is a tedious job as it involves recognizing text of different fonts, styles and with different background noise. Also recognizing handwritten text is even more complicated due to differences in letter size, orientation and spacing between letters which varies from one person to another. Thus there is a need to develop an automated text recognition system which can identify the text component present in an image or scene and convert it into a machine recognizable format.

The process of text recognition starts with capturing the image of the required document, preprocessing it to acquire the desired portion and then segmenting it to extract the text content present in it. This paper discusses different stages in the task of text recognition from images.

II. LITERATURE REVIEW

This section provides a brief overview of the existing work carried out in the field of text recognition. Text recognition has been in existence since a very long time.

In [1], images with colorful background is considered and a preprocessing method is described which improves upon the performance of the Tesseract Optical Character Recognition (OCR) engine. Here first text segmentation is done to separate the text from the colorful background by dividing the original image into k images. Then a classifier recognizes the image containing text. There was an improvement of about 20% compared to the Tesseract OCR performance by employing preprocessing.

Work by S. Akopyan, O.V. Belyaeva, T.P. Plechov and D.Y. Turdakov [2] is based on a text extraction pipeline which is used to extract text from varied quality of images obtained from social media. Their work mainly focuses on dividing the input images into various classes and then preprocessing is done depending on the classes. This is followed by text recognition using OCR engine. The dataset collected from the social media is made use of in this work.

OCR is used to identify the text component present in images. In [3], the authors have proposed an algorithm to extract text from the scanned document. In this work, for segmentation Otsu's algorithm has been used and for skew detection Hough transform has been used. Also OCR technique has been applied to identify characters. They carried out the experiments and validated the proposed algorithm on various images taken from different sources. The average accuracy was found to be 93%.

K. Karthick, K.B. Ravindrakumar, R. Francis and S.Ilankannan [4] have discussed the various steps in text detection in detail highlighting the different techniques used

Special Issue - 2020

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
NCCDS - 2020 Conference Proceedings

for the same. They have also emphasized on handwritten text recognition which is one of the complex field. From their study it has been found that best results can be had with reduced computation time and it is possible to segment multilingual characters and enhance the character recognition rate.

Anupriya Shrivastava, Amudha J., Deepa Gupta and Kshitij Sharma [5] in their work have developed a system based on Convolutional Neural Network and Long Short-Term Memory. The developed model identifies the texts from images which is horizontal, curved or oriented style. The model has four components. The first component performs feature extraction at the low level. The second component uses shared convolution approach to extract high level features. Irrelevant feature are ignored by the third component. The fourth component predicts the character sequences.

Pratik Madhukar Manwatkar and Dr. Kavita R. Singh [6] have reviewed in their paper various methods to extract characters from images. Basic architecture of the system for text recognition from images is also described in their work. They have also discussed the sequence of image processing techniques to extract text from scanned image. Various fields of applications are also described in their paper.

III. TEXT RECOGNITION MODULE

The text recognition module has to perform a number of tasks. The input to the module is the image containing text. The output of the module is the text information in machine readable form. The text recognition module has to perform the following tasks: preprocessing, segmentation, feature extraction and classification. Fig. 1 shows the various tasks involved in text recognition.
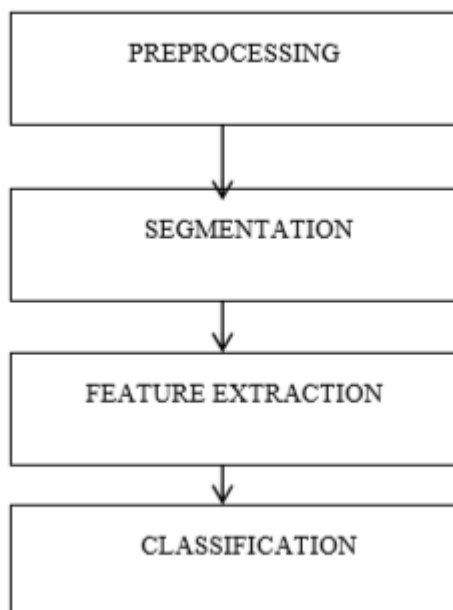

Fig. 1 Various tasks in text recognition

The scanned document is usually in the form of an image. The first step is preprocessing which is to convert the image into a format suitable for further processing. The text image may contain noise or it may be skewed. In this step the image is enhanced by noise removal and then converted to binary. The noise present in the image has a major role to play in successful text recognition. Noise removal increases the probability of accurate text recognition and generates more accurate output. Various filters such as Gaussian filter, mean filter can be used for noise removal. Then normalization is done to ensure uniformity which is followed by binarization to convert the gray image into a binary image.

After the preprocessing is done, the individual characters are separated using segmentation process. Then the vital data is retrieved from the raw data using the feature extraction step. Different techniques like Principle Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), Histogram etc. can be used for segmentation purpose [6].

The next step is classification which involves recognizing each character and allocating it to the right character class thus converting the text into machine readable form. Different classifiers based on Artificial Neural Network (ANN), Support Vector Machine (SVM) could be used for this purpose.

Post processing involves storing the recognized text in a format suitable for further processing.

IV. APPLICATION

In the recent years several applications of text detection have gained prominence. The major one being automated detection of text from the scanned document. Recent technological advances have further given a boost to the text recognition techniques. Text recognition enables automation in various fields. It can be used for automatic license plate reading at toll booths; automated reading of signatures on check leafs; image tagging and analysis of scene data.

Text recognition also enables automated storage and access of huge documents in the health care industry, offices. It enables creation of a database wherein the text can be searched and located easily. The created databases can be accessed and updated effortlessly which saves a lot of paper work. Automated text detection with voice assistance is also a boon for the visually impaired people. It can also be used for recognizing the text component in the videos.

Automated text identification can also be used to make the transport systems intelligent. It can also be used in airports for passport verification and information extraction. Text detection also enables automated data entry for business documents.

Text detection also makes a vast array of books to be available online which allows preservation and sharing of

Special Issue - 2020

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
NCCDS - 2020 Conference Proceedings

knowledge. Text recognition also enables automation in industries by supporting automated reading of labels, numbers. With the further advances in the technology text recognition has found a vast array of applications in almost every field.

V. SUMMARY

This paper presents a brief summary of various steps used in text recognition from images. The work carried out in this field has also been discussed briefly. A review of basic model of text recognition system is also given which describes the flow of text recognition from images. Finally the various fields where text recognition could be used are discussed.

REFERENCES

[1] Matteo Brisinello, Ratko Grbić, Dejan Stefanović and Robert Pečkai-Kovač, "Optical Character Recognition on images with colorful background", 2018, IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin).

[2] M.S. Akopyan, O.V. Belyaeva, T.P. Plechov and D.Y. Turdakov, "Text recognition on images from social media", 2019, Ivannikov Memorial Workshop (IVMEM).

[3] Neha Agrawal, Arashdeep Kaur, "An Algorithmic Approach for Text Recognition from Printed/Typed Text Images", 2018, 8th International Conference on Cloud Computing, Data Science & Engineering.

[4] K. Karthick, K.B. Ravindrakumar, R. Francis, S. Ilankannan, "Steps Involved in Text Recognition and Recent Research in OCR; A Study", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019.

[5] Anupriya Shrivastava, Amudha J., Deepa Gupta, Kshitij Sharma, "Deep Learning Model for Text Recognition in Images", 10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur, Kanpur, India.

[6] Pratik Madhukar Manwatkar, Dr. Kavita R. Singh, "A Technical Review on Text Recognition from Images", IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.