INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1814)
Available online at: https://www.ijariit.com

Smart Scan – Text Recognition System

Divya S.
rsdivya1999@gmail.com
New Horizon College of Engineering, Bengaluru, Karnataka

ABSTRACT

In today's world every information is in the digital form. Technology has made things much easier. Digitization is the process of transforming information into digital format. Ancient record, personal documents are scanned to store as images. Such scanned images of documents are modified and characters are recognized for further processing. Documents can contain either typed characters or handwritten characters. Extracting the characters in scanned images of documents helps for easier access to contents that need to be further modified. In this project, the models are built to recognize the printed text as well as handwritten text from images. Optical Character Recognition (OCR) technology can be used for printed text recognition and models using Artificial Neural Networks (ANN) can be created to extract handwritten characters. The mode of storage of the information plays a major role. This system allows the users to store the recognized text in desired formats.

Keywords—*Machine Learning, Optical Character Recognition, Neural Networks, Handwritten Text, Python*

1. INTRODUCTION

Digitization is transforming the available information into digital format. Technology today has widely spread across the world to obtain solutions for evolving problems and issues. There were times, when information was stored in the form of pen and paper. All such information loses its originality and perishes by time. In today's world, emerging technologies can be used to preserve such information. Ancient historical records, bank statements are scanned and stored to make it available for the future generation. In such cases, we can come across the form of the information. The characters in the documents can be printed or handwritten in case of past documents. This also helps in evaluating cheques and filled application forms.

Before any recognition process, images are to be pre-processed to avoid the noise in the images. Recognition of characters in the scanned documents can be performed using the machine learning and artificial intelligence models to get efficient results. Printed text usually has a format and makes the training process easier. Optical Text Recognition (OCR) model can be used for printed character recognition in images. Handwritten text appears in wide range of formats and consumes time to train the model. The Neural Network model can be implemented for handwritten text recognition. Also, it includes a large training dataset.

2. RELATED WORK

A. ***Shushant Chak, Ambalika Sharma, 2014, Handwritten Character Recognition using ANN, IJERT, Volume 03, Issue 07, July 2014.***

This journal talks about recognition of handwritten characters which is in current demand. The ability to efficiently process samples in applications like cheques and envelopes drives towards current research. In the approach, an artificial neural network is trained to identify similarities and patterns among different handwriting samples. This is for optimal handwritten English word recognition system based on character recognition. A number of techniques are available for feature extraction and training of CR systems in the literature, each with its own superiorities and weaknesses.

B. ***Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar, "Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014.***

This journal mainly focusses on the Optical Character Recognition and processes associated with the purpose – No more retyping, make searches digitally quicker, save space and edit text in a comfortable manner. Among the various steps, the pre-processing step for removing the noise or unnecessary details from the image. The character recognition step for processing the image obtained and includes binding relationship between the current image and the pre-trained images. Post processing is for converting the recognized text into desired file format and also adding the words to the local dictionary.

C. ***M. Zimmermann and H. Bunke, "Automatic segmentation of the IAM off-line database for handwritten English text," Object recognition supported by user interaction for service robots, 2002.***

This journal presents an automatic segmentation scheme for cursive handwritten text lines using the transcriptions of the text

lines and a hidden Markov model (HMM) based recognition system. The system uses IAM database that contains offline images of cursively handwritten English text. Applying the segmentation scheme on 417 pages of handwritten text a correct word segmentation rate of 98% has been achieved, producing correct bounding boxes for over 25,000 handwritten words.

3. METHODOLOGY

A. *Optical Character Recognition*

Optical Character Recognition or OCR is technically known as conversion of printed or handwritten text from the images into text format. Images that are scanned or captured as a photo image are used as the input for the system. It is emerged as a common method of digitizing the documents. Several models can be designed using python and machine learning algorithms. Tesseract OCR is one such open source model available for recognition of printed text from images.

B. *Neural Networks*

Neural Networks or also called as Artificial Neural Networks, are computing systems inspired by the neurons in human brain. They are a set of machine learning algorithms that can be used for solving and modelling complex data patterns. Some of the applications are forecasting, data validation, and recognition systems. In recognition systems, recognition and classification of characters from image is required.

C. HTR using Tensorflow

This is the most challenging problem. One of the solutions is using neural networks comprising of 5 convolutional NN (CNN) layers, 2 recurrent NN (RNN) layers and a Connectionist Temporal Classification (CTC) layer. Here the dataset used to train the neural network model is IAM dataset.

D. IAM Dataset

The IAM Handwriting Database contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments.

The database contains forms of unconstrained handwritten text, which were scanned at a resolution of 300dpi and saved as PNG images with 256 grey levels. All forms and also all extracted text lines, words and sentences are available for download as PNG files, with corresponding XML meta-information included into the image files. All texts in the IAM database are built using sentences provided by the LOB Corpus

The words have been extracted from pages of scanned text using an automatic segmentation scheme and were verified manually. All form, line and word images are provided as PNG files and the corresponding form label files, including segmentation information and variety of estimated parameters are included in the image files as meta-information in XML format which is described in XML file and XML file format (DTD).The database was first published at the ICDAR 1999.

D. OpenCV for Pre-processing

OpenCV is a Python cross-platform library which is used for real-time computer vision applications. It mainly focuses on image processing, video capture and analysis including features like face detection and object detection. It has various functions like image cropping and editing. It also allows to add filters to images for increasing the accuracy of the recognition systems.

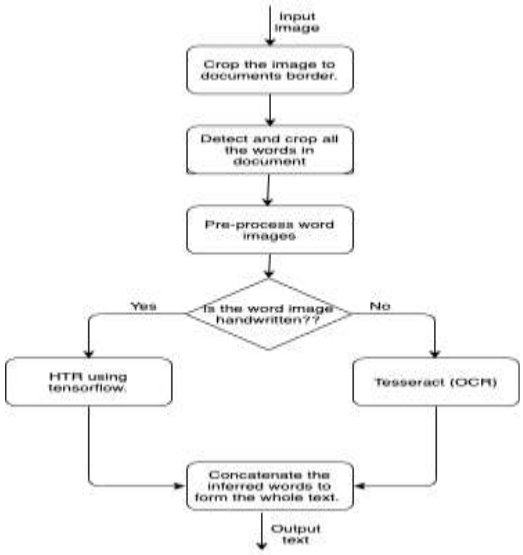

Fig. 1: Outline Text Recognition Process

4. PROJECT WORK AND RESULT

The entire project is built in Python and uses a graphical user interface. The interface built makes any user to interact and avail the services of the system in an effective manner. The images can be manipulated to make it more effective to view and for further operations. Any images that contain printed text can be processed to detect text using the tesseract OCR. The accuracy of printed text recognition ranges more than 95% in case of pre-processed image that is the image without any noise. The model also provides an option to search for a word or a phrase in the image of the printed text. When the word to be searched is given as input, if found the word is highlighted in the input image.

The model to recognize handwritten text in the images, it includes many steps using the neural networks model. First the model is trained with the IAM handwritten dataset which contains thousands of images written in different styles. Later the model is validated to check the validation criteria. The character error rate obtained is around 10% and word accuracy is around 75%. The system also allows the option to the user to save any modified image or text recognized to save into the local system in desired file format.

5. CONCLUSIONS AND FUTURE WORK

There is wide scope and future work of this system as many enhancements can be done for the system which we are implementing such as including different languages, to increase the accuracy rate of the handwritten text recognition. Along with these functionalities, this system can be enhanced in a way such that it can also has options to share recognized text and document. The system can be modified into a web application or an mobile application that attracts wide range of users.

6. ACKNOWLEDGMENT

We take the opportunity to thank all the people who have helped us in this project and whom we might not have mentioned here. Our guide for the most effective guidance and valuable support throughout the project session. And a great acknowledge to all us who were involved in this project. They always up front to motive and encourage us for bringing out this project.

7. REFERENCES

[1] Shushant Chak, Ambalika Sharma, 2014, Handwritten Character Recognition using ANN, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 03, Issue 07 (July 2014).

[2] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar, "Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014.

[3] M. Zimmermann and H. Bunke, "Automatic segmentation of the IAM off-line database for handwritten English text," Object recognition supported by user interaction for service robots, 2002.

[4] Ajinkya Khalwadekar, "Building OCR and Handwriting Recognition for document images", The Journal Blog.

[5] R. R. Ingle, Y. Fujii, T. Deselaers, J. Baccash and A. C. Popat, "A Scalable Handwritten Text Recognition System," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 17-24, doi: 10.1109/ICDAR.2019.00013.

[6] Muna Ahmed Awel, Ali Imam Abidi. "Review On Optical Character Recognition", International Research Journal of Engineering and Technology (IRJET), 2019.

[7] J. E. M. Adriano, K. A. S. Calma, N. T. Lopez, J. A. Parado, L. W. Rabago and J. M. Cabardo, "Digital Conversion Model For Hand-filled Forms Using Optical Character Recognition (OCR)", IOP Conference Series: Materials Science and Engineering, 2019.

