International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

Image to Text Conversion Using Tesseract

Nisha Pawar[1], Zainab Shaikh[2], Poonam Shinde[3], Prof. Y.P. Warke[4]

[1,2,3,4]*Dept. of Computer Engineering, Marathwada Mitra Mandal's Institute of Technology, Maharashtra, India*

Abstract - *In the current world, there is a great increase in the utilization of digital technology and various methods are available for the people to capture images. Such images may contain important textual data that the user may need to edit or store digitally. This can be done using Optical Character Recognition with the help of Tesseract OCR Engine. OCR is a branch of artificial intelligence that is used in applications to recognize text from scanned documents or images. The recognized text can also be converted to audio format to help visually impaired people hear the information that they wish to know.*

Key Words: **Artificial Intelligence, Optical Character Recognition, Tesseract, text-to-speech**

1. INTRODUCTION

Textual information is available in many resources such as documents, newspapers, faxes, printed information, written notes, etc. Many people simply scan the document to store the data in the computers. When a document is scanned with a scanner, it is stored in the form of images. But these images are not editable and it is very difficult to find what the user requires as they will have to go through the whole image, reading each line and word to determine if it is relevant to their need. Images also take up more space than word files in the computer. It is essential to be able to store this information in such a way so that it becomes easier to search and edit the data. There is a growing demand for applications that can recognize characters from scanned documents or captured images and make them editable and easily accessible.

Artificial intelligence is an area of computer science where a machine is trained to think and behave like intelligent human beings. Optical Character Recognition (OCR) is a branch of artificial intelligence. It is used to detect and extract characters from scanned documents or images and convert them to editable form. Earlier methods of OCR used convolutional neural networks but they are complicated and usually best suitable for single characters. These methods also had a higher error rate. Tesseract OCR Engine makes use of Long Short Term Memory (LSTM) which is a part of Recurrent Neural Networks. It is open-source and is more suitable for handwritten texts. It is also suitable at recognizing larger portion of text data instead of single characters. Tesseract OCR Engine significantly reduces errors created in the process of character recognition.

Tesseract assumes that the input image is a binary image and processing takes place step-by-step. The first step is to recognize connected components. Outlines are nested into blobs. These blobs are organized into text lines. Text lines are broken according to the pitching. If there is a fixed pitch between the characters then recognition of text takes place which is a two-pass process.

An adaptive classifier is used here. Words that are recognized in the first pass are given to the classifier so that it can learn from the data and use that information for the second pass to recognize the words that were left out in the previous pass. Words that are joined are chopped and words that are broken are recognized with the help of A* algorithm that maintains a priority queue which contains the best suitable characters. Then, the user can store this information in their computers by saving them in word documents or notepads that can be edited any time they want.

It is difficult for visually impaired people to read textual information. Blind people have to make use of Braille to read. It would be easier for them to simply listen to the audio form of the data. This application can be used to convert textual data to audio format so that it is easier for people to hear the information. Google Text-To-Speech API is used to convert the text information into audio form.

2. EXISTING SYSTEM

There are various methods for OCR. Some of them are:

2.1 Connected components based method

It is a well known method used for text detection from images. The connected components are extracted with help of algorithm. The resulting components are then partitioned into clusters. This approach detects pixel differences between the text and the background of the text image. It can extract and recognize the characters too.

2.2 Sliding window based method

This method is also known as text binarisation process. It classifies individual pixels as text or background in the textual images. The method acts as bridge between localization and recognization by OCR.

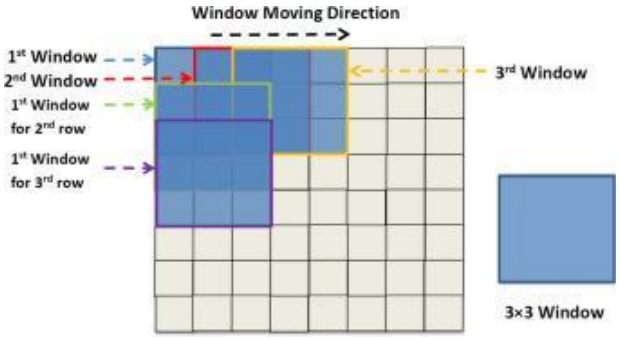

Fig -1: Sliding window based method

2.3 Hybrid method

Hybrid method is used for text classification. This approach detects and recognizes texts in CAPTCHA images. The strength of CAPTCHA can be checked. This method efficiently detects and recognizes the text with a low false positive.

Fig -2: Captcha

2.4 Edge based method

This method is also known as image processing technique, which finds boundaries of the images or any other objects within the images. It works by detecting discontinuities in brightness. This approach is also used for image segmentation and data extraction in areas such as image processing, machine vision and computer vision.

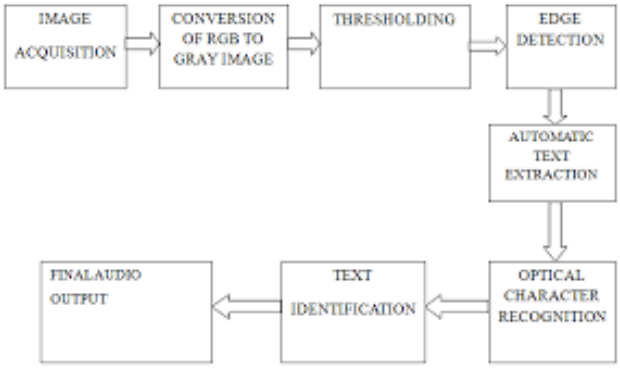

Fig -3: Edge-based method

2.5 Color based method

Color based approach is used for clustering. It consists of two phases: text detection phase and text extraction phase. In text detection phase, two features are considered – homogeneous color and sharp edges, and color based clustering is used to decompose color edge map of image to several edge maps, which makes text detection more accurate. In text extraction phase, the difference between the text and background in image is considered.

2.6 Texture based method

It is another approach used for detecting texts in images. This approach uses Support Vector Machine (SVM) to analyze the textural properties of texts. This method also uses continuously adaptive mean shift algorithm (CAMSHIFT) that results in texture analysis. It combines both SVM and CAMSHIFT to provide robust and efficient text detection.

2.7 Corner based method

Corner based method is used for text extraction method. It has three stages - a) Computing corner response in multi-scale space and thresholding it to get the candidate region of text; b)verifying candidate region by combining color and size range features and; c) locating the text line using bounding box. It is two-dimensional feature point which has high curvature in region boundary.

2.8 Stroke based method

This approach is used to detect and recognize text from the video. It uses text confidence using an edge orientation variance and opposite edge pair feature. The components are extracted and grouped into text lines based on text confidence maps. It can detect multilingual texts in video with high accuracy.

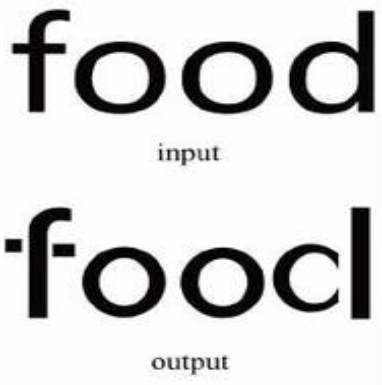

Fig -4: Stroke-based method

2.9 Semi automatic ground truth generation method

This approach is also used for detecting and recognizing text from the videos. It can detect English and Chinese text of different orientations. It has attributes like: line index, word index, script type, area, content, type of text and many more. It is most efficient method to detect texts from the videos.

3. LITERATURE REVIEW

Text detection from images is useful in many real world applications. The data that is stored in text is huge and there is need to store this data in such a manner that can be searched easily whenever required. Elimination of the use of paper is one of the steps to progress towards a world of electronics. Also, data that can be converted to audio form is a way to ease the lives of visually impaired people.

In [2007] Ray Smith published an overview of Tesseract OCR Engine. It stated that Tesseract started as a PhD project sponsored by HP in 1984. In 1987, a second person was assigned to help build it better. In 1988, HPLabs joint with Scanner Division project. In 1990, the scanner product was cancelled and four years ahead HPLabs project was cancelled too. From the year 1995 till the year 2005, Tesseract was in its dark ages. But in the year 2005, it was open sourced by HP. In 2006, Google took over it. In 2008, Tesseract expanded to support six languages. By the year 2016, it was developed further to makes use of LSTM for the purpose of OCR.

In [2015] Pratik Madhukar Manwatkar and Dr. Kavita R. Singh published a technical review on text recognition from images. It emphasizes on the growing demand of OCR applications as it necessary in today's world to store information digitally so that it can be edited whenever required. This information can later be searched easily as it is in digital format. The system takes image as an input, processes on the image and the output is in the form of textual data.

In [2016] Akhilesh A. Panchal, Shrugal Varde and M.S. Panse proposed a character detection and recognition system for visually impaired people. They focused on the need of people that are visually impaired as it is difficult for them to read text data. This system can be used to extract text data from shop boards or direction boards and convey this information to the user in audio form. The main challenges are the different fonts of the texts on the natural scene images.

In [2017] Nada Farhani, Naim Terbeh and Mounir Zrigui published a paper that stated the conversion of different modalities. Human beings have different modalities such as gesture, sound, touch and images. It is vital that they can convert the information between these modalities. The paper focuses on conversion of image to text and also on text-to-speech so that the user can hear the information whenever required.

In [2017] Azmi Can Özgen, Mandana Fasounaki and Hazım Kemal Ekenel published a paper that stated how text data can be extracted from both natural scene images and computer generated images. They make use of Maximally Stable Extermal Regions for the purpose of text detection and recognition. This method eliminated the non-text part of the image so that OCR can be done more efficiently.

In [2018] Sandeep Musale and Vikram Ghiye proposed a system that is a smart reader for visually impaired people. Using this system, they can convert the text information to audio format. This system has an audio interface that the people with visual problems can use easily. It uses a combination of OTSU and Canny algorithms for the purpose for character recognition.

In [2018] Christian Reul, Uwe Springmann, Christoph Wick and Frank Puppe proposed a method to reduce the errors generated during OCR process. It includes cross fold training and voting to recognize the words more accurately. As LSTM is introduced now, it is easier to recognize words of old printed books, handwritten words, blurry or uneven words with high accuracy. A combination of ground truths and confidence values are used in this method for optimal recognition of the characters.

4. CONCLUSION AND FUTURE SCOPE

In this age of technology, there is a huge amount of data and it keeps on increasing day by day. Even though much of the data is digital, people still prefer to make use of written transcripts. However, it is necessary to store this data in digital format in computers so that it can be accessed and edited easily by the user. This system can be used for character recognition from scanned documents so that data can be digitalized. Also, the data can be converted to audio form so as to help visually impaired people obtain the data easily.

In the future, we can expand the system to that is can recognize more languages, different fonts and also handwritten notes. Various accents can also be added for audio data.

REFERENCES

[1] Ray Smith, "An overview of the tesseract OCR engine," 2005.

[2] 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.

[3] Akhilesh A. Panchal, Shrugal Varde, M.S. Panse, "Character detection and recognition system for visually impaired people," IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 20-21, 2016.

[4] Nada Farhani, Naim Terbeh, Mounir Zrigui, "Image to text conversion: state of the art and extended work," IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017.

[5] Azmi Can Özgen, Mandana Fasounaki, Hazım Kemal Ekenel, "Text detection in natural and computer-generated images," 2017.

[6] Sandeep Musale, Vikram Ghiye, "Smart reader for visually impaired," Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7.

[7] Christian Reul, Uwe Springmann, Christoph Wick, Frank Puppe, "Improving OCR accuracy on early printed books by utilizing cross fold training and voting," 13th IAPR International Workshop on Document Analysis Systems, 2018.

[8] U. Springmann and A. Ludeling, "OCR of historical ¨ printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus," Digital Humanities Quarterly, vol. 11, no. 2, 2017.

[9] J. C. Handley, "Improving OCR accuracy through combination: A survey," in Systems, Man, and Cybernetics, IEEE, 1998.

[10] F. Boschetti, M. Romanello, A. Babeu, D. Bamman, and G. Crane, "Improving OCR accuracy for classical critical editions," Research and Advanced Technology for Digital Libraries, pp. 156–167, 2009.

[11] Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru, "Show and tell: A neural image caption generator", In CVPR, 2015.

[12] Rafeal C. Ginzalez and Richard E. Woods, "Digital Image Processing", Pearson Education, Second Edtition, 2005.