International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

EXTRACTION AND DETECTION OF TEXT FROM IMAGES

Anshul Arora
Department of Computer Engineering
Bharati Vidyapeeth
(Deemed to be University)
College of Engineering, Pune

Rajat Singh
Department of Computer Engineering
Bharati Vidyapeeth
(Deemed to be University)
College of Engineering, Pune

Ashiq Eqbal
Department of Computer Engineering
Bharati Vidyapeeth
(Deemed to be University)
College of Engineering, Pune

Ankit Mangal
Department of Computer Engineering
Bharati Vidyapeeth
(Deemed to be University)
College of Engineering, Pune

Prof. S.U Saoji
Department of Computer Engineering
Bharati Vidyapeeth
(Deemed to be University)
College of Engineering, Pune

Abstract : In today's world there has been a huge increase in the use of digital technology and various methods of capturing images are available. These images may contain important textual data that the user may need to digitally edit or archive. This can be done by optical character recognition using the Tesseract OCR engine. The main important concept behind this technology is something called OCR - Optical Character Recognition. Using OCR, we can search and recognize text in electronic documents and easily convert it to readable text. It converts the text of electronic documents into a relative ASCII character and, if the document is handwritten, OCR uses the database to recognize which character it is and resolve it with the utmost precision. In this article we have discussed and analyzed different methods for recognizing text from images. The purpose of this review article is to summarize the known methods for a better understanding of the reader.

Keywords- OCR, recognized, Tesseract, IDA.

Introduction

Nowadays, there is a growing demand for software systems to recognize characters on the computer when information is scanned through paper documents, because we know that there are various historical and mythological books and magazines in print. Every day they are damaged by changes in the atmosphere or improper use. Therefore, today there is a strong demand to "save the information available in these paper documents to a computer storage drive and reuse it later through a search process." An easy way to store information from these paper documents on a computer system is to scan the documents first. Whenever we scan documents through the scanner, the documents are saved as images in the computer system. These images contain text that the user cannot edit. However, to reuse this information, it is very difficult for the computer system to read the individual contents and search the contents of these documents line by line and word by word. The reason for this difficulty is that the characteristics of paper documents differ from those of the computer system. This prevents the computer from recognizing characters while reading. This concept of storing the contents of paper documents in a computer storage location and then reading and searching for the contents is known as document processing. From time to time in developing this document we may have to deal with information related to languages other than English around the world. This process is also known as document image analysis (DIA). Researchers have proposed many approaches to manage IDA in recent years. Presentation of the text recognition system

In this section we briefly describe the general architecture of the text recognition system, as shown in Figure 1. A text recognition system receives input in the form of an image containing text information. The output of this system is in electronic format, which means that the textual information of the image is stored in a computer readable format. The text recognition system can be divided into the following modules: (A) Pre-processing (B) Text recognition (C) Post-processing. Each module is described below:

A. Pre-treatment module

The optical scanner typically scans the paper document and converts it into an image. An image is the combination of elements of the image, also known as pixels. At this point, we have the data as an image and this image can be further analyzed so that we can retrieve the important information. Therefore, to improve thequality of the input image, only some image enhancement operations are performed, such as. Noise reduction, normalization, binarization, etc.

OCR software often "pre-processes" images to improve the chances of successful recognition. Techniques include:[15]

Skewed: If the document was skewed when scanning, you may need to tilt it a few degrees clockwise or

counterclockwise to make the lines of text be perfectly horizontal or vertical.

• Eliminate smudging: removes positive and negative points, softens edges

• Binarization - Converts a color or grayscale image to black and white (called a "binary image" because there are two colors). The binarization task is an easy way to separate text (or any other image component you want) from the background. [16] The binarization task itself is necessary because most commercial recognition algorithms only work on binary images, as it turns out to be simpler. [17] Furthermore, the efficiency of the binarization step significantly influences the quality of the character recognition step, and careful decisions are made when choosing which binarization to use for a particular type of. input image; since the quality of the binarization method used to obtain the binary result depends on the type of input image (scanned document, scene text image, historically degraded document, etc.).

Line removal – Cleans up non-glyph boxes and lines

• Layout or "zoning" analysis: identify columns, paragraphs, captions and more. as separate blocks. Especially important in tables and multi-column layouts.

• Word and Line Detection: Establishes a baseline for word and character shapes, separating words as needed.

• Script Recognition: In multilingual documents, the script can change at the word level, and therefore, script identification is required before the appropriate OCR can be called to handle the specific script.

• Character isolation or "segmentation" - For character-by-character OCR, multiple linked characters must be separated due to image artifacts; Unique characters divided into multiple pieces by artifacts must be linked together.

• Normalize aspect ratio and scale

A. Text recognition

There are two basic types of basic OCR algorithms, which can produce a classified list of candidate characters. [2. 3]

Matrix matching involves comparing one image pixel by pixel with a stored glyph; it is also known as "pattern matching", "pattern recognition" or "image correlation". This is based on the fact that the input glyph is well isolated from the rest of the image and that the saved glyph has a similar font on the same scale. This technique works best with typed text and does not work well when new characters are detected. This is the technique implemented directly by the first photocell-based physical OCRs.

Function extraction splits glyphs into "attributes" such as lines, closed loops, line direction, and line intersections. The extraction functions reduce the dimensionality of the representation and make the recognition process computationally efficient. These attributes are compared to an abstract vector representation of a character, which can be reduced to one or more glyph prototypes. Common feature detection techniques in computer vision apply to this type of OCR, which is commonly seen in "smart" handwriting recognition and indeed in most modern OCR software. More stringent classifications, such as the k nearest neighbors algorithm, are used to compare image attributes with stored glyph attributes and choose the closest match.

Software such as Cuneiform and Tesseract use a two-step approach to character recognition. The second step is known as "adaptive recognition" and uses the letter shapes recognized with great confidence in the first step to better recognize the remaining letters in the second step. This is useful for unusual fonts or low quality scans where the font is distorted (such as fuzzy or faded).

Modern OCR software, such as OCRopus or Tesseract, uses trained neural networks to recognize entire lines of text instead of focusing on individual characters.

A new technique known as iterative OCR automatically cuts a document into sections based on the page layout. OCR is performed on individual sections using variable character confidence level thresholds to maximize page-level OCR precision. The United States Patent Office has granted a patent for this method [26].

The OCR result can be saved in the standardized HIGH format, a special XML schema maintained by the US Library of Congress. Other commonly used formats are hOCR and PAGE XML**.**

B. Post-processing form

The output of the text recognition engine is in the form of text data that the computer can understand. Therefore, it should be saved in a suitable format (for example, text or MS-Word) for later use, for example to Text recognition techniques

Many techniques have been developed from various studies on text and image recognition. All of these techniques are roughly divided into three basic techniques:

1) Method based on textures: this method uses the properties of textures based on the Fourier transform, local intensity, filter response and wavelet coefficients to distinguish the text part from the non-text part of natural images .

2) Region Based Method: This method uses properties such as color, intensity, and border matching to distinguish between text and non-text parts in natural images.

It is divided into three types:

Contour Based: - This method uses an edge detector. Operator to detect the edges of images.

In general, two types of edge detection methods are used, such as the Canny operator and Soble Edge. Connected **components**: - This method identifies character components using edge detection and grouping methods. One of the most important techniques in this method is the extremely stable maximum area.

Lineweight: This method uses text functions. It can be recognized by the indent of the coins. Constant line art components are treated as text and treated as if they were not text. The bar width transform operator is used in the text for this operation.

Hybrid method: pushing the limits of any combination of the above techniques Two or more of the so-called hybrid techniques are used Technology

Literature review

In [2007], Ray Smith published an overview of the Tesseract OCR engine. He claimed that Tesseract started in 1984 as an HP sponsored doctoral project. In 1987 a second person was appointed to help with the implementation of the project. In 1988, HPLabs joined the Scanner Division project. In 1990 the scanner was discontinued and four years earlier the HPLabs project was abandoned. From 1995 to 2005, Tesseract was in a dark period. But in 2005 it was open source from HP. In 2006, Google took over. In 2008, Tesseract expanded to six languages.

In [2015] Pratik Madhukar Manwatkar and Dr. Kavita R. Singh a magazine specializing in the recognition of text from images. The growing demand for OCR applications is highlighted because in today's world information must be stored in digital format so that it can be edited as needed. This information can be easily viewed later as it is in digital format. The system takes the image as input, processes the image, and the output is in the form of text data. In 2016, it was developed to use LSTM for OCR purposes.

In 2016, Akhilesh A. Panchal, Shrugal Varde and M.S. Panse have proposed a system of recognition andrecognition of signs for the blind. They focused on the needs of blind people as they have difficulty reading textual data. This system allows you to extract textual data from memory cards or addresses and transmit this information to the user in audio form. The main challenges are the different sources of the texts in the images of the natural scene.

In 2017, Nada Farhani, Naim Terbeh and Mounir Zrigui published an article advocating the conversion of different modalities. People have different modalities such as gestures, sounds, touch and images. It is important to convert information between these modes. The document focuses on converting images to text and converting text to speech so that the user can hear the information when they need it.

APPLICATION-

Text recognition technology can be used in all industries and is revolutionizing the document management process. With this technology, scanned documents can become more than image files and become documents with textual content that computers can recognize and search. This technology eliminates the need for users to manually enter important documents when accessing electronic databases. Instead, the text recognition system extracts the relevant information and inserts it automatically. The result thus obtained is precise and the effective processing time of the information is much shorter.

bank

The use of image text recognition varies by region. A well-known application is in the banking industry. It allows checks to be processed without human intervention. A check can be typed, the text it contains is immediately scanned, and the correct amount is transferred. This technology is nearly perfect for paper exams and is also quite accurate for handwritten exams, although manual confirmation is sometimes required. In general, this will reduce waiting times at many banks.

Legal

Significant progress has also been made in the legal field in the digitization of paper documents. To save space and avoid searching in paper archive boxes, documents are scanned and entered into computer databases. Image text recognition further simplifies the process by allowing you to search for text in documents, making it easier to find and use in the database. Lawyers now have quick and easy access to a large library of electronic documents that you can find by entering a few keywords.

Healthcare

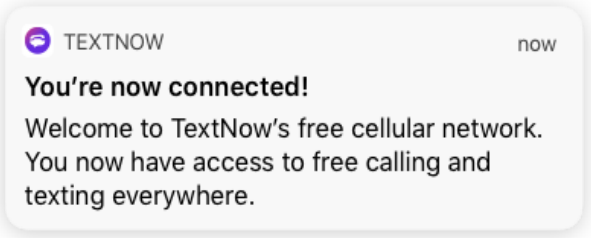
Healthcare also uses image recognition technology. Process paper documents. Healthcare professionals should forever Process large numbers of forms for each patient, including insurance and general health forms. the following with all this information follows; It is useful

to enter the relevant data into an accessible electronic database if necessary. They can extract files using image recognition technology. Extract information from forms and enter it into databases so that all patient data is captured instantly. As a result, healthcare providers canfocus on providing the best possible service to each patient.

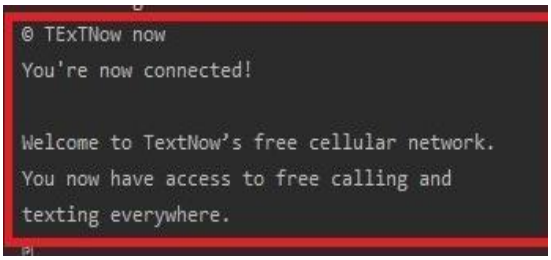
RESULT-

Based on the techniques described earlier in the upper section , we will write a code to recive an image as input and will provide text as the output.

Image which we sent as input-



Output on the console-



SUMMARY

In this article, we have examined and discussed various methods of finding text characters from scene frames. We have reviewed the basic architecture for recognizing text from images. In which we discuss various image processing techniques in a specific order to recognize text from a scanned image. We also discuss some applications of the text recognition system.

References-

[1] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and JingXu, "A fast adaptive binarization method for complex scene images," 19th IEEE International Conference on Image Processing (ICIP), 2012.

[2] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N-Grams," IEEE, 2012.

[3] Gur, Eran, and ZeevZelavsky, "Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic," IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012.

[4] Rhead, Mke, "Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems." IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[5] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." IEEE International Conference on Document Analysis and Recognition, 2012.

[6] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network," IEEE, 2012.

[7] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep Features for Text Spotting_jaderberg14."

[8] Zhang, Z., Zhang, C., Shen, Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-Oriented Text Detection with Fully Convolutional Networks," Cvpr, pp. 4159–4167,2016.

[9] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao, "Scene Text Detection via Holistic, Multi-Channel Prediction," pp. 1–10, 2016.

[10] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9912 LNCS, pp. 56–72, 2016.

[11] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," pp. 4161–4167, 2016.

[12] Y. Qu, X. Yang, and L. Lin, "Scene text detection with text statistical characteristics and deep neural network," Commun.Comput. Inf. Sci., vol. 773, pp. 245–254, 2017.

[13] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. NotesBioinformatics), vol. 8692 LNCS, no. PART 4, pp. 497–511, 2014

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

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