Volume 23 Issue 1 2022 ***CENTRAL ASIA AND THE CAUCASUS*** ***English Edition***

From Text Image To Text Document

Yan Puspitarani
Viddi Mardiansyah

DOI: https://doi.org/10.37178/ca-c.23.1.111

***Yan Puspitarani**, Universitas Widyatama*
Email: yan.puspitarani@widyatama.ac.id

***Viddi Mardiansyah**, Universitas Widyatama*
Email: viddi.mardiansyah@widyatama.ac.id

Abstract

The storage of printed documents such as letters in digital form will not be meaningful if documents containing these characters are only stored in the form of images. The conversion of images into character collections using Optical Character Recognition (OCR) is needed so that information in documents can be further processed to produce knowledge. There are so many processes that must be passed an image in Optical Character Recognition (OCR). This paper will discuss the processes that occur in Optical Character Recognition in theory.

***Index Terms** digital documents, images, Optical Character Recognition, stage of Optical Character Recognition*

INTRODUCTION

A dvances in technology in the digital age make the need for digital data storage is even greater. The documents in the form of printed letters will be stored in digital form using a scanner. However, storing documents in the form of images will be difficult to reprocess to produce meaningful knowledge. Therefore we need a process that can identify letters in the image so that it can be stored in a database to be processed to the next process such as knowledge management systems, data mining, business intelligent, and so on. This identification process is known as Optical Character Recognition (OCR).

Optical character recognition (OCR) is the process of converting letter images into ASCII characters that are recognized by the computer. The letter images in question can be in the form of document scans, print-screen results of web pages, photographs, and so on[1]. Basic principles OCR imitate the way humans are reading: visually scanning a objects that contain text, process these objects, and interpret text contained in the object. This text is then stored in digital form [2].

This paper will explain the stages of Optical Character Recognition with some popular algorithm based on literature review.

Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is a PC application that is utilized to recognize the picture of letters and numbers to be changed over into composing records. This letter acknowledgment framework can expand the adaptability or capacity and knowledge of a PC framework. The keen letter acknowledgment framework is extremely useful for the enormous scope business that is at present being done by many gatherings, specifically the work of digitizing data and information, for instance in making computerized library assortments, advanced antiquated abstract assortments, and others.

The stages of the OCR System are as follows [3]:

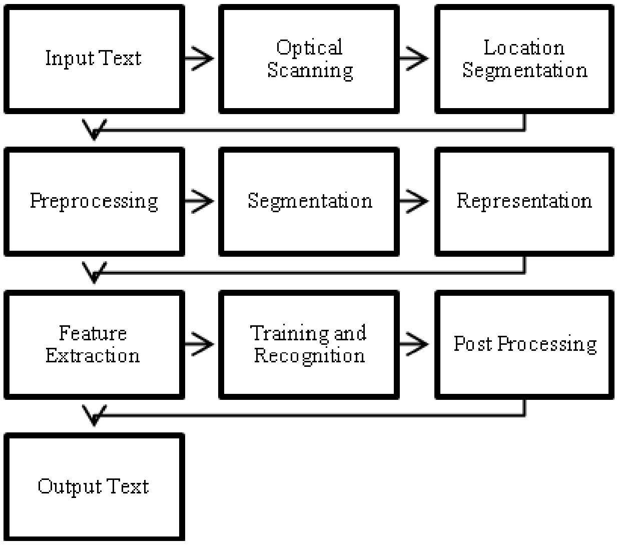

Fig. 1. The stages of Optical Character Recognition (OCR)

The initial step is to digitize simple report utilizing an optical scanner. At the point when locales containing text are found every image is separated through division measure. The removed images are pre-handled, taking out clamor to work with include extraction. The personality of every image is found by contrasting removed provisions and depictions of image classes got through a past learning stage. At long last context oriented data is utilized to remake words and quantities of the first text.

A. Optical Scanning

The primary part in OCR is optical scanning. Through examining measure computerized picture of unique report is caught. When performing OCR staggered picture is changed over into bi-level high contrast picture. This cycle known as thresholding is performed on scanner to save memory space and computational exertion. The thresholding system is significant as the aftereffects of acknowledgment are absolutely subject to nature of bi-level picture.

B. Location Segmentation

The following OCR part is Location Segmentation. Segmentation have printed information and are recognized from figures and illustrations. For instance, when performing programmed mail figuring out envelopes address should be found and isolated from different prints like stamps and friends logos, before acknowledgment.

When applied to message, division is disengagement of characters or words. The fundamental issues in division are: (a) extraction of contacting and divided characters (b) recognizing commotion from text (c) confounding illustrations and calculation with text as well as the other way around.

C. Preprocessing

The third OCR part is preprocessing. The crude information relying upon the information procurement type is exposed to various fundamental preparing steps to make it usable in the engaging phases of character examination. The picture coming about because of checking interaction might contain certain measure of clamor. The principle goals of pre-handling can be pointed as [3, 4] : (a) commotion decrease (b) standardization of the information and (c) pressure in the measure of data to be held.

The initial step is the transformation of any sort of picture into a Paired picture (the one having pixel esteems as '0' and '1' in particular). 'Binarization' changes over any picture into a progression of Dark text composed on a White foundation. Along these lines, it actuates consistency to every one of the info pictures. Different impacts like difference, sharpness and so forth can likewise be effortlessly taken care of once the picture has been binarized. The ANN utilized in the framework utilizes 'Element Vectors' as its feedback. Thus, each character is fragmented out from the pre-prepared picture. This division happens in two stages. To begin with, each line is isolated in the info picture. Then, at that point each character is isolated out in each line. It could be noticed that the progression of choosing out a 'Up-and-comer Square' is required where just a piece of picture contains 'text' which should be recognized[5].

Normalization is done to adjust the input image data with image data in the database. The normalization process is adjusted to the needs of the recognition process used. One of the simplest process of normalization is image size normalization [6].

D. Segmentation

The preprocessing stage yields a perfect person picture as in an adequate measure of shape data, high pressure, and low commotion on a standardized picture is acquired. The following OCR part is division. Here the person picture is sectioned into its subcomponents. Division is significant on the grounds that the degree one can reach in partition of the different lines in the characters straightforwardly influences the acknowledgment rate. Inside division is utilized here which detaches lines and bends in the cursively composed characters. However a few noteworthy strategies have created previously and an assortment of procedures have arisen, the division of cursive characters is a perplexing issue. The person division methodologies are partitioned into three classes: (a) explicit segmentation (b) implicit segmentation and (c) mixed strategies.

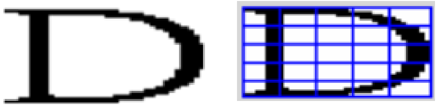

Fig. 2 Character Segmentation with 25 segment

E. Representation

The fifth OCR part is portrayal. The picture portrayal plays one of the main jobs in any acknowledgment framework. In the most straightforward case, dark level or twofold pictures are taken care of to a recognizer.

The person picture portrayal techniques are by and large sorted into three significant gatherings:

a. global change and series extension. The normal strategy that is utilized in this gathering are fourier change, gabor change, wavelets, minutes and karhunen loeve development.

b. statistical portrayal. A portion of the major measurable components utilized for character portrayal are drafting, intersections and distances, and projections.

c. geometrical and topological portrayal. The topological and mathematical portrayals can be assembled into removing and counting topological constructions, estimating and approximating the mathematical properties, coding, and charts and trees.

F. Feature Extraction

The 6th OCR part is feature extraction. The goal of component extraction is to catch fundamental attributes of images. Component extraction is acknowledged as one of the most troublesome issues of example acknowledgment. The most straight forward method of portraying character is by genuine raster picture. Another methodology is to extricate certain components that portray images yet leaves the insignificant qualities. The strategies for extraction of such provisions are separated into three gatherings : (a) appropriation of focuses (b) changes and series developments and (c) underlying examination.

Feature Extraction fills two needs; one is to separate properties that can recognize a person interestingly. Second is to extricate properties that can separate between comparable characters [7]. A bunch of various sorts of elements has been utilized to distinguish the characters, in our calculation. These incorporate Amount of pixels along the even lines drawn at different distances along the person stature as displayed in Figure 3. These boundaries vary starting with one person then onto the next dependent on its width profile variety along the stature.

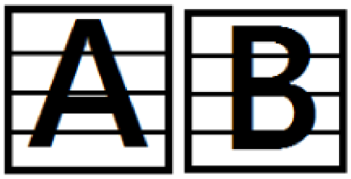

Fig. 3 Horizontal Lines at various heights

Additionally, a bunch of vertical lines drawn at different distances along the width, portraying the amount of pixels, can likewise fill in as one more list of capabilities, as displayed in Figure 4.

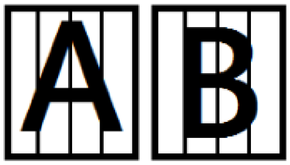

Fig. 4 Vertical Lines through the width

Symmetry is one more boundary that can be utilized to decrease equivocalness among characters, for example, '8' and 'B' can be separated dependent on its Level Evenness while 'I' and 'J' can be separated handily dependent on their Upward Balance. It ought to be noticed that these boundaries show the 'Level of balance', for example a decimal worth between 0 (No evenness) to 1 (Amazing balance), as opposed to 'Valid' or 'Bogus'. For this, we make a network, say M having the main half (even or vertical) part to be the identical representation of the subsequent half. Then, at that point, the relationship is found among 'M' and 'I'. This degree of connection gives us the measure of balance the person has.

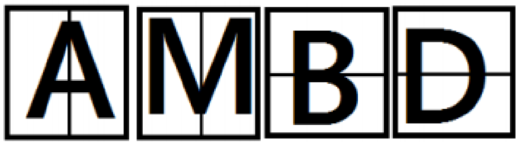

Fig. 5 Horizontal & Vertical Symmetry in characters

One more worldview of character acknowledgment is the quantity of shut regions in its shape. Characters, for example, 'A','P','D' and 'Q' have one shut region, while others, for example, 'B' and '8' have two. There likewise exist characters which are open, for example, 'H', '7', 'C' and so on This boundary additionally serves to comprehensively arrange characters dependent on its transparency or closeness.

The fundamental thought behind computing various boundaries is to expand the distinctions among the characters, in order to make the acknowledgment simpler.

G. Training and Recognition

The seventh OCR part is preparing and acknowledgment. OCR frameworks widely utilize the techniques of example acknowledgment which allots an obscure example into a predefined class. The OCR are explored in four general methodologies of example acknowledgment as proposed in [3, 4]: (a) template matching (b) statistical techniques (c) structural techniques and (d) ANNs.

The use of ANN and Template Matching as the most widely used recognition model and produce the best recognition rate. Character recognition with the Arial font results in an recognition rate of 92.2% for ANN and 94.6% for Template Matching . Meanwhile the worst results with an recognition rate of 33.33% for ANN on the Georgian font but the recognition rate reaches 100% when using SOM Neural Network [7].

H. Post Processing

The eighth OCR part is post-processing. A portion of the normally utilized post-handling exercises incorporate gathering and blunder location and rectification. In gathering images in text are related with strings. The aftereffect of plain image acknowledgment in text is a bunch of individual images. Notwithstanding, these images don't typically contain sufficient data. These singular images are related with one another making up words and numbers. The gathering of images into strings depends on images' area in record. The images which are adequately close are gathered together. For textual styles with fixed pitch gathering measure is simple as position of each character is known. For typeset characters distance between characters are variable. The distance between words are altogether enormous than distance among characters and gathering is subsequently conceivable. The issues happen for manually written characters when text is slanted. Until gathering each character is dealt with independently, the setting where each character seems has not been taken advantage of. Be that as it may, in cutting edge optical text acknowledgment issues, framework comprising just of single person acknowledgment isn't adequate. Indeed, even best acknowledgment frameworks won't give 100% right ID, all things considered. Just a portion of these blunders are distinguished or revised by the utilization of setting. There are two primary methodologies. The first uses the chance of groupings of characters showing up together. This is finished by utilizing rules characterizing grammar of word. For various dialects the probabilities of at least two characters showing up together is succession can be figured and is used to identify mistakes. For instance, in English language likelihood of k showing up after h in a word is zero and if such a mix is identified a mistake is expected. Another methodology is word references use which is most effective blunder location and revision technique. Given a word where a blunder is available and the word is gazed upward in word reference. On the off chance that the word isn't in word reference a blunder is recognized and is rectified by changing word into most comparable word. The probabilities got from order assists with recognizing character wrongly grouped. The mistake changes word starting with one legitimate word then onto the next and such blunders are imperceptible by this technique. The drawback of word reference techniques is that inquiries and correlations are tedious.

II. OPTICAL CHARACTER RECOGNITION APPLICATION TODAY

Optical Character Recognition has been applied to various applications. OCR has empowered examined archives to turn out to be something other than picture records, transforming into completely accessible reports with text content perceived by PCs. Optical Character Recognition extricates the applicable data and naturally enters it into electronic data set rather than the customary method of physically retyping the text. Optical Character Recognition is a huge field with various differed applications like Useful applications, Receipt imaging, Lawful industry, Banking, Medical care, and so forth OCR is additionally broadly utilized in numerous different fields like Captcha, Institutional repositories and digital libraries, Optical Music Recognition without any human correction or human effort, Automatic number plate recognition, Handwritten Recognition and Other Industries[9].

Technically, OCR was developed as an API to complement other more complex and useful applications. The first OCR module developed was tesseract. Tesseract was developed by HP between 1984 and 1994. HP launched it as open source in 2005[8]. From this emergence came various other OCR modules such as OpenCV and google vision. This module is widely used to build useful applications.

III. CONCLUSION

This paper has provided an explanation of the stages carried out in Optical Character Recognition (OCR). Based on this process it can be seen that the image preprocess greatly determines the resulting recognition rate. It is also known that most studies using ANN and Template Matching are known to produce the best recognition rates. Beside that, feature extraction is one of the one of the most troublesome issues in pattern recognition. OCR has also been widely used in various applications for data storage. This data storage can be used to find knowledge from patterns generated by the stored data set.

REFERENCES

1. Mohammad, F., et al., *Optical character recognition implementation using pattern matching.* International Journal of Computer Science and Information Technologies, 2014. **5**(2): p. 2088-2090.
2. Cheriet, M., et al., *Character recognition systems: a guide for students and practitioners, vol 2, pp, 1-321*. 2007: John Wiley & Sons.DOI: https://doi.org/10.1002/9780470176535.
3. Chaudhuri, A., et al., *Optical character recognition systems*, in *Optical Character Recognition Systems for Different Languages with Soft Computing*. 2017, Springer. p. 9-41.DOI: https://doi.org/10.1007/978-3-319-50252-6_2.
4. Arica, N. and F.T. Yarman-Vural, *An overview of character recognition focused on off-line handwriting.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2001. **31**(2): p. 216-233.DOI: https://doi.org/10.1109/5326.941845.
5. Shrivastava, V. and N. Sharma, *Artificial neural network based optical character recognition.* arXiv preprint arXiv:1211.4385, 2012.DOI: https://doi.org/10.5121/sipij.2012.3506.
6. Ashburner, J. and K.J. Friston, *Nonlinear spatial normalization using basis functions.* Human brain mapping, 1999. **7**(4): p. 254-266.DOI: https://doi.org/10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G.
7. Yoruk, E., et al., *Shape-based hand recognition.* IEEE transactions on image processing, 2006. **15**(7): p. 1803-1815.DOI: https://doi.org/10.1109/TIP.2006.873439.
8. Smith, R. *An overview of the Tesseract OCR engine*. IEEE.