Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Advanced Digital Image Processing Technique based Optical Character Recognition of Scanned Document

S. Iwin Thanakumar Joseph

Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, Andhra Pradesh, India

E-mail: iwineee2006@gmail.com

Abstract

For many years, the field of image processing and pattern approval has included handwriting approval among its most intriguing and rigid analytical fields. In this article, the steps necessary to convert text from a paper document to a computer-readable format has been discussed. This is the most tedious and labor-intensive task. For nearly three decades, scientists have been trying to figure out how to make a computer read like a human. In Optical Character Recognition (OCR), a scanned picture is converted mechanically or electronically into an image that may be read as handwritten, typed, or printed text. It's a way to turn paper documents into digital files that can be searched and utilised in automated procedures. To facilitate applications like machine translation, text-to-speech, and text mining, OCR encodes the pictures as machine-readable text. It's an easy and inexpensive approach to make OCR that can read any document in a standard font size and with standard handwriting.

Keywords: Character recognition, scanned document, image processing, text line detection, character segmentation

1. Introduction

Automatically converting an image of a page into a text file using symbols is the goal of Optical Character Recognition (OCR). Images of input documents might be sourced from several different sources. A document's image format might be anything digital, fax, scanned, machine printed or handwritten. A document's text is extracted by an OCR system and saved as a symbolic text file. In addition to the page layout, font size, style, document area type,

DOI: https://doi.org/10.36548/jiip.2022.3.007
Received: 20.08.2022, received in revised form: 17.09.2022, accepted: 02.10.2022, published: 13.10.2022

confidence level for the identified characters etc., it also provides additional descriptive information.

1.1 Optical Character Recognition

The field of pattern recognition and artificial intelligence has found its most widespread application in OCR. There are a number of commercial OCR systems available, and it is used in a variety of contexts; nonetheless, computers still can't quite keep up with human readers. One common use for scanners is scanning character pictures on forms. Images are sent into the OCR system's recognition engine, which then converts the readable text into ASCII code [1-4].

1.2 Digital Creation

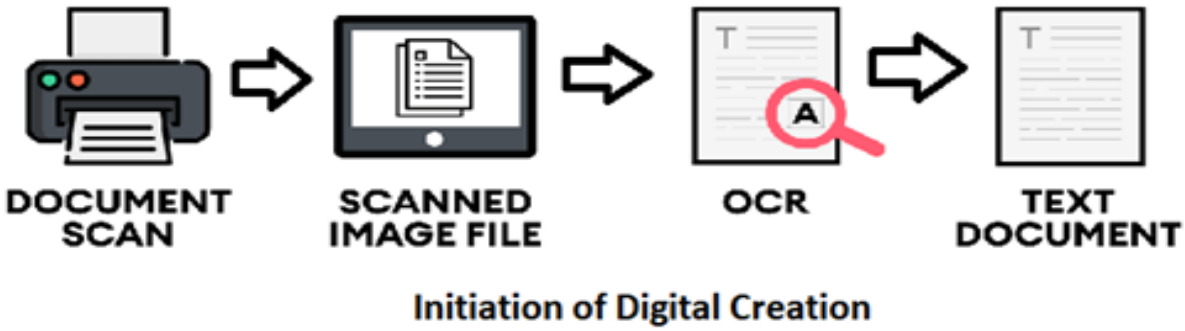

Figure 1. Initial Stages for Digital Creation

Printing, archiving, and reading texts have all seen significant shifts in methodology since their inception, majorly due to technological developments. Digital document generation and printing have replaced older methods like woodblock printing and typing. Document scanning became a viable method along with this development. Almost any paper document may be scanned nowadays and worked with like a digital file, but raw scans often lack any information that might be useful for further analysis. Although it begins as a high-quality scan of a printed page, it may be converted into a file in which individual composites of the text, figures, and tables can be used to reconstruct the original document. As you may imagine, this has far-reaching implications for how history is stored and accessed. For this reason, numerous antique historical documents have been scanned using scanning technology, since every physical copy may be converted to digital form. It contains a wide variety of sources, from newspapers published in the 1920s in Finland [1] and Australia [2] to historical medical writings published in Britain [3] and much more. The textual information

included in these papers might be extracted using an OCR programme for ease of reading. Figure 1 shows some simple steps of the initial stages for digital creation.

1.3 Initial stages

With OCR, printed text can be digitized from an image. The use of OCR may streamline various processes and increase the amount of searchable and accessible material. A high-quality picture, well-lit areas, legible writing, and other factors are all necessary for the tool to function well. Unfortunately, mistakes may still occur even if all those conditions are satisfied. Accordingly, such standards are seldom met when an old document has aesthetic flaws. Disadvantages such as archaic terminology, unique typefaces, irregularity, and/or strange printing of characters accompany the era of digitalizing physical paper [5, 6]. In addition to these limitations, there is also the possibility that sentiments have faded, wrinkles have formed as a result of folding, and the paper itself has been damaged in some way, such as by holes or torn edges. Stains and other colour shifts or fadings on the paper may also have developed over time.

1.4 Scanned document analysis

Even more of a problem arises when considering archived, form-type documents from the past, since they often include one-of-a-kind form values. These variables may be a name, ID number, or something else entirely, and they likely seemed visibly different in past form-type documents since they weren't mechanised. It is usual to see letters overlapping each other, lines crossing over, and doubled characters when equipment like typewriters were utilised. OCR is useful, yet it has certain limitations. Extracting text from a document is known as document analysis. Quality of the source document and the quality of the registered picture are prerequisites for accurate character segmentation and identification. Image editing techniques such as sharpening, blurring, and desaturating may rescue images from mediocre source material or noisy scans [7, 8].

1.5 Conventional image processing approach

The approaches used to improve images have an emphasis on identifying text from background. Image processing may remove printed guidelines and other lines that could be emotionally upsetting and get in the way of character identification using noise reduction and underline removal. The ability to extract characters from a graphical representation of text is

crucial for document analysis. This is accomplished in many OCR systems by the usage of interconnected parts [9,10].

2. Literature Survey

Based on the work of Peng et al. [11], OCR relies heavily on the accuracy of the text line segmentation procedure. Several popular strategies, including projection-based and stochastic techniques, have been proposed to accomplish this goal. The pictures in palm leaf manuscripts of Dai are of low quality and contain smudges, wrinkles, stroke distortion, and character touches, rendering them unsuitable for direct application of most current technologies to their processing.

Scientists T. A. Sanjrani and associates [12], reported that Sindhi has a script similar to those of Arabic and Persian. It has been around for at least 2500 years and is now spoken in a number of Asian nations. They presented an OCR system that can read handwritten strings of Sindhi numbers instead of relying on the usual keyboard and data storage input methods. Character recognition is at the heart of the investigations; this technology has potential applications in areas as diverse as tutoring, children's arithmetic games, and the automated conversion of phone digits from signs in India and Pakistan.

A group led by S. F. Rashid and others [13] proposed that there is a widespread belief that OCR of papers written with a Latin script on a machine may be achieved. Still, it's difficult for state-of-the-art OCR systems to perform flawless OCR on deteriorated or noisy text. The majority of modern methods for recognising text include segmentation. Degraded text makes this difficult since it is difficult to divide it into individual words.

According to Siebra Lopes et al. [14], there is a rising need for digital manuscript identification in a variety of contexts, including the automatic rerouting of mail via the recognition of handwritten postal address digits and the acceptance of nominal values in bank cheques. Variation within a class of handwritten numbers is notoriously difficult to handle because of the wide range of possible writing styles and character inclinations. OCR software analyses scanned pictures of printed text in order to decipher it. This functionality is already widespread in scanners and mobile devices.

In order to retrieve images from documents, Hassan and his associates [15], suggested an innovative multi-modal framework that draws on textual and graphical information. The approach employed many hashing formulations based on kernel learning to produce

multimodal composite document indexes. Moving further, a new multi-modal document indexing architecture that makes use of learning to integrate OCR text and image-based representation, with the aim of recovering old and damaged text documents was proposed.

The research by Mantoro et al. [16], suggested a system for server-based processing of OCR on mobile devices. The proposed work focused on OCR based processing to obtain higher accuracy and less execution timing.

2.1 Research questions

The research framework aims to address the following research questions:

- Can the existing findings of an OCR tool be improved by the use of image processing methods, such that scanned historical documents are simpler to read?
- Secondly, is the picture quality consistent amongst scanned historical manuscripts where the text is not always legible?

3. Character Recognition of Scanned Document

3.1 Optical Scanning

A digital picture is acquired from the original paper using optical scanning technique. OCR makes use of optical scanners, which typically have a transport mechanism and a sensor device that maps light intensity to grayscale [17]. Commonly, papers that are printed have black text on a white backdrop. In order to execute OCR, it is a normal practice to first transform the multilevel picture to a bi-level image in black and white. Often referred to as "thresholding", this operation is carried out by the scanner to save data storage and processing power.

3.2 Preprocessing

Preprocessing may improve the original image's suitability for further computing by the use of techniques like thresholding, binarizing, filtering, edge detection, gap filling, segmentation, and so on. When an object is scanned, the resultant picture might have some noise [18]. The degree to which the characters are blurred or broken depends on the scanner's resolution and the effectiveness of the thresholding procedure. A preprocessor may smooth the digitised characters, fixing some of the flaws that might otherwise lead to low recognition rates. Preprocessing often includes normalisation along with filing. Applying normalisation

ensures that all characters have the same height, width, and orientation. Accurately describing, the procedure requires determining the optimal starting point. In order to identify skew in rotated pages and lines of text, versions of the Hough transform are often utilised. Until a character has been identified, however, it is difficult to know how it rotates.

3.3 Segmentation

In order to identify the parts of a picture, a segmentation procedure must be performed. Identifying the textual locations of data and differentiating them from those of other visual elements is essential. Taking out individual letters, or even whole words, from a text is called segmentation. There is a plethora of word-segmenting algorithms that break down words into their constituent characters. Isolating each interdependent part is important to the segmentation process. Figure 2 contains the blocks of the proposed improved image processing approach through hybrid model.

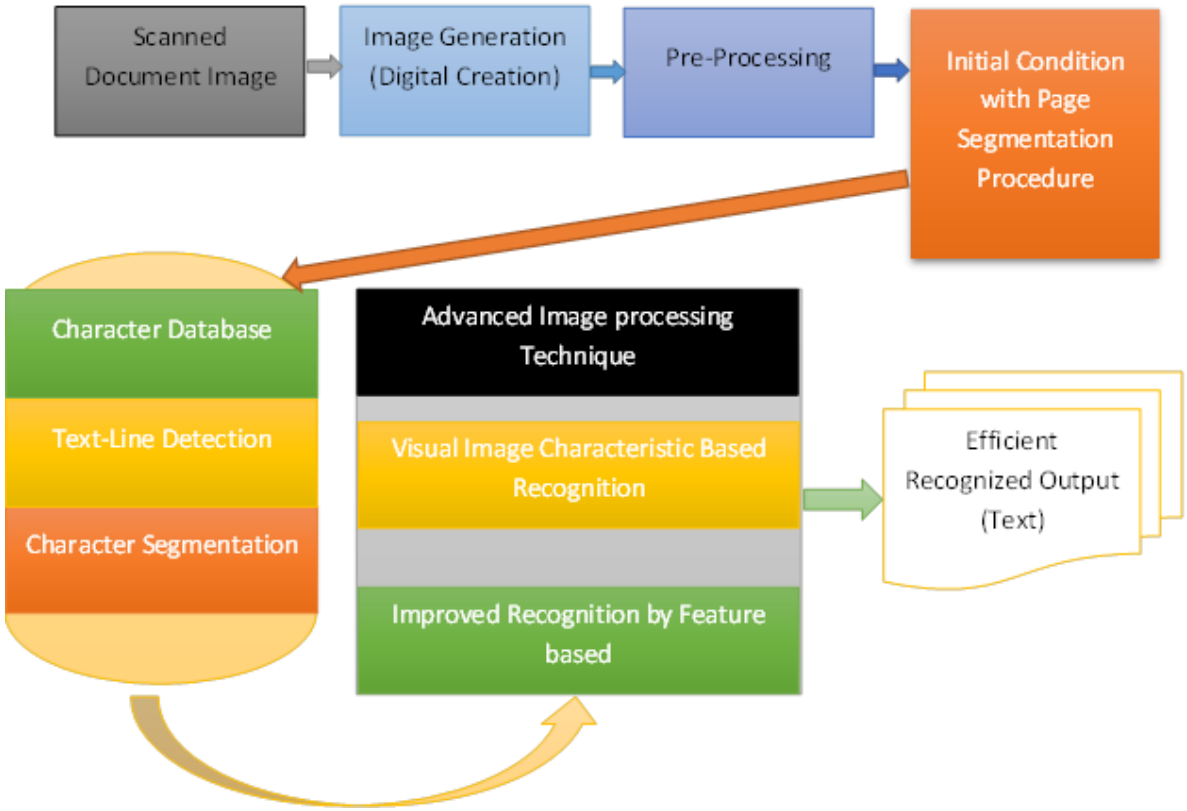

Figure 2. Improved image processing approach through hybrid model

3.4 Page Segmentation

In an OCR system, page segmentation is an essential preliminary processing step. The term "image segmentation" refers to the procedure of breaking up a document picture into

smaller, more manageable sections that each include the same sort of information, whether it is a text, a table, a figure, or a halftone image. In many scenarios, the efficacy of an OCR system is directly tied to the precision of the page segmentation algorithm [19]. In order to preserve just the relevant fragments of a given text, a page segmentation technique is used. Without the need for horizontal scrolling, shorter sections of pages may be constructed.

This section of the document, rather than the whole thing, may be forwarded for further processing. This strategy is very useful for preserving the straight margins of papers. The OCR may also analyse the layout of a document to ensure that the text remains in the correct sequence for reading. Consequently, OCR systems will be able to recognise text when photos cannot [20 -22].

Step1: Text Line Detection

There are three main components to the approach used for text line segmentation.

- In the first stage, the binary picture is processed by extracting related components and estimating the average character height.
- In the second stage, potential text lines are identified using a block-based Hough transform.
- In the third and final stage, any necessary splitting is performed i.e., any text lines that were missed in the second step are found, and any vertically related letters are separated and placed on separate text lines.

Step2: Word and Character Segmentation

Once the text lines are identified, projection profiles built on top of them are utilised to identify the words. Then, the following process is used to decipher the symbols into alphabetic form. Specifically, it is based on the segmentation method described for detecting numbers that touch. The essential notion is that the feature points on the word's skeleton and its backdrop may be used to locate probable segmentation pathways [23].

Step3: Character Database

Here, it chooses a sample of photos to use for training, and then the segmentation method is used to those images in order to pull out the characters. However, no OCR approach based on supervised learning that requires a training set of labelled patterns can be employed since there is no classification of these characters. As a result, it is first focused on "revealing" how the characters are grouped into "reasonable" categories. Once all necessary

steps have been taken to repair any mistakes in the clusters, they are given labels, and the character database is formed [24].

4. Comparative Observation Procedures

4.1 Recognition

Segmented image is matched to all device-preloaded templates. After the correlation is complete, the most highly associated template is claimed to represent the subject of the photograph. Every non-training-phase document picture is now ready to be transformed into a text file. Using the method outlined, in which each character is represented as a feature vector, characters are extracted, and finally, all characters are categorised using the database constructed. According to the cluster, each character may be mapped to a feature vector. A character's place in the cast is determined by their entry in the collection's database.

4.2 Visual Image Characteristics

Each picture in the collection will have its attribute classification result shown here, along with a pie chart indicating the percentage of images with imperfect OCR. If all characters are successfully recovered, the OCR process may be considered a success. If even one character is right, or if one or more characters are accidentally added to an already valid registration number, then the attempt has succeeded to some extent. Finally, if no characters are successfully scanned, the outcome is considered to be failed. Different groups of findings, those for blue- and red-themed photos, are compiled.

5. Conclusion

This article offers a brief overview of the packages across several disciplines and experiments in a few areas. The suggested method is very environmentally friendly for extracting all types of bimodal pictures including blur and illumination. For scholars just starting out in the field of individual optical repute, this study will serve as a solid literature overview. These complications stem from the various character forms, top bars, and quit bars. In addition, there are differences in vowels and compound letters. Recognition errors in the OCR system are a common cause of low accuracy. Constrained OCR may become less necessary in the long run. This is due to the fact that, when the production process is under control, the document is frequently created using information already saved in the computer. Information may be sent electronically or printed in a more machine-readable form, such as a

barcode, if a machine-readable version already exists. Recognizing documents in settings where production quality control is not an option, is where OCR systems of the future will find their greatest use. This may be the case for older content that could not be created electronically at the time of creation, or in situations where the receiver is shut off from an electronic version and has no influence over the production process. This indicates that understanding printed text will be the focus of future OCR systems. A further vital use of OCR is in the reading of handwritten materials.

References

[1] O. Joshua, T. Ibiyemi, and B. Adu, "A Comprehensive Review On Various Types of Noise in Image Processing," Int. J. Sci. Eng. Res., vol. 10, no. November, pp. 388–393, 2019.

[2] U. Garain, A. Jain, A. Maity, and B. Chanda, "Machine reading of camera-held low quality text images: An ICA-based image enhancement approach for improving OCR accuracy," Proc. - Int. Conf. Pattern Recognit., pp. 1–4, 2008, doi: 10.1109/icpr.2008.4761840.

[3] C. H. Keerthana, P. S. S, S. S. Pai, V. A. Meda, and M. D. T H, "Character Recognition of Handwritten Text Using Machine Learning and Image Processing," J. Opt. Commun. Electron., vol. 5, no. 2, pp. 11–16, 2019, doi: 10.5281/zenodo.2705098.

[4] B. Altinoklu, I. Ulusoy, and S. Tari, "A probabilistic sparse skeleton based object detection," Pattern Recognit. Lett., vol. 83, pp. 243–250, Nov. 2016, doi: 10.1016/j.patrec.2016.07.009.

[5] M. Koistinen, K. Kettunen, and T. Pääkkönen, "Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing," Proc. 21st Nord. Conf. Comput. Linguist., no. May, pp. 23–24, 2017.

[6] S. Mandal, S. Chowdhury, A. Das and B. Chanda, "A Simple and Effective Table Detection system from Document Images," Int'l J. Document Analysis and Recognition, vol. 8, nos. 2-3, pp. 172-182, June 2006.

[7] F. Shafait, D. Keyser and T.M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images," Proc. 18th Int'lConf. Pattern Recognition, pp. 872-875, Aug. 2006.

[8] F. Shafait, J. van Beusekom , D. Keysers, and T.M. Breuel, “Page Frame Detection for Marginal Noise Removal from Scanned Documents,” Proc. Scandinavian Conf. Image Analysis, pp. 651-660, June 2007

[9] N. Stamatopoulos, B. Gatos and A. Kesidis, “Automatic Borders Detection of Camera Document Images,” Proc. Second Int’l Workshop Camera-Based Document Analysis and Recognition, pp. 71-78, Sept. 2007.

[10] D. Keysers, F. Shafait and T.M. Breuel, “Document Image Zone Classification—a Simple High-Performance Approach,” Proc. Second Int’l Conf. Computer Vision Theory and Applications, pp. 44-51, Mar. 2007.

[11] G. Peng, P. Yu, H. Li and L. He, "Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai," 2016 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2016, pp. 336-340.

[12] A. Sanjrani, J. Baber, M. Bakhtyar, W. Noor and M. Khalid, "Handwritten Optical Character Recognition system for Sindhi numerals," 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, 2016, pp. 262-267.

[13] S. F. Rashid, F. Shafait and T. M. Breuel, "Scanning Neural Network for Text Line Recognition," 2012 10th IAPR International Workshop on Document Analysis Systems, Gold Cost,QLD,2012,pp.105-109. doi: 10.1109/DAS.2012.77.

[14] G. Siebra Lopes, D. Clifte da Silva, A. W. Oliveira Rodrigues and P. P. Reboucas Filho, "Recognition of handwritten digits using the signature features and Optimum-Path Forest Classifier," IN IEEE Latin America Transactions, vol. 14, no. 5, pp. 2455-2460, May 2016.

[15] E. Hassan, S. Chaudhury and M. Gopal, "Multi-modal Information Integration for Document Retrieval," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, 2013, pp. 1200-1204.

[16] T. Mantoro, A. M. Sobri and W. Usino, "Optical Character Recognition (OCR) Performance in Server-Based Mobile Environment," 2013 International Conference on Advanced Computer Science Applications and Technologies, Kuching, 2013, pp. 423-428.

[17] Samantaray, R. K., Panda, S., & Pradhan, D. (2011). Application of Digital Image Processing and Analysis in Healthcare Based on Medical Palmistry. IJCA Special Issue on Intelligent Systems and Data Processing, 56-59.

[18] Sharma, D. V., Saini, G., & Joshi, M. (2012). Statistical Feature Extraction Methods for Isolated Handwritten Gurumukhi Script. International Journal of Engineering Research and Application, 2(4), 380-384.

[19] M. Zhang, F. Xie, J. Zhao, R. Sun, L. Zhang, and Y. Zhang, "Chinese license plates recognition method based on a robust and efcient feature extraction and bpnn algorithm," Journal of Physics: Conference Series, vol. 1004, p. 012022, 2018.

[20] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, and N. Komodakis, "A robust and efcient approach to license plate detection," IEEE Transactions on Image Processing, vol. 26, no. 3, pp. 1102–1114, 2017.

[21] Y. Wiseman, "Vehicle identifcation by OCR, RFID and Bluetooth for toll roads," International Journal of Control and Automation, vol. 11, no. 9, pp. 1–12, 2018.

[22] A. Saini, S. Chandok, and P. Deshwal, "Advancement of trafc management system using RFID," in Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1254–1260, Madurai, June 2017

[23] Q. Wang, "License plate recognition via convolutional neural networks," in Proceedings of the 2017 8th IEEE International Conference on Sofware Engineering and Service Science (ICSESS), pp. 926–929, Beijing, China, November 2017.

[24] R. Fu, "Te research and design of vehicle license plate recognition system in trafc management system," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 9, no. 3, pp. 445–456, 2016.

Author's biography

S. Iwin Thanakumar Joseph works in K L University, Vaddeswaram, Guntur, India. He has done his Ph.D in the Department of Computer Science and Engineering in Annamalai University, Chidambaram, Tamilnadu, India. His research interest includes artificial intelligence, machine learning, image processing, internet of things and soft computing techniques. He published more than 15 research papers in this research area. He is a lifetime member of ISTE professional community.