



# **Review of the Character Recognition System Process and Optical Character Recognition Approach**

**Jaswinder Kaur, Mrs. Rupinder Kaur**

CSE & PTU, India

[jass1153@gmail.com](mailto:jass1153@gmail.com), [rupy.wahla@gmail.com](mailto:rupy.wahla@gmail.com)

---

*Abstract— Character recognition in pictures could be an analysis space that makes an attempt to develop a computing system with the power to mechanically scan the text from pictures. Currently there's a large demand in storing data the knowledge the data out there in paper document format into a hardware disk and so later reusing this information by looking method. One easy thanks to store data from these paper documents in to computing system is to 1st scan the documents and so store them as pictures. However to utilize this data it's terribly troublesome to scan the individual contents and looking the contents kind these documents line-by-line and word-by-word. The challenges concerned during this the font characteristics of the characters in paper documents and quality of pictures. Thanks to these challenges, PC is unable to acknowledge the characters whereas reading them. Therefore, there's a requirement of character recognition mechanisms to perform Document Image Analysis (DIA) that transforms documents in paper format to electronic format. During this paper, they have reviewed and analyzed completely different ways for text recognition from pictures. The target of this review paper is to summarize the well-known ways for a higher understanding of the reader*

*Keywords— Optical Character Recognition (OCR), Classification, Digital Image process.*

---

## **I. INTRODUCTION**

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website. OCR could be a system that gives full alphabetical recognition of written or written characters at merely scanning the document. Documents area unit scanned employing a scanner and area unit given to the OCR systems that acknowledges the characters within the scanned documents and converts them into code information. The collateral is completed either arbitrarily or chronologically by human Intervention [1]. The area unit a of OCR is changing into associate integral a part of document are scanners, and is employed in several applications cherish communication process, script recognition, banking, security (i.e. passport

authentication) and language identification. The analysis during this space has been current for over a century and also the outcomes are astounding with palmy recognition rates for written characters prodigious ninety nine, with vital enhancements in performance for written cursive character recognition wherever recognition rates have exceeded the ninetieth mark [2]. In OCR a info is employed at the rear finish for recognition. The method consists of following process steps:

- (1) Scanning of Image,
- (2) Preprocessing of Image
- (3) Character Extraction
- (4) Feature Extraction and Recognition
- (5) Post-Processing. [3].

Optical Character Recognition is classed into 2 sorts, Offline recognition and on-line recognition. In offline recognition the supply is either a picture or a scanned variety of the document whereas in on-line recognition the sequential points area unit painted as a perform of your time and also the order of strokes also are offered [4]. The OCR system provides the subsequent Features:

- No a lot of retyping,
- Fast Digital Searches,
- Edit Text,
- Save area.

Signature verification is a very important analysis space within the field of authentication of someone yet as documents [5] in E-Commerce and banking. The importance of signature verification arises from the actual fact that it's long been accepted in government, legal, and industrial transactions as an appropriate methodology of verification [6]. Recognition is finding the identification of the signature owner. Verification is that the call regarding whether or not the signature is real or forgery. During this call section the forgery pictures is classified in 3 groups: (i) random, (ii) easy, (iii) ball-hawking [7]

## II. RELATED WORK

**Rhead et al. [8]** has thought of planet kingdom range plates and relates these to ANPR. It considers aspects of the relevant legislation and standards once applying them to planet range plates. The various producing techniques and varied specifications of element components also are noted. The various fixing methodologies and fixing locations area unit mentioned still because the impact on image capture.

**Badawy, W. et al. [9]** has mentioned the automated vehicle plate recognition (ALPR) is that the extraction of auto vehicle plate data from a picture or a sequence of pictures. The extracted data is used with or while not a info in several applications, like electronic payment systems (toll payment, parking fee payment), and superhighway and blood vessel observance systems for traffic police work. The ALPR uses either a color, black and white, or infrared camera to require pictures.

**Jawahar et al. [10]** have proposed a recognition theme for the Indian script of script. Recognition accuracy of Devanagari isn't nevertheless like its Roman counterparts. This is often in the main because of the quality of the script, genre etc. Our answer uses a continual Neural Network called bifacial Long- Short Term Memory (BLSTM). Our approach doesn't need word to character segmentation that is one in all the foremost common reason for top word error rate.

**Ntirogiannis et al. [11]** has studied that the document image linearization is of nice importance within the document image analysis and recognition pipeline since it affects additional stages of the popularity method. The analysis of a linearization technique aids in finding out its recursive behavior, still as validator its effectiveness, by providing qualitative and quantitative indication of its performance. This paper addresses a pixel-based linearization analysis methodology for historical handwritten/machine-printed document pictures. Within the planned analysis theme, the recall and preciseness analysis measures area unit properly changed employing a coefficient theme that diminishes any potential analysis bias.

**Malakar et al. [12]** has represented that extraction of text lines from document pictures is one in all the vital steps within the method of associate degree Optical Character m Recognition (OCR) system. Just in case of written document pictures, presence of skew, touching or overlapping text line(s) makes this method a true challenge to the research worker. The current technique extracts eighty seven.09% and 89.35% text lines with success from the aforesaid databases severally.

### III. ARCHITECTURE OF A GENERAL CHARACTER RECOGNITION SYSTEM

The major steps concerned in recognition of characters embrace, preprocessing, segmentation, feature extraction and classification (fig. 1)

A. *Preprocess* - The sequences of pre-processing steps square measure as follows

- *Noise Removal* - Noise is outlined as any degradation within the image thanks to external disturbance. Quality of written documents depends on varied factors together with quality of paper, aging of documents, quality of pen, color of ink etc. Some samples of noises square measure mathematician noise, salt and pepper noise. These noises square measure removed to sure extent mistreatment filtering technique. Technical details of filtering are determined in [13].



Original Image Binary Image After Preprocessing

Fig 1. Preprocessing of Image

- *Thresholding* - The task of thresholding is to extract the foreground (ink) from the background (paper) [14]. Given a threshold, T between zero and 255, replace all the pixels with grey level under or adequate T with black (0), the remainder with white (1). Fig. 2. Design of a personality recognition system If the edge is simply too low, it's going to cut back the amount of objects and a few objects might not be visible. If it's too high, we tend to might embrace unwanted background info. The acceptable threshold price chosen is applied globally or regionally. Otsu's [15] algorithmic program is that the unremarkably used world thresholding algorithmic program..

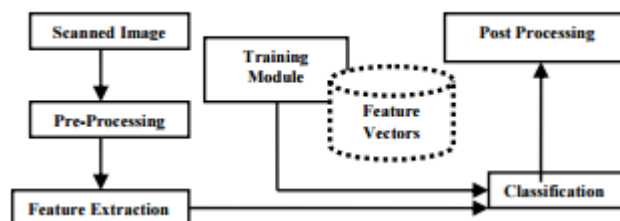
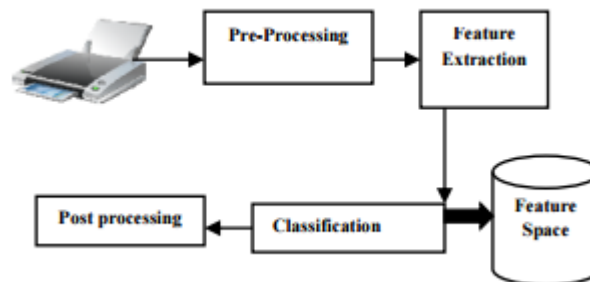


Fig 2. Architecture of a character recognition system

- *Skeletonization* - Skeletonization is a picture preprocessing technique performed to create the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region. a whole survey of cutting methodologies is mentioned in [16]

B. *Segmentation* - Segmentation step contains word segmentation, character segmentation and line segmentation. strategies for character segmentations [17] square measure supported i) White area and pitch ii) Projection analysis and iii) connected element labeling



C. *Standardisation* - it's the method of changing the random sized image into normal sized image. In this, size standardisation avoids inhome category variation among characters. Bilinear, Bi-cubic interpolation techniques square measure a number of strategies for size standardisation.

*D. Feature Extraction* - options square measure a group of numbers that take the salient characteristics of the metamer image. There square measure completely different feature extraction strategies for character recognition [18].

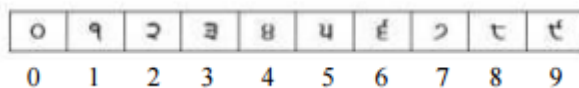


Fig 3. Punjabi (Gurmukhi) Number from 0 to 9

*E. Classification* - The feature vector obtained from previous section is appointed a category label and recognized mistreatment unsupervised and supervised methodology. The information set is split into coaching set and take a look at set for every character. Character classifier is Thomas Bayes classifier, nearest neighbor classifier, Radial basis operate, Support vector machine, linear discriminant functions and Neural networks with or while not back propagation.

*F. Post-processing* - Post-processing step involves grouping of symbols. The method of performing arts the association of symbols into strings is noted as grouping.

#### IV. APPLICATION

Text recognition technology is also applied throughout the complete spectrum of industries, revolutionizing the document management method. This technology modify scan documents to become quite simply image files, turning into totally searchable documents with text content that's recognized by computers. With the assistance of this technology, folks now not have to be compelled to manually retype necessary documents once coming into them into electronic databases. Instead, Text recognition system extracts relevant data and enters it mechanically. The result's correct, economical information science in less time. Within the following, we have a tendency to summary some applications of text recognition system,

##### A. Banking

The uses of image text recognition vary across totally different fields. One wide proverbial application is in banking, it's accustomed method checks while not human involvement. A check may be inserted into a machine, t he writing thereon is scanned instantly, and also the right amount of cash is transferred. This technology has nearly been formed for written checks, and is fairly correct for written checks further, although it sometimes needs manual confirmation. Overall, this reduces wait times in several banks.

##### B. Legal

Within the legal business, there has conjointly been a major movement to change paper documents. So as to save lots of house and eliminate the requirement to sift through boxes of paper files, documents ar being scanned and entered into pc databases. Image text recognition additional simplifies the method by creating documents text-searchable, so they're easier to find and work with once within the info. Legal professionals currently have quick, easy accessibility to a large library of documents in electronic format that they'll notice just by typewriting during a few keywords.

##### C. Healthcare

Healthcare conjointly use of image text recognition technology to method work. Health care professionals continually ought to take care of massive volumes of forms for every patient, together with insurance forms likewise as general health forms. to stay up with all of this info, it's helpful to input relevant information into AN on-line database that may be accessed as necessary. By victimisation image recognition technology they're able to extract info from forms and place it into databases, so each patient's information is promptly recorded. As a result, health care suppliers will specialise in delivering the simplest attainable service to each patient.

##### D. Image text recognition in alternative Industries

Image text recognition technology is wide utilized in several alternative fields, together with education, finance, and government agencies. This technology has created innumerable texts accessible on-line, saving cash for college kids and permitting information to be shared. Invoice imaging applications square measure utilized in several businesses to stay track of economic records and stop a backlog of payments from stilt up. In government agencies and freelance organizations, image text recognition technology simplifies information assortment and analysis, among alternative processes.

#### V. CONCLUSION

In this paper we've got reviewed and analyzed totally different ways to seek out text characters from scene pictures. We've got reviewed basic design of text recognition from pictures. During which we have a tendency to mention totally different techniques of image process especially sequence for text recognition from scan image. Also, we've got mentioned some application of text recognition system.

## REFERENCES

- [1] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1, Jan-Feb 2012.
- [2] "Pre-processing Techniques in Character Recognition", Yaseer Alginahi, 17, ISBN 978-953-307-105-3, August, 2010.
- [3] Singh, Guriqbal, and Vikas Mongia. "Recognition Of Punjabi Script Character And Number For Multiple Fonts."
- [4] "Document Analysis and Recognition", Alon, Jonathan, 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005.
- [5] "Personal Authentication Using Signature Recognition", Diana Kalenova, Department of Information Technology, Laboratory of Information Processing, Lappeenranta University of Technology.
- [6] "An introduction to biometric recognition", A.K. Jain, A. Ross, and S. Prabhakar, IEEE Trans. on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4–20, January 2004.
- [7] "Signature Verification: Increasing Performance by a Multi-Stage System", Sansone and Vento, Pattern Analysis & Applications, vol. 3, pp. 169–181, 2000.
- [8] 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.
- [9] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." IEEE International Conference on Document Analysis and Recognition, 2012.
- [10] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, 2012.
- [11] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis. "A Performance Evaluation Methodology for Historical Document Image Binarization." IEEE International Conference on Document Analysis and Recognition, 2013.
- [12] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2012
- [13] Lim, Jae S., "Two-Dimensional Signal and Image Processing", Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.
- [14] P. K. Sahoo, S. Soltani, A.K.C Wong and Y C Chen, "A survey of Thresholding Techniques", Computer Vision, Graphics and Image processing, vol 41, pp 233-260, 1988.
- [15] Otsu.N, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cybernetics, vol.9, pp.62-66, 1979
- [16] L. Lam, S.W. Lee and C.Y.Suen, "Thinning Methodologies: A Comprehensive Survey", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.14, pp 869-885,1992
- [17] Richard G. Casey And Eric Lecolinet " A Survey Of Methods And Strategies In Character Segmentation", IEEE Trans. On Pattern Analysis And Machine Intelligence, Vol 18, Pp 690-706,1996
- [18] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 19a96.