International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)

ISSN: 2456-236X

Vol. 04 Issue 02 |2020

A survey of Offline Handwriting Recognition using Optical Character Reader

Srividya B G[1], Varun H M[2]

[1, 2] Student, Department of MCA, Jain Deemed-to-be University, Bengaluru, Karnataka, India

ABSTRACT

Handwriting Recognition among the first few biometrics to be employed even before computers arrive. It allows a person to write something on a piece of paper, and then translate it into text. There are countless ways in which a character can be portrayed if we look into practical reality. Such types can be combined by themselves to produce more styles. Recognition of the handwriting involves matching the template with the scanned document created from the unknown image. Using this approach, we got 98.44 per cent accuracy. A technique for offline handwritten detection, based on optical character recognition (OCR) is proposed in the present paper.

Keyword: - Biometrics, OCR

1. INTRODUCTION

When we pass from stone to papyrus to paper the handwriting has the server for transmitting the authors ' personal message. The handwriting-based authentication method is one of the most commonly used security procedures, it may be a good idea to use handwriting to identify people to identify the author of a given document.

Knowing that each person has exclusive handwriting but people can learn to replicate the handwriting of other people to falsify it. Analysis of the handwritten document is, therefore, necessary, a handwriting recognition biometric device scans the document and takes the input of the scanned document, in this method they also look at the form of each letter and analyze the writing act, the pen strength, the pace and the rhythm with which they write. This is a part of the methods of biometric behavioural recognition. Personal identification based on handwriting has a wide range of uses including bank check processes, certificate signatures, postal code verifications, forensics, etc. Handwriting-focused biometric approaches may be split down into three areas: handwriting detection, forensic identification and device authentication.

1.1 Types of recognition areas in Handwriting

There are two primary types of recognition areas in Handwriting:

Static: In this form, the person writes on paper, the document (book) is digitized via an optical scanner or a camera, and the biometric device recognizes the text evaluating its shape, involves the automated translation of the text in the picture into letter codes that can be used in the computer and text processing. The data obtained from this process is known as static handwriting representation, and this type of handwritten recognition is also called offline.

- Writing speed
- Direction
- Duration
- Height and width
- slant angle
- blank pixels.

Dynamic: Throughout this method, the individual writes throughout a digitizing tablet which acquires the message in actual time, involving instant translation of the text while it is handwritten on a specific digitizer or PDA where a sensor picks up the pen tip motions while well as pen-up / pen-down flipping. The data collected from this procedure is called the interactive depiction of handwriting, and this form of handwritten identification is often named online. Typically, complex knowledge is as follows:

- Spatial coordinate x(t)

- Spatial coordinate y(t)
- Pressure p(t)
- Azimuth az(t)
- Inclination in(t)

Off-line and on-line recognizers require different equipment, while off-line recognizers need only a scanner to digitize written text, on-line recognizers need a transducer to capture the writing as it is written.

1.2 Offline handwriting recognition

Recognition of Word and Character is the activities of offline handwriting. The handwritten text review is a preliminary step in the recognition where different approaches have been introduced to offline recognition which has led to the present technical competence in the document being scanned and used as an input to obtain Word and Character.

2. OVERVIEW OF OPTICAL CHARACTER RECOGNITION

Optical Character Recognition frequently called as an Optical Character Reader (OCR) is a strategy that capacities like human eyes and fills in as an optical gadget for distinguishing the picture seen by the mind's eye input. OCR is an examination field that incorporates design acknowledgement, man-made reasoning, and PC vision. For improved execution and effectiveness, it expels the human communications. OCR is a PC's investigation of composed characters. Records are checked to utilize a scanner and given to OCR frameworks that recognize the characters in the records being filtered and transform them into ASCII information that a machine can comprehend. Programmed content acknowledgement utilizing OCR is, at the end of the day, the strategy for deciphering a picture of content records into its computerized literary partner. The advantage is that it is conceivable to alter the literary substance which is in any case unrealistic in examined archives in which these are picture documents. The portrayal of the content itself may either be machine-printed or manually written or a blend of both. PC frameworks furnished with such an OCR framework speed up, decline some conceivable human mistakes and take into consideration minimized stockpiling, quick recovery and other document control. The scope of utilizations incorporates postal code distinguishing proof, robotized information section into an expansive managerial framework, bookkeeping, programmed cartography and perusing helps for the outwardly disabled when interfaced with a Voice synthesizer.

There are two types of optical recognition, offline recognition and online recognition. The source is either a picture or an examined adaptation of the record in offline recognition while the progressive focuses are delineated as a component of time in online recognition, and the request for strokes are additionally unmistakable. The principle highlights describing a decent OCR gadget are exactness, flexibility and speed. Here right now offline recognition is managed.

2.1 Components of an OCR

In OCR a database is used at the back end for recognition. An OCR system consists of several components. In the proposed system the process consists of following processing steps

2.1.1 Scanning of Image

In the archive examining step, a scanner is utilized to check the written by hand or printed records. The nature of the filtered archive relies up upon the scanner. Along these lines, a scanner with fast and shading quality is attractive. The scanners by and large comprise of a vehicle component in addition to a detecting gadget that changes over light power into dim levels. The picture is taken and is changed over to the greyscale picture. The greyscale picture is then changed over to the binary picture. This procedure is called Digitization of picture (Binarization). For all intents and purposes, any scanner isn't great; the filtered picture may have some commotion. This commotion might be because of some superfluous subtleties present in the picture. Printed records normally comprise of black print on a white foundation, the filtered picture is changed over from staggered picture into a bi-level picture of high contrast. Frequently this procedure is known as thresholding.

2.1.2 Location and segmentation

Segmentation is one of the most significant stages in OCR advancement that decides the basis of a picture. The division is the disconnection of characters or words. Most of optical character recognition calculations portion the words into confined characters which are perceived exclusively. Typically, this division is performed by disengaging each associated segment that is each associated dark zone. This procedure is anything but difficult to actualize, during the procedure of segmentation they experience various stages, they are:

- **Line segmentation:** wherein detachment of individual line of content is taken consideration.
- **Word segmentation:** wherein disengagement to printed content is finished.
- **Character segmentation:** is the disconnection of individual character ordinarily those that are composed discretely instead of cursively. Be that as it may, the issues happen if characters contact or if characters are divided and comprise of a few sections.

The primary issues in the division might be partitioned into four gatherings: Extraction of touching and fragmented characters. Such mutilations may prompt a few joint characters being deciphered as one single character, or that a bit of a character is accepted to be a whole image. Joints will happen if the archive is a dull photocopy or in the event that it is checked at a low threshold. Likewise, joints are normal if the textual styles are serried. The characters might be part of the record comes from a light photocopy or is examined at a high threshold.

- Distinguishing noise from the text.
- Specks and accents might be confused with clamour and the other way around.
- Mistaking graphics or geometry for text.
- This prompts no content being sent to acknowledgement.
- Mistaking text for graphics or geometry. Content won't be passed to the acknowledgement organize. This frequently occurs if characters are associated with illustrations.

2.1.3 Preprocessing of Image

The picture coming about because of the checking procedure may contain a specific measure of clamour. Clamour in pictures has numerous causes, for example, corrupted info pictures, blemished catch gadgets, just as inappropriate utilization of the recently utilized gadgets, pressure and transmission blunder. Prior to utilizing the pictures, the optical irregularities must be made up for. Commotion decrease targets expanding the sign to clamour proportion and Depending on the goals on the scanner and the accomplishment of the applied method for thresholding, the characters might be spread or broken. A portion of these deformities, which may later reason poor acknowledgement rates, can be killed by utilizing a preprocessor to smooth the digitized characters. The smoothing suggests both filling and diminishing. Filling dispenses with little breaks, holes and openings in the digitized characters, while diminishing lessens the width of the line. The most well-known strategies for smoothing, move a window over the parallel picture of the character, applying certain standards to the substance of the window. Notwithstanding smoothing, preprocessing, as a rule, incorporates standardization. The standardization is applied to acquire characters of uniform size, inclination and turn. To have the option to address for the turn, the edge of pivot must be found. For pivoted pages and lines of content, variations of Hough change are normally utilized for identifying slant. Nonetheless, to discover the pivot edge of a solitary image is beyond the realm of imagination until after the image has been perceived.

At this stage, we have the information as a picture and this picture can be additionally broke down with the goal that's the significant data can be recovered. This examination procedure can incorporate the accompanying focuses: Stages of preprocessing

- **Noise Reduction:** When the record is checked and is put away in picture design there are the odds that is the commotion is presented in the picture. Clamour is characterized as any debasement in the picture because of outer aggravation. The commotion can be presented by optical checking gadget or by the composing instrument. Because of the clamour, there can be the disengaged line fragment, enormous holes between the lines and so forth so it is fundamental to expel these mistakes so's the data can be recovered in the most ideal manner these commotions can be expelled to certain degree utilizing sifting strategy.
- **Normalization:** It is the way toward changing over the arbitrary estimated picture into the standard measured picture. The standardization is done with the goal that's the character is organized in a legitimate way. For the most part, the individual character is fitted inside the network. The grid can be of 32 x 32 or 64 x 64 with the goal that's the all characters can have a similar size. This size standardization maintains a strategic distance from bury class variety among characters. Bilinear, Bicubic addition procedures are a couple of strategies for size standardization.
- **Thresholding:** The thresholding is a procedure where the dark scale picture or any shading picture is changed over into the double picture. Given a limit, T somewhere in the range of 0 and 255, supplant all the pixels with a dim level lower than or equivalent to T with dark (0), the rest with white (1). On the off chance that the threshold is excessively low, it might decrease the number of items and a few articles may not be unmistakable. On the off chance that it is excessively high, we may incorporate undesirable foundation data. The suitable threshold esteem picked can be applied all around or locally.

Through the scanning, the procedure is the computerized picture of the original record is caught. While OCR optical scanners by and large comprise of a vehicle system in addition to a detecting gadget that changes over light force into grey levels. Printed reports, as a rule, comprise of dark print on a white foundation;

henceforth, when performing OCR, it is a regular practice to change over the staggered picture into a bi-level picture of highly contrasting. Frequently, this procedure is known as thresholding, is performed on the scanner to spare memory space and computational exertion

- **Binarization:** Binarization is a strategy for changing a dim scale picture into a highly contrasting picture through Thresholding.
- **Skew Detection:** For a record filtering process, there can be the skewness. The skewness ought to be evacuated in light of the fact that it decreases the exactness of the report. The slanted point is determined and with the assistance of slant edge, the slanted lines are made level.
- **Thinning:** To discover the highlights of the articles, the limit recognition of the picture is finished. For limit identification, different capacities can be applied to the items which are accessible in the MATLAB.

2.1.4 Feature Extraction and Recognition

The pre-prepared picture fills in as the contribution to this and every single character in the picture is discovered. The picture from the extraction arrange is corresponded with all the layouts which are preloaded into the framework. When the connection is finished, the layout with the greatest related worth is announced as the character present in the picture. The systems for extraction of such highlights are regularly isolated into three principle gatherings, where the highlights are found from:

- The conveyance of focuses.
- Transformations and arrangement extensions.
- Structural examination.

The various gatherings of highlights might be assessed by their affectability to commotion; also, disfigurement and the simplicity of execution and use.

2.1.5 Post-Processing

After the acknowledgement arranges, if there are some unrecognized characters found, those characters are given their importance in the post-handling stage. Additional formats can be added to the framework for giving a wide scope of similarity checking in the framework's database

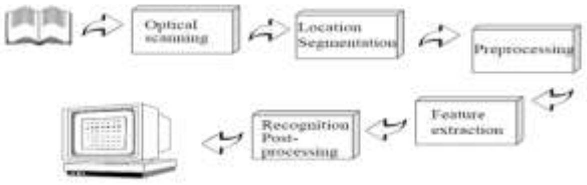

Fig 1: Components of OCR

2.2 Comparison of OCR techniques

Various techniques used for the design of OCR by their characteristics.

- **Matrix Matching:** Matrix Matching transforms each symbol inside a matrix into a sequence, then matches the sequence to an index of recognized characters. This awareness is highest on single-column pages with monotype and uniformity.
- **Fuzzy Logic:** Fuzzy logic is a custom-valued logic which allows for the interpretation of optimal variables amongst traditional assessments such as yes/no, true/false, black/white etc. The effort is made to relate the machine code to a more human-like way of rational thought. Fuzzy reasoning is used where the solutions have no clear truth or false meanings and are concerned uncertainly.
- **Feature Extraction:** This approach describes increasing character by the existence or absence of main characteristics like height, weight, length, loops, curves, stems and other individual character traits. Features extraction is a successful option for OCR paper, plasma processing, and top-quality photographs.
- **Structural Analysis:** Structural Analysis defines characters by analyzing the graphic, sub-vertical, and horizontal histograms of their sub-feature shapes. Its power to restore character is perfect for the text of poor quality and newsprints.
- **Neural Networks:** This technique simulates how the biological neural network functions; it tests the pixels in each illustration and compares them to a defined index of pixel patterns of personality. Modified papers and broken text are perfect for the ability to understand the characters by abstraction. Neural networks are suitable for specific challenges, such as analyzing stock exchange data or identifying similarities in visual patterns. Neural Networks are effective in many of these methods than others.

3. WORKING OF OCR

LSTMs are perfect for studying sequences but decelerate considerably once the lot of states becomes too huge. There are scientific findings that indicate asking an LSTM to discover a long series is better than that of a short segment of various classes. Tesseract was built in Python from OCRopus concept, which was a fork of a C++ LSMT named CLSTM. CLSTM is an application in C++ of the LSTM recursive machine learning framework utilizing the computational computation library Eigen.

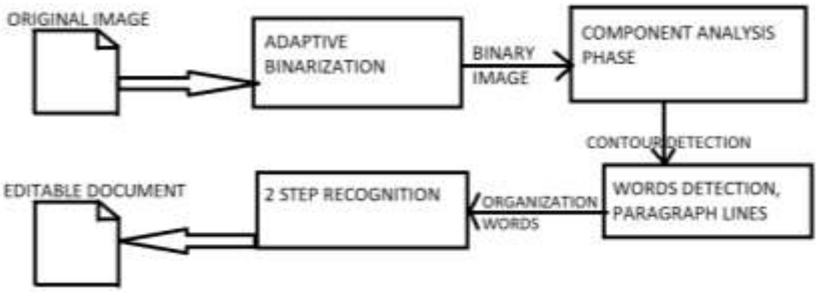

Fig 2: working mechanism

3.1 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) could be an information handling worldview that's enlivened by the approach natural sensory systems, for instance, the mind, method information. The key element of this worldview is that the novel structure of the info handling framework. it's created out of innumerous deeply interconnected making ready elements (neurons) operating mutually to require care of specific problems. ANNs, kind of like people, learn by model. associate ANN is meant for a selected application, for instance, pattern recognition or data characterization, through a learning procedure. Learning in natural frameworks includes acclimations to the conjugation associations that exist between the neurons. this is often valid for ANNs additionally.

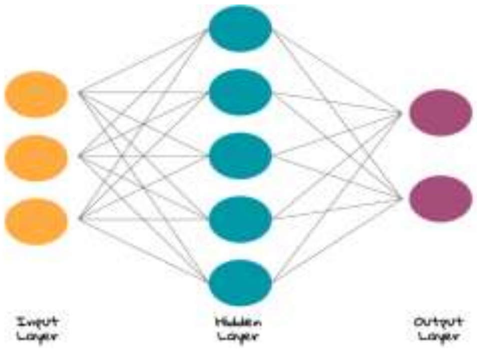

Fig 3: ANN Architecture

3.2 OCR process flow using Tesseract

Tesseract is an open-source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images. It supports a wide variety of languages. Tesseract doesn't have a built-in GUI. Tesseract is compatible with many programming languages and frameworks through wrappers. It can be used with the existing layout analysis to recognize text within a large document, or it can be used in conjunction with an external text detector to recognize text from an image of a single text line.

Tesseract 4.00 includes a new neural network subsystem configured as a text line recognizer. It has its origins in OCRopus' Python-based LSTM implementation but has been redesigned for Tesseract in C++. The neural network system in Tesseract pre-dates TensorFlow but is compatible with it, as there is a network description language called Variable Graph Specification Language (VGSL), that is also available for TensorFlow.

To recognize an image containing a single character, we typically use a Convolutional Neural Network (CNN). Text of arbitrary length is a sequence of characters, and such problems are solved using RNNs and LSTM is a popular form of RNN.

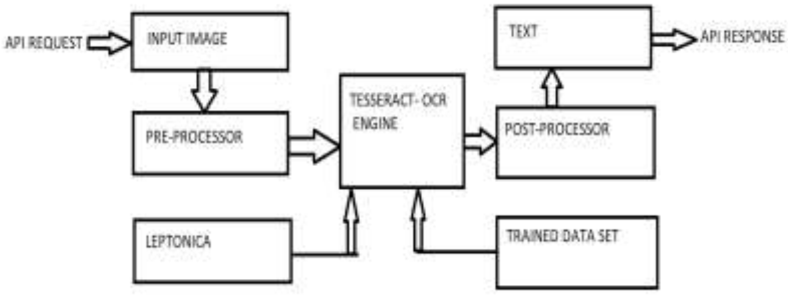

Fig 4: Tesseract process flow.

4. CONCLUSIONS

This paper says about OCR program for identification the of offline handwritten characters. The devices are expected to produce excellent performance. In the handwriting recognition schemes, preprocessing techniques used in text photos is introduced as an initial stage. The most significant of these is the attribute extraction stage in optical character recognition.

5.ACKNOWLEDGEMENT

We are sincerely thankful to Jain University for providing me with the opportunity to write a research paper on this topic. We are also thankful to Mr Subarna Panda, Assistant Professor and Dr Lakshmi JVN, Associate Professor for guiding me in every single stage of this research paper and supporting throughout the process in preparing a meaningful paper. We are also thankful to Dr.N.M Nachappa (Head of School-CS& IT) of Jain (Deemed-to-be University) who have helped me during the course of this research paper in different ways. Through this paper, we have learnt how information security & cyber forensics function in different platform.

6. REFERENCES

[1] “Feature-Based Recognition of Handwritten Kannada Numerals – A Comparative Study”, International Conference on Computing, Communication and Applications (ICCCA),22-24 Feb 2012.
[2] “Handwritten Character Recognition of South Indian Scripts: A Review”, National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011.
[3] Design and comparison of segmentation driven and recognition driven Devanagari OCR. In Proceeding 2nd Int. Conf. Document Image Anal. Libraries, 2006,pp.1-7.
[4] Offline Recognition of Devanagari Script: A Survey. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol.41, No.6, 2011, pp.782-796.
[5] A complete OCR System for Continuous Bengali Characters. T ENCON 2003, IEEE conference on convergent Technologies for Asia-Pacific Region Vol.4, 2003, pp.1372-1376.
[6] Historical Review of OCR Research and Development. Proceeding IEEE, Vol.80, No.7, 1992, pp.1029-1058. [2] Chaudhari, A.A., Ahmad, E.
[7] http://www.ehow.com/about_6552517_definition-handwriting-recognition.html
[8] http://www.ehow.com/info_8085619_software-handwriting-recognition.html
[9] http://www.cse.yorku.ca/course_archive/2004-05/F/4441/HandwritingRecognition.pdf
[10] artificial neural network-based character recognition using backpropagate (International Journal of Computers & Technology www.ijctonline.com ISSN: 2277-3061 Volume 3, No. 1, AUG, 2012).