Text Recognizer Using Optical Character Recognition Technique

Madhuram M [1], Sandeep M [2], Praveen Hari Krishna N [3], Aditya S [4]
[1]Assistant Professor, Department of Computer Science, SRM Institute of Science and Technology, Chennai-600089, Tamil Nadu, India.
[2,3,4] Department of Computer Science, SRM Institute of Science and Technology, Chennai-600089, Tamil Nadu, India

Abstract – In the running world, there is a growing demand for the machines to acknowledge characters in automatic data processing system once data is scanned through paper, documents as we all know that we have range of newspapers and books that are in written format relating to completely different subjects. Of late, there is an enormous demand in storing the data in these paper documents in to a memory disk then later reusing this information by looking process. One easy ways to store data from these paper documents in to automatic data processing system is to first scan the documents then store them as pictures. However to recycle a data it is terribly troublesome to browse the individual contents and looking the contents type these documents line-by-line and word-by-word. The explanation for this problem is that the font characteristics of the characters in paper documents are completely different to font of the characters in automatic data processing system. As a result, pc is unable to acknowledge the characters whereas reading them. This idea of storing the contents of paper documents in memory place then reading and looking the content is named DOCUMENT process. Generally during this document process we want to method the knowledge that is relating to languages although nation within the world. For this document process, we want a software known as character recognition system. This method is additionally known as document image analysis (DIA).

Index Terms – Optical Character recognition, Neural Network, Hand written character recognition, Feature Extraction.

1. INTRODUCTION

Optical character recognition refers to the branch of technology that involves reading text from paper and translating the pictures into a type that the pc will manipulate (for example, into computer code codes). Associate in Nursing OCR system permits you to require a book or an article, feed it directly into Associate in nursing computer file, and so edit the file employing an applications program. All OCR systems embody Associate in nursing optical scanner for reading text, and complex computer code for analyzing pictures. Most OCR systems use a mixture of hardware (specialized circuit boards) and computer code to acknowledge characters, though some cheap systems sleep with entirely through computer code. Advanced OCR systems will scan text in massive type of fonts, however they still have problem with written text. It is the mechanical or electronic translation of scanned pictures of written, written or written text into machine-encoded text. Its wide accustomed convert books and documents into electronic files, to computerize a recordkeeping system in Associate in nursing workplace, or to publish the text on a web site. OCR makes it doable to edit the text, hunt for a word or phrase, store it additional succinctly, show or print a replica freed from scanning artifacts, and apply techniques admire MT, text-to-speech and text mining to that. OCR could be a field of analysis in pattern recognition, AI and pc vision. OCR systems need standardization to scan a selected font; early versions required to be programmed with pictures of every character, and worked on one font at a time. "Intelligent" systems with a high degree of recognition accuracy for many fonts area unit currently common. Some systems area unit capable of reproducing formatted output that closely approximates the initial scanned page as well as pictures, columns and different non-textual elements. Optical character recognition (OCR) is one in every of the foremost widespread areas of analysis in pattern recognition because of its vast application potential. However, most of the accessible strategies agitate the popularity of the Roman script and a few of the oriental scripts like Kanji, Kana, etc. The aim of this OCR system is to require written English characters as input, method the character, train the neural network formula, to acknowledge the pattern and modify the character to a beautified version of the input. This work is restricted to English characters and numerals solely. It is conjointly useful in recognizing special characters. It is any developed to acknowledge the characters of various languages. One in every of the first means that by that computers area unit endowed with human skills is thru the utilization of a neural network. Neural networks area unit significantly helpful for determination issues that cannot be expressed as a series of steps, admire recognizing patterns, classifying them into teams, series prediction and data processing...

2. RELATED WORK

Various techniques used for the look of OCR by their characteristics. Maintaining the Integrity of the Specifications

- Fuzzy Logic
- Neural Network

2.1. Fuzzy Logic

Fuzzy logic is associate approach to computing supported "degrees of truth" instead of the standard "true or false" (1 or 0).

mathematical logic includes zero and one as extreme cases of truth (or "the state of matters" or "fact") however additionally includes the varied states of truth in between so, as an example, the results of a comparison between 2 things might be not "tall" or "short" however ".38 of tallness."

Fuzzy logic looks nearer to the manner our brains work. we tend to mixture information and kind variety of partial truths that we tend to mixture more into higher truths that successively, once bound thresholds ar exceeded, cause bound more results like motor reaction. the same quite method is employed in neural networks, knowledgeable systems and different computer science applications. mathematical logic is crucial to the event of human-like capabilities for AI, typically remarked as artificial general intelligence: the illustration of generalized human psychological feature talents in package so, long-faced with associate unknown task, the AI system may realize an answer..

2.2. Neural Network

Neural Networks square measure a key piece of a number of the foremost sure-fire machine learning algorithms. The event of neural networks are key to teaching computers to suppose and perceive the planet within the method that humans do. Primarily, a neural network emulates the human brain. Brains cells, or neurons, square measure connected via synapses. this can be abstracted as a graph of nodes (neurons) connected by weighted edges (synapses).

3. PORPOSED MODELLING

The proposed method consists of four steps. They are image acquisition, preprocessing of an image, segmentation of characters, feature extraction of characters and character recognition.

3.1. Image Acquisition

The written character to be recognized is no heritable by exploitation Associate in nursing optical scanner.

3.2. Image Preprocessing Pre-processing includes many operations over the scanned image, in order that input image becomes appropriate and comfortable for applying to additional sub sections. Essentially the target of pre-processing is to boost the standard of scanned input image. Noise removal, mathematical operations may be processed during this Preprocessing section. It includes binarization, boundary detection, segmentation, thinning. It performs the many operations over the scanned computer file.

3.2.1 Binarization

Binarization plays a very important role in preprocessing. It is necessary to convert a color image into black and white format. Thus, we will method over that black and white image. Essentially separation of background and actual image space referred as foreground of a scanned image is named binarization

Phase	Description	Approaches
Acquisition	The process of acquiring image	Digitization, binarization, compression
Pre-processing	To enhance quality of image	Noise removal, Skew removal, thinning, morphological operations
Segmentation	To separate image into its constituent characters	Implicit Vs Explicit Segmentation
Feature Extraction	To extract features from image	Geometrical feature such as loops, corner points Statistical features such as moments
Classification	To categorize a character into its particular class	Neural Network, Bayesian, Nearest Neighborhood
Post-processing	To improve accuracy of OCR results	Contextual approaches, multiple classifiers, dictionary based approaches

3.2.2 Boundary Detection

The binarized image is currently applicable for boundary detection. During this operation the boundaries of scanned image is detected. It detects all the boundaries of image. It is necessary to notice the boundaries thus on choose a private character

3.2.3 Segmentation

This is necessary operation of OCR as rate of recognition is directly proportional to segmentation. During this method, each individual character is separated. This isolates the various sub-parts of a picture. It is wont to separate pixels of a picture

as per the contents in knowledge like words, paragraph etc. Figures and Tables

3.2.4 Thinning

Thinning is employed to scrub the scanned input image. This method deletes the dark points within the image

3. Feature Extraction

For the accuracy of OCR system, the suitable Feature Extraction methodology ought to be elect. Whereas process over the image some options ought to be separated. The standard options square measure Edges, Corners, Ridges, etc. This methodology of separation is named as Feature Extraction. The accuracy of AN OCR technique depends on choice of correct feature extraction methodology

4. Classification

The feature-extracted knowledge should have versed the method of Classification. This method classifies the extracted individual character in correct manner.

5. Post-Processing

This is the last and a crucial part of OCR technique. It includes completely different operations like Grouping, Error detection and correction. Regardless of the knowledge being operated through completely different operations admire, binarization, segmentation, Feature extraction, Classification etc. is fed to post-processing. Meaning completely different options of input scanned image square measure extracted. That feature extracted knowledge is a personal character. It is unable to urge careful info from that individual character. Therefore, it is necessary to gather individual character in applicable and consecutive manner. The method of grouping individual characters of constant contents to make a string is termed as Grouping. By victimization error detective work and correcting algorithms, errors may also be eliminated. Finally, we have a tendency to get the recognized output character

4. RESULTS AND DISCUSSIONS

We take three characters with its 5 pattern and examine that character set.

Character	No. of patterns given	Recognition	Not Recognized	Rate (%) of Recognition
L	5	5	0	100%
M	5	5	0	100%
O	5	5	0	100%

Table 1 Resultant Graph of the Proposed System

TWDB		
Test Set	Train Set	Recognition Rate
2793 chars	11173 chars	95.44%
HWDB		
Test Set	Train Set	Recognition Rate
1351 chars	5407 chars	94.62 %

Table 2 Printed and handwritten document results

REFERENCES

[1] N. venkata rao, Dr. A.S.C.S.Sastry, A.S.N.Chakravarthy, kalyanchakravarthi P, "Optical character recognition technique algorithm", in Journal of Theoretical and Applied Information Technology 20th January 2016. Vol.83. No.2..

[2] Najib ali Mohamed Isheawy, "Optical Recognition System(OCR) System" in IOSR Journal of computer Engineering, vol. 17 Issue 2, Version II, March-April 2015

[3] Chattopadhyay, T., Ruchika Jain, and Bidyut B. Chaudhuri. "A novel low complexity TV video OCR system." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[4] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." Communications, Devices and Intelligent Systems (CODIS), 2012 International Conference on. IEEE, 2012.

[5] Bhagirath Kumar, Niraj Kumar, Charulata Palai, Pradeep Kumar Jena, Subhagata Chattopadhyay, "Optical Character Recognition using Ant Miner Algorithm: A Case Study on Oriya Character Recognition", International Journal of Computer Applications, 2013, volume 61-No.3.

[6] Muthumani.I, Uma Kumari C.R, "Online Character Recognition of Handwritten Cursive Script", International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.

[7] Bhansali, M., & Kumar, P, 2013, An Alternative Method for Facilitating Cheque Clearance Using Smart Phones Application. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2(1), 211-217.

[8] Swapnil A. Vaidya, Balaji R. Bombade, A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186.

[9] Aradhana A Malanker , Prof. Mitul M Patel , Handwritten Devanagari Script Recognition: A Survey, IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 9, Issue 2 Ver. II (Mar – Apr. 2014), PP 80-8.

[10] Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav[May 2013] Optical Character Recognition using MATLAB International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)Volume 2, Issue 5, May 2013.