Optical Character Recognition System

Shashank Sharma[1], Ms. Anamika Jain[2]

[1]*Student, Computer Science Department, Poornima Group of Institutions, Jaipur*

[2]*Asst. Prof., Dept. of Computer Science, Poornima Group of Institutions, Jaipur*
[1]*2014pgicsshashank@poornima.org,* [2]*anamika.jain@poornima.org*

Abstract

An Optical Character Recognition (OCR) System is an advanced system by which embedded textual information is repossessed through digitizing from graphical media like an image. Developing such software system for handwritten characters with different handwriting is still a challenge. Sometimes it contains different letters in other than English language. This paper gives an entire Optical Character recognition (OCR) device for digital camera captured picture/pictures embedded textual documents for handheld gadgets. At the start, text areas are extracted and skew corrected. Then, the one's areas are binariezed and segmented into strains and characters. Characters are surpassed by the popularity module. Experimenting with a set of one hundred enterprise card photographs, captured by using way of cell phone camera, we've got accomplished a most reputation accuracy of approx 90%. In contrast to Tesseract, an open deliver computer-primarily based powerful OCR engine, present recognition accuracy is genuinely really worth contributing. Furthermore, the superior approach is computationally inexperienced and consumes low reminiscence as a way to be applicable accessible held gadgets.

Keywords: *text recognition, character analysis, OCR*

1. Introduction

Optical Character Recognition (OCR) is a chunk of a software program that converts published textual content and snapshots right into a digitized form such that it is able to be manipulated by a device. Not like human mind which has the capability to very without difficulty recognize the textual content/ characters from a photograph, machines aren't smart enough to perceive the statistics available within the photograph. Consequently, a large quantity of studies efforts had been recommends that attempts to convert a document photograph to format understandable for the machine. OCR is a complicated trouble due to the form of languages, fonts and styles wherein textual content may be written, and the complicated guidelines of languages etc. for this reason, techniques from exceptional disciplines of computer technology (i.e. photo processing, sample type and herbal language processing and so on. are hired to deal with exclusive challenges. This paper introduces the reader to the hassle. It enlightens the reader with the ancient views, applications, demanding situations and strategies of OCR.

2. Literature Survey

Character Recognition isn't always a brand new trouble however its roots can be traced again to structures before the inventions of computer systems. The earliest OCR structures were no longer computer systems however mechanical gadgets that have been able to understand characters, however very gradual pace and occasional accuracy. In 2016, Noman Islam, Zeeshan Islam and Nazia Noor did the survey on Optical Character Recognition System [1], in which an outline of numerous strategies of OCR has been studied. An OCR is not an atomic technique but accommodates various levels which include acquisition, preprocessing, segmentation, characteristic extraction, type and submit-processing. Each of the steps is discussed in detail in this paper. The usage of a combination of these techniques, a green OCR system may be developed as a destiny work. The OCR system also can be utilized in one of kind practical packages which include quantity-plate recognition, clever libraries and diverse different real-time programs.

Najib Ali Mohamed Isheawy And Habibul Hasan did the research on Optical Character Recognition System [2], in which, The Grid infrastructure used within the implementation of Optical Character Recognition System can be successfully used to quick up the interpretation of picture-primarily based documents into text documents. The community has been educated and examined for some of the extensively used fonts. The popularity of recent font characters via the machine may be very clean and quick. We can edit the facts of the documents greater conveniently and we are able to reuse the edited data as and when required.

Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri did the research on OCR system for handheld devices [3], in which the complete study of the OCR is done with the output efficiency of 92.74% for handheld devices.

3. Background of OCR

OCR may be implemented both off-line and online. Within the off-line recognition, the writing is generally captured optically with the aid of a scanner and the completed writing is to be had as a photo. However, in the on-line device, the two-dimensional coordinates of success factors are represented as a characteristic of the time. And the orders of strokes made by means of the author are also available. The online techniques have been shown to be advanced to their offline opposite numbers in spotting handwritten characters because of the temporal facts available with the former. The input for the OCR hassle is pages of scanned text. To carry out the OCR, our utility has to undergo three critical steps:

1) Segmentation: when an image is given as an input, discover character glyphs (primary devices representing one or extra characters, typically contiguous).

2) Feature Extraction: From every glyph picture, extract capabilities for use as entering of ANN. this is the most vital a part of this approach.

3) Classification: train the ANN by using the training pattern. Then given new glyph, classify it.

The second step is the hardest in the sense that there may be no obvious way to gain those capabilities. To gain better know-how, strategies, and answers regarding the approaches that we need to follow, we studied the numerous research papers on current OCR structures. These types of research helped us with clarifying our target desires.

The simple steps involved in Optical Character Recognition are:-

a) Image as an input or image acquisition
b) Image preprocessing
c) Analysis and character segmentation
d) Characteristics extraction
e) Learning and analysis
f) Inage Post-processing

3.1. Types of Character Recognition System

There have been multiple directions in which research on OCR has been accomplished all through past years. This phase discusses one-of-a-kind sorts of OCR structures have emerged because of these researchers. We can categorize those systems based totally on picture acquisition mode, textual connectivity, font-regulations etc. Figure 1 categorizes the OCR gadget.

Handwriting character reputation is a completely tough activity because of the extraordinary writing style of the person as well as distinct pen moves by the person for the equal individual. These systems can be divided into two sub-classes i.e. online and rancid-line systems. The former is completed in actual-time while the customers are writing the individual. They are much less complicated as they are able to seize the temporal or time based totally facts i.e. speed, curves, the range of strokes made, the course of the writing of strokes and many others. Similarly, there no want for thinning strategies as the trace of the pen is few pixels extensive. The offline popularity structures operate on static statistics i.e. the entry is a bitmap. Consequently, it is miles very difficult to perform popularity.

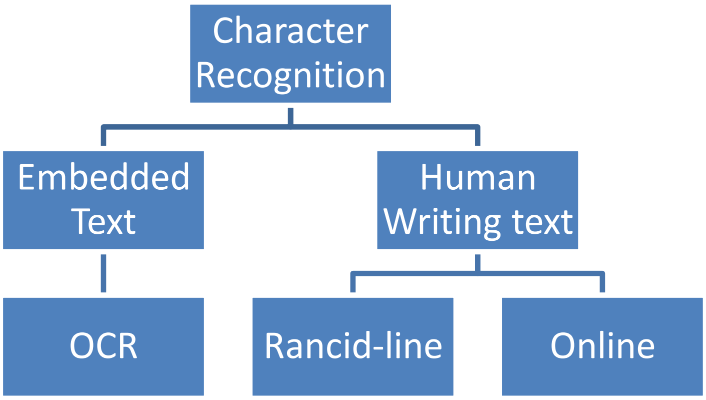

Figure 1. Types of OCR

4. Methods in OCR

An OCR system can be completed using following phases:

4.1. Image Input or Image acquisition

Picture acquisition is the preliminary step of OCR that contains acquiring a virtual photograph and changing it into the proper form that may be without difficulty processed through the laptop. This can involve quantization as well as compression of the photo. A unique case of quantization is binarization that involves handiest degrees of the image. In the maximum of the cases, the binary image suffices to represent the photograph. The compression itself may be lossy or lossless. An overview of numerous picture compression techniques has been provided in.

4.2. Image Preprocessing

Subsequent to picture acquisition is pre-processing that pursuit to beautify the quality of the picture. One of the pre-processing strategies is thresholding that ambitions to binaries the photograph primarily based on a few threshold cost. The threshold cost may be set at neighborhood or international stage.

Special forms of filters which include averaging, min and max filters may be implemented. As an alternative, extraordinary morphological operations which include erosion, dilation, commencing and last can be finished.

4.3. Analysis and Character Segmentation

In this phase, the photograph is segmented into characters earlier than being passed to category phase. The segmentation can be achieved explicitly or implicitly as a byproduct of class section. Further, the alternative phases of OCR can assist in supplying contextual information useful for segmentation of the picture.

4.4. Character Extraction

In this phase, numerous functions of characters are extracted. Those functions uniquely discover characters. The choice of the right capabilities and the whole number of features to be used is an important research question. Specific kinds of features which includes the photo itself, geometrical features (loops, strokes) and statistical characteristic (moments) can be used. Subsequently, various techniques which include predominant element analysis can be used to lessen the dimensionality of the picture.

4.5. Learning and Analysis

In this phase, it is miles described because the method of classifying a character into its suitable class. The structural approach to classification is based on relationships present in photo additives. The statistical approaches are based totally on the use of a discriminate function to categories the photo. Some of the statistical type procedures are the Bayesian classifier, selection tree classifier, neural network classifier, nearest neighborhood classifiers and many others. Finally, there are classifiers based on a syntactic approach that assumes a grammatical technique to compose an image from its substituent.

4.6. Image Post-processing

Once the textual content has been categorized, there are numerous tactics that can be used to enhance the accuracy of OCR effects. One of the tactics is to apply a couple of classifier for the class of the picture. The classifier may be used in cascading, parallel or hierarchical style. The consequences of the classifiers can then be mixed using various approaches.

With a view to improving OCR outcomes, the contextual analysis also can be done. The geometrical and file context of the picture can help in lowering the possibilities of mistakes. Lexical processing based totally on Markov fashions and dictionary also can assist in improving the effects of OCR.

All of the above phases can be described by using following images in which all if the phases are applied to a different types of embedded text which is not a standard text at all. When the image is processed through the OCR system then text is extracted as following.

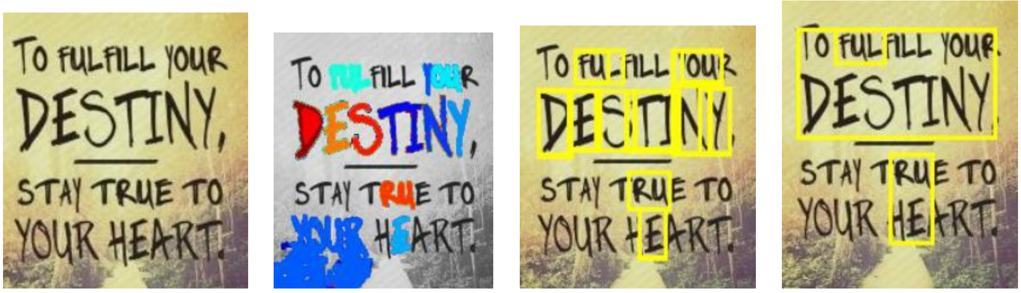

Figure 2. (a) Input Image (b) Character extraction (c) Analysis (d) Post-Processing

4. Conclusion

In this paper, a top-level view of diverse strategies of OCR has been presented. An OCR is not an atomic manner but incorporates numerous phases including acquisition, preprocessing, segmentation, characteristic extraction, and category and post-processing. Each of the steps is mentioned in detail in this paper. The usage of an aggregate of those strategies, an efficient OCR device can be evolved as a future work. The OCR gadget can also be used in one-of-a-kind sensible programs consisting of number-plate popularity, clever libraries and various different actual-time packages.

No matter the good-sized amount of studies in OCR, the popularity of characters for the language consisting of Arabic, Sindhi and Urdu nonetheless remain an open assignment. An overview of OCR techniques for those languages has been deliberate as a destiny work. Some other essential region of research is multi-lingual person popularity gadget. In the end, the employment of OCR systems in realistic programs remains an active region of research.

Acknowledgement

The authors would like to thank Department of Computer Science of Poornima Group of Institutions, Jaipur for the kind support.

References

[1] Character Recognition System for Camerabased Handheld Devices", IJCSI International Journal of Computer Science Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri, "Design of an Optical Issues, Vol. 8, Issue 4, No 1, July 2011

[2] Najib Ali Mohamed Isheawy And Habibul Hasan, "Optical Character Recognition (OCR) System", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. II (Mar – Apr. 2015), PP 22-26

[3] Noman Islam, Zeeshan Islam and Nazia Noor, "A Survey on Optical Character Recognition System", Journal of Information & Communication Technology-JICT Vol. 10 Issue. 2, December 2016

[4] Niall Anderson, Editor, "Optical Character Recognition", Released under Creative Commons - Attribution-NonCommercial-ShareAlike v3 Unported (International) (2011)

[5] Qadri, M.T., & Asif, "Automatic Number Plate Recognition System for Vehicle Identification Using Optical Character Recognition", International Conference on Education Technology and Computer, Singapore, (2009)