International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

PHOTO OPTICAL CHARACTER RECOGNITION MODEL

Shubhangi Singh[1], Pranjal Sakargayan[2], Ajitesh Singh[3]

[1,2,3]*Jaypee University of Engineering and Technology, Guna, MP*

Abstract - *Optical Character Recognition (OCR) is a topic in which many have shown interest for research and study in the past years. It's the process of converting images of typed, handwritten or printed text into machine-encoded text which can be in text format or as per required. Intensive research has been done on optical character recognition (OCR) and many articles have been published on this topic during the last few decades. Different techniques for preprocessing and segmentation have been already been surveyed n number of times in the past years. This paper aims at summarizing the research done so far done in the field of OCR. It provides an overview of different aspects of OCR and discusses possible futures scopes of our version of OCR model.*

Key Words: **OCR, Preprocessing, Classifier, Image Segmentation, Handwritten Character Recognition, Image processing, Feature extraction, Neural Network.**

1. INTRODUCTION

Do you ever find it difficult to read a doctor's prescription or maybe your friend's class notes ? That happens quite a number of times and when it comes processing the hand-written documents manually on a large scale, it gets even more monotonous. Mostly computers have things done their way only if we "talk" them through various devices such as keyboards so they can figure out what we want them to do. But when it comes to processing more human kinds of information, like a printed book or a letter scribbled by a child, or more such handwritten material; computers work much harder. That's where optical character recognition (OCR) comes in to use. It's a type of software (program) that can automatically analyze printed text and turn it into a form that a computer can process more easily like a text file, pdf etc. Character recognition is not a new problem, but its roots can be traced back to time even before the inventions of computers. The earliest OCR systems were devices that were able to recognize all characters, but very slow speed and low accuracy hence resulting in not so trustworthy results. In 1951, M. Sheppard invented a reading and robot that can be considered as the earliest work on modern OCR [1]. It can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. This survey is restricted to offline systems. Presence of standard database is also necessary for handwriting recognition research. As per our model, MNIST data has been put to use.

In this paper, we present a review of the work done by the OCR model on English language scripts and digits from 0-9. The review is organized into **VII Sections**. *Section I* covers the introduction. Then we move on to describe the complete overview of the method used in the model in *Section II* and then further discussing the various phases involved in *Section III*. Further, we talk about the various applications of Photo OCR model in *Section IV*. And last but not the least, in *Section V*, we discuss the scope of future work in our model and then conclude the paper.

II. The complete Methodology

Considering that there's only one single letter in the entire world. But still there are n number of ways to write/print that single letter.

Even then, the work of an OCR model would be difficult. Now imagine the same for so many characters actually present in the real world. Even with printed text, there's an issue, because books and other documents can be printed in subtly different fonts.

For e.g. In the picture above, every 'A' has been written in different form and in different angle but if one notices there's a very clear common thing between them and i.e. every 'A' shown in the picture above is made by two bent lines facing in the opposite direction and a mid-line connecting them. And that'll be our key to recognize the letter 'A'.

The above example of 'A' was just to give an overview of the way we'll try to recognize the letters. We'll follow almost a similar pattern for all such characters.

And the steps followed are discussed further in this paper in the later sections.

III. Major Phases involved in OCR:

Acquisition	The process of acquiring image.	Digitization, binarization, compression.
Pre-processing	To improve the 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 (loops, curves) from image.	Geometrical feature such as loops, corner points.
Classification	To categorize a character into the appropriate class.	Neural Network, Bayesian, Nearest Neighborhood.
Post-processing	To improve accuracy of the results produced by OCR.	Contextual approaches, multiple classifiers, dictionary based approaches.

A. Acquisition

Image acquisition is the initial step of OCR that includes obtaining a digital image and converting it into suitable form that can be easily processed by the computer. This can involve quantization as well as compression of image [3]. A special case of quantization is binarization (converting a pixel image to binary image). A binary image is one which has only two possible value for each pixel and the two colors used are black and white. However, compression of an image on the other hand can be a loss itself in the quality of the image. An overview of various image compression techniques has been provided in [2].

B. Pre-Processing

After we've acquired the image the next phase is called pre-processing that aims to improve the quality of the image. One of the pre-processing techniques is thresholding that aims to binaries the image based on some threshold value [2]. In this phase, there is a series of operations performed on the scanned input image which enhances the image making it suitable for the next phase called segmentation which takes the the gray-level character image and normalizes it into a window sized image.

C. Segmentation

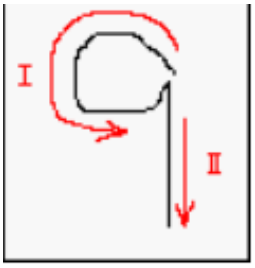
It's the process of partitioning a digital image to multiple segments. Our goal is simplify the representation of the image into something more meaningful and easy to analyze. This phase involves 2 major levels to help us segment a page containing text which are broadly classified as: page decomposition and word segmentation. Page decomposition might identify sets of logical components of a page, whereas word segmentation separates the characters of a sub-word. The segmentation can be performed explicitly or implicitly as a byproduct of classification phase [4].

D. Feature Extraction

Also known as Feature Detection or Intelligent Character Recognition (ICR). The segmented characters are then processes to extract different features. Based on these features, the characters are recognized. The selection of the appropriate features (loops, curves, lines etc.) which are to be used in this phase which'll help us identify various similarities and differences between all the characters is an important question for research and study. The number of features which one should use for feature extraction should also be chosen wisely. Different types of features can be considered for such a process such as the image itself, geometrical features (loops, strokes) and statistical feature (moments).

Trying feature extraction for the capital letter A:

Trying feature extraction for the digit 9:



E. Classification

After the characters have been extracted, they're then classified in the appropriate category. The classification stage is the major decision maker of the complete OCR model because when classifying a pattern, this phase often produces a set of hypothetical solutions instead of generating a unique solution. The later stage called as post-processing uses higher level information to select the correct solution. This phase follows two main procedures: syntax based (or structural) and statistical (or decision theoretic) classification. In the recent times, recognition using neural networks has provided a third procedure which is used widely now. Some of the statistical classification approaches are Bayesian classifier, decision tree classifier, neural network classifier, nearest neighborhood classifiers etc. as discussed in [5].

F. Post-Processing

The post-processing stage is the final stage which improvises upon the image recognition done by earlier phases by refining and filtering the decisions taken in the previous stage. In order to improve the results produced by the OCR model, context based analysis can also be performed. Contextual document analysis of the image can help in

reducing the chances of errors. Lexical processing based on Markov models and dictionary can also help in improving the results of OCR [5]. Post-processing is also above all the phases and has high level information to check the decisions made by the various classifiers in the classification phase. Most common job of this phase is spell check and correcting them. Spell checking can be as simple as looking up words in a dictionary.

IV. Applications

OCR helps in implementing a number of tasks such as mail sorting, bank cheque reading and signature verification [6]. OCR can also be used by organizations for automated form processing in places where data has to be available in printed form on a large scale. Other uses of OCR include processing utility bills, passport validation, pen computing and automated number plate recognition etc. [7]. Another useful application of OCR is helping blind and visually impaired people to read text [8].

Also can be used in CAPTCHA. It's a program that can generate and grade tests that human can pass but current computers programmers' i.e a robot cannot. In CAPTCHA, an image consisting of series of characters is generated which is made unclear by image distortion techniques, size and font variation, random segments, highlights, and introducing noise in the image. This model can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems.

Can be used for Invoice imaging which is widely used by many businesses applications and e-commerce websites to keep track of financial records and prevent a backlog of payments from pilling up. In government agencies and independent organizations, OCR simplifies data collection and dealing with such data on a large scale and reduces manual work.

V. Scope of Project

There's a huge scope for significant amount of research in OCR regarding recognition of characters for languages such as Arabic, Sindhi and Urdu; which is still an open challenge for many.

VI. Conclusion

In this paper, an overview of the phases involved in the working of an OCR model has been presented. An OCR model is not a one-step process but involves a number of steps which have briefly discussed in this paper. Using a combination of these along with some improved features can result in a much better OCR model which is better at recognizing hand written scripts.

From the survey, it is noted that the errors in recognizing handwritten English characters are mainly due to incorrect character segmentation of touching or broken characters due to poor feature extraction. Because of upper and lower modifiers of English text, many portions of two consecutive lines may also overlap and for such text proper segmentation and classification is needed for better accuracy in results produced. Many suggest that if in the post-processing stage we go by integrating a dictionary with the OCR system, then it can significantly reduce the misclassifications in printed as well as handwritten word recognitions.

VII. References

[1] Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141.

[2] Lund, W.B., Kennard, D.J., & Ringger, E.K. (2013). Combining Multiple Thresholding Binarization Values to Improve OCR Output presented in Document Recognition and Retrieval XX Conference 2013, California, USA, 2013. USA: SPIE

[3] Lazaro, J., Martín, J.L, Arias, J., Astarloa, A., & Cuadrado, C, 2010, Neuro semantic thresholding using OCR software for high precision OCR applications. Image and Vision Computing, 28(4), 571-578.

[4] Shaikh, N.A., Shaikh, Z.A., & Ali, G, 2008, Segmentation of Arabic text into characters for recognition presented in International Multi Topic Conference, IMTIC, Jamshoro, Pakistan, 2008. Pakistan: Springer.

[5] Ciresan, D.C., Meier, U., Gambardella, L.M., & Schmidhuber, J, 2011, Convolutional neural network committees for handwritten character classification presented in International Conference on Document Analysis and Recognition, Beijing, China, 2011. USA: IEEE.

[6] Shen, H., & Coughlan, J.M, 2012, Towards A Real Time System for Finding and Reading Signs for Visually Impaired Users. Computers Helping People with Special Needs. Linz, Austria: Springer International Publishing

[7] Bhavani, S., & Thanushkodi, K, 2010, A Survey On Coding Algorithms In Medical Image Compression. International Journal on Computer Science and Engineering, 2(5), 1429-1434.

[8] Bhammar, M.B., & Mehta, K.A, 2012, Survey of various image compression techniques. International Journal on Darshan Institute of Engineering Research & Emerging Technologies, 1(1), 85-90

[9]MNIST Database of Handwritten digits: http://yann.lecun.com/exdb/mnist/

AUTHORS

Author 1 : Shubhangi Singh

Author 2 : Pranjal Sakargayan

Author 3 : Ajitesh Singh