ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

Implementing a System for Recognizing Optical Characters

Hewa Majeed Zangana

Department of Computer Science / College of Computer Science and IT / Nawroz University / Kurdistan Region of Iraq

Abstract: In the current paper we present a system of characters recognition by taking the photo of character with the identity of symbolic. In the proposed system we are going to make a scan in kind of optical for input character in order to be digitized. After that every character will be segmented and located and after that it will be obtained as a photo to be processed for normalization and even for reducing noise. After that it will be classified. Then from the obtained extraction we can find various techniques like weakness and strengths. Next step will be grouping the characters which identified in order to obtain the original string of symbols and we can apply the context in order to fix and detect false. The results show us that the system is working well and the recognition is really good.

Keywords: Characters, Recognition, Context

1. INTRODUCTION

The system proposed in a program, developed in Matlab environment, which provides the ability to insert a character in an image. It is agree that making a machine to do what human can do is a dream, for example reading is one of the most important functions that humans are doing. However, this dream is becoming true day by day and researchers and working on this by many ways, where nowadays artificial intelligence is focusing on pattern recognition and in this field it is also focusing on the applications of character recognition and even many organizations and companies are designing systems for character recognition by many application and even that it is facing some challenges to make machines be able to read like humans and have the same capabilities. Recognizing characters is challenging some problems with the optical characters. Although, it is performed to be off line optical recognition for characters especially after completing the printing and writing, and to be online recognition to recognize characters as they have been drawn or written. Printed characters and even hand written characters could be recognized, but what we are always looking for is the performance where especially it is depending on the quality of files that been entered. Next step of challenging reviewed by many researchers is the online and the offline cursive writing. To get new ideas in the recognition of pattern, the classifying of characters could be tested, but where the experiments results are conducted on isolated characters, here the results are not necessary in case of immediately relevant to optical character recognition. Maybe more striking than the improvement of the accuracy and limit in methods of classification has been decreased in cost. The old devices of optical character recognition equipments were some optical hardware like the optical page reader of the company of IBM in order to read typed earning reports at the social security administration which cost more than two million dollars and some electronic and some high expensive scanners. Nowadays, the software of optical character recognition is often add on to scanner of desktop which is not costly. The main goal is to examine some details in examples of the false which committed by the proposed system.

2. PROPOSED SYSTEM

The general technique is very simple to describe. The proposed optical character recognition system will contain some components and they are presented in figure 1. The install is illustrated, where to digitize the analog file by the optical scanner will be the first step in the system. After that the area which containing characters will be located and every symbol extracted by the process of segmentation. After that applying a preprocessing on the extracted symbols and then we are going to reduce the noise and eliminate it in order to make it easier the feature extraction to be prepared for the coming step. After that we are going to comparing the description of the classes of symbols which are gained by a phase of previous learning with the extracted features in order to find the identity of the symbol. Then to reconstruct the numbers and words of the original string we are going to use the contextual information.

DOI 10.17148/IJARCCE.2018.71013

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

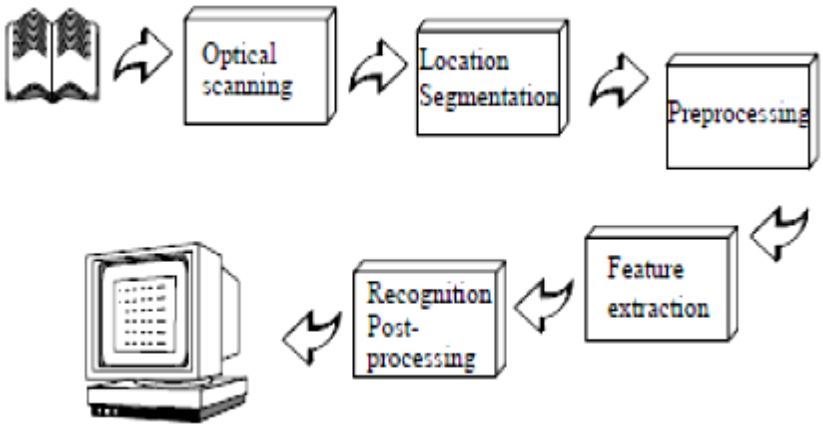

Fig. 1: Proposed System Steps

Now we are going to explaining the steps of the proposed system one by one.

2.1. Optical Scanning

Here the digital photo or image of the prime document is taken in the process of scanning and scanning in optical is used in the OCR where in general it is consists of a transport mechanism plus a sensing device which converts the intensity of light into levels of gray. Usual documents which are printed consist of a white background and black print. Hence, when performing OCR, it is common practice to convert the multilevel image into a bilevel image of black and white. Often this process, known as thresholding, is performed on the scanner to save memory space and computational effort. The thresholding process is important as the results of the following recognition are totally dependent of the quality of the bi-level image. Still, the thresholding performed on the scanner is usually very simple. A fixed threshold is used, where gray-levels below this threshold is said to be black and levels above are said to be white. For a high-contrast document with uniform background, a pre chosen fixed threshold can be sufficient. However, a lot of documents encountered in practice have a rather large range in contrast.

2.2. Location and segmentation

Segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. The main problems in segmentation may be divided into four groups:

3. *Extraction of touching and fragmented characters.*
4. *Distinguishing noise from text.*
5. *Mistaking graphics or geometry for text.*
6. *Mistaking text for graphics or geometry.*

2.3. Preprocessing

The image resulting from the scanning process may contain a certain amount of noise. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. The most common technique for smoothing is moves a window across the binary image of the character, applying certain rules to the contents of the window. The normalization is applied to obtain characters of uniform size, slant and rotation. To be able to correct for rotation, the angle of rotation must be found. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew.

2.4. Feature Extraction

The techniques for extraction of such features are often divided into three main groups, where the features are found from:

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

• The distribution of points.
• Transformations and series expansions.
• Structural analysis.

In MATLAB mat2cell command is used for the extraction of image in form of a cell for correlating with the saved templates.Fig.2 shows extraction of character in Matrix form.

2.5. Template-matching and correlation techniques

These techniques are different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines. However, this technique is sensitive to noise and style variations and has no way of handling rotated characters.

Fig.2 (a) Character extraction in form of Matrix

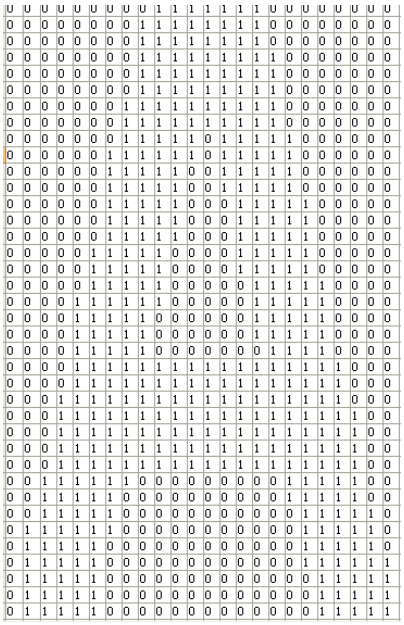

Fig.2 (b) Character extraction in form of Matrix

2.6. Post Processing

It encompasses grouping, error detection and correction techniques. The result of plain symbol recognition on a document, is a set of individual symbols. However, these symbols in themselves do usually not contain enough information. Instead we would like to associate the individual symbols that belong to the same string with each other, making up words and numbers. The process of performing this association of symbols into strings, is commonly referred to as grouping. The grouping of the symbols into strings is based on the symbols location in the document. Symbols that are found to be sufficiently close are grouped together Up until the grouping each character has been treated separately, and the context in which each character appears has usually not been exploited. However, in

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

advanced optical text recognition problems, a system consisting only of single-character recognition will not be sufficient. Even the best recognition systems will not give 100% percent correct identification of all characters, but some of these errors may be detected or even corrected by the use of context.

3. WHY MATLAB?

MATLAB stands for Matrix Laboratory. Here you play around with matrices. Hence, an image (or any other data like sound, etc.) can be converted to a matrix and then various operations can be performed on it to get the desired results and values. Image processing is quite a vast field to deal with. We can identify colors, intensity, edges, texture or pattern in an image. In this tutorial, we would be restricting ourselves to detecting colors (using RGB values) only. Using MATLAB you can solve technical computing problems faster than with traditional programming language, such as C, C++, JAVA, FORTRAN. There is a wide range of applications, including signal and image processing, image accusation, Neural Network, etc.

4. OCR PERFORMANCE EVALUATION

No standardized test sets exist for character recognition, and as the performance of an OCR system is highly dependent on the quality of the input, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often given, and usually presented as the percentage of characters correctly classified. However, this does not say anything about the errors committed. Therefore in evaluation of OCR system, three different performance rates are investigated:

• ***Recognition rate.***

It is the proportion of correctly classified characters.

• ***Rejection rate***.

It is the proportion of characters which the system was unable to recognize. Rejected characters can be flagged by the OCR-system, and are therefore easily retraceable for manual correction.

• ***Error rate.***

The proportion of characters erroneously classified. Misclassified characters go by undetected by the system, and manual inspection of the recognized text is necessary to detect and correct these errors. There is usually a tradeoff between the different recognition rates. A low error rate may lead to a higher rejection rate and a lower recognition rate. Because of the time required to detect and correct OCR errors, the error rate is the most important when evaluating whether an OCR system is cost-effective or not. The rejection rate is less critical. An example from barcode reading may illustrate this. Here a rejection while reading a bar-coded price tag will only lead to rescanning of the code or manual entry, while a miss decoded price tag might result in the customer being charged for the wrong amount. In the barcode industry the error rates are therefore as low as one in a million labels, while a rejection rate of one in a hundred is acceptable. In view of this, it is apparent that it is not sufficient to look solely on the recognition rates of a system. A correct recognition rate of 99%, might imply an error rate of 1%. In the case of text recognition on a printed page, which on average contains about 2000 characters, an error rate of 1% means 20 undetected errors per page. In postal applications for mail sorting, where an address contains about 50 characters, an error rate of 1% implies an error on every other piece of mail.

5. RESULTS

To illustrate the accuracy of proposed English handwritten and sample text images OCR algorithm by using MATLAB, performance was measured using the samples. Figure 3 and 4 shows the sample document scanned from HP desk jet scanner at 300 dpi. The images were then filtered, binarized, clipped and resized. Lines of text were then extracted from the images. The font size was identified; segmentation was performed on each line to segment characters taking in consideration the characteristics of English Verdana fonts templates. MATLAB (R2012.a/64-bit) is used to implement the proposed OCR algorithm. The recognition accuracy was 85% to 90% due to improper hand written characters. The templates of all Characters and numbers are of 24X42 pixels.

Fig.3. Handwritten Sample and its output

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 10, October 2018

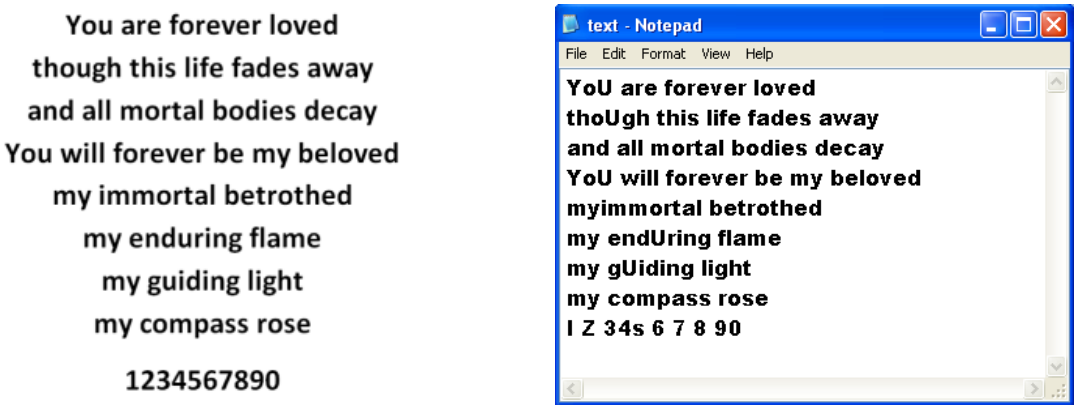

Fig.4. Text Image Sample and its output

6. FUTURE SCOPE

New methods for character recognition are still expected to appear, as the computer technology develops and decreasing computational restrictions open up for new approaches. There might for instance be a potential in performing character recognition directly on grey level images. However, the greatest potential seems to lie within the exploitation of existing methods, by mixing methodologies and making more use of context. Integration of segmentation and contextual analysis can improve recognition of joined and split characters. Also, higher level contextual analysis which looks at the semantics of entire sentences may be useful. Generally there is a potential in using context to a larger extent than what is done today. In addition, combinations of multiple independent feature sets and classifiers, where the weakness of one method is compensated by the strength of another, may improve the recognition of individual characters. The frontiers of research within character recognition have now moved towards the recognition of cursive script that is handwritten connected or calligraphic characters. Promising techniques within this area, deal with the recognition of entire words instead of individual characters.

7. CONCLUSIONS

Today optical character recognition is most successful for constrained material that is documents produced under some control. However, in the future it seems that the need for constrained OCR will be decreasing. The reason for this is that control of the production process usually means that the document is produced from material already stored on a computer. Hence, if a computer readable version is already available, this means that data may be exchanged electronically or printed in a more computer readable form, for instance barcodes. The applications for future OCR-systems lie in the recognition of documents where control over the production process is impossible. This may be material where the recipient is cut off from an electronic version and has no control of the production process or older material which at production time could not be generated electronically. This means that future OCR-systems intended for reading printed text must be Omni font. Another important area for OCR is the recognition of manually produced documents. Within postal applications for instance, OCR must focus on reading of addresses on mail produced by people without access to computer technology. Already, it is not unusual for companies etc., with access to computer technology to mark mail with barcodes. The relative importance of handwritten text recognition is therefore expected to increase.

REFERENCES

[1]. H.S. Baird & R. Fossey. A 100-Font Classifier. Proceedings ICDAR-91, Vol. 1, p. 332-340, 1991.
[2]. R. Bradford & T. Nartker. Error Correlation in Contemporary OCR Systems. Proceedings ICDAR-91, Vol. 2, p. 516-524, 1991.
[3]. J-P. Caillot. Review of OCR Techniques. NR-note, BILD/08/087.
[4]. R. G. Casey & K. Y. Wong. Document-Analysis Systems and Techniques. Image Analysis Applications, eds: R. Kasturi& M. Tivedi, p. 1-36.
[5]. Product help: http://www.mathworks.com/pl_homepage