Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 2, Number 2; January-March, 2015, pp. 136-139

http://www.krishisanskriti.org/jbaer.html

A Review on Optical Character Recognition Using Various Techniques

Dipti Singh[1], Atul Bansal[2] and Neha Bansal[3]

[1]M.Tech Scholar Dept. of ECE GLAU Mathura
[2]Dept. of ECE GLAU Mathura
[3]Dept. of ECE GLAU Mathura
E-mail: [1]dipti0409@gmail.com

Abstract—*Computer vision, artificial intelligence and pattern recognition are vital areas of research in the field of electronics and image processing. Optical character recognition (OCR) is one of the main aspects of computer vision and has evolved to a great extent since its commencement. Optical Character Recognition (OCR) has gained so much significance among the researchers now a day as it is a prominent sector for a Human Computer Interaction (HCI) System. OCR is a process in which readable characters are recognized from optical data obtained digitally. In character recognition techniques symbolic identity relate with image of character. In a very first stage of typical OCR systems, optical scanner digitized the input image. After that location and segmentation is performed on each character, and the ensuing character image which contains some noise is put into a pre-processor for noise reduction and normalization. For classify characters in classification stage certain features are extracted from the character. After classification, the recognized characters are grouped to renovate the original symbol strings, and context may then be applied to detect and correct errors. Using different approaches various methodologies and algorithms have been developed for this reason. In this paper various techniques have been reviewed.*

Keywords: *OCR, Computer vision, Artificial intelligence, ANN, MSER.*

1. INTRODUCTION

OCR is a technique that automatically extracts text from paper documents. Computer vision, pattern matching and artificial intelligence all these fields involve OCR. In OCR the printed and handwritten characters or text are detected according to their shapes [4].

There are types of character recognition one is offline and other is online character recognition. Offline character recognition scanned the input image and stored it in digital form. The variation in the human handwriting is the difficult problem during the recognition in OCR [3].

Another area where OCR can be used is Data Entry. This part covers technologies for entering huge amounts of limited data. They are designed to read data like, article numbers, amounts of money, customer's identification, account numbers etc [1]. Text entry is another field provided by OCR [1]. They are page readers mostly used in office automation for text entry. In a word processing environment the reading machines are used to enter large amounts of text. Other applications include Aid for blind, automatic number plate readers, Form readers, Signature verification and identification etc [1].

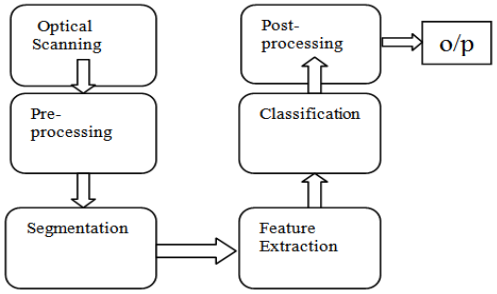

Fig. 1: General block diagram of OCR.

1.1 Optical scanning

A digital image of the original document is captured through the first stage scanning. In OCR optical scanners are used, which consist of a transport mechanism and a sensing device to convert light intensity into gray-levels. Input image generally contains black print on a white background. Therefore, when performing OCR, its common to convert the multilevel image into a bilevel image of black and white which is known as thresholding [2]. The thresholding process is totally dependent of the quality of the bilevel image which is important as the result of the following recognition. By putting fixed threshold, the gray-levels below this threshold are black and levels above this threshold are white. A document with uniform background having a high contrast, a pre-chosen fixed threshold can be sufficient. However, a lot of documents encountered in practice comprise a quite large range in contrast. In these cases to obtain a good result more sophisticated methods are required for thresholding [2].

1.2 Pre-processing

The scanned image contains a certain amount of noise. The characters may be smeared or broken which is depend on the resolution on the scanner and the success of the applied technique for thresholding. In preprocessor some of these defects, which may later cause poor recognition rates, can be eliminated to smooth the digitized characters [2].

The smoothing implies both filling as well as thinning. Filling removes small gaps, breaks, and holes in the digitized characters, whereas thinning may help to reduce the width of the line. The most commonly used technique for smoothing is to moves a window across the binary image of the character, applying certain rules to the contents of the window [2].

In addition toward smoothing, preprocessing usually involves normalization. To obtain characters of uniform size, slant and rotation generally normalization is applied. The angle of rotation must be found to correct for rotation. For detecting skew hough transform are commonly used. However, it is not possible to find the rotation angle of a single symbol until after the symbol has been recognized [2].

1.3 Segmentation

In segmentation the constituents of an image are determines. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition [2].

Applied to text, 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.

1.4 Feature extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes [2].

1.5 Classification

The classification is the process of identifying each character and assigning to it the correct character class. In the following sections two different approaches for classification in character recognition are discussed. First decision-theoretic recognition is treated. These methods are used when the description of the character can be numerically represented in a feature vector [3].

We may also have pattern characteristics derived from the physical structure of the character which are not as easily quantified. In these cases the relationship between the characteristics may be of importance when deciding on class membership. For instance, if we know that a character consists of one vertical and one horizontal stroke, it may be either an “L” or a “T”, and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed [1].

1.6 Post processing

Grouping

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 [1].

Error-detection and correction

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 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 [1].

2. OPTICAL CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK

M. Usman Raza et al. [4] proposed that classification was based on the letter width basis, because different characters have varying widths i.e. 5 pixels, 10 pixels and 15 pixels. Letters of all the font sizes were sampled up/down to a fixed height of 10 pixels, and to a relative width. For this purpose, famous image processing algorithm 'Nearest Neighbor Algorithm' was used to produce patterns according to fixed height, but relative width. The resultant patterns were thickened, if necessary, to fit in the specific category. The thickening process involves the filling up of the pattern with the light colored pixels to have the same size as the category has. For example, if there is a category of 10-pixels width, and

some pattern is 7 pixels wide, of course, it does not fit in the 5- pixels width category; so this pattern is thickened to 10-pixels width. At the end, three categories of patterns were obtained: 10x5 pixels, 10x10 pixels, and 10x15 pixels where height is fixed to 10, but had different widths of 5, 10 and 15. as it was shown in Fig. 2.

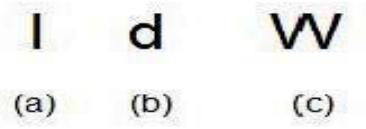

Fig. 2: Classification on letter-width basis

Different topologies of backpropagation techniques could be used but for this study the three layer topology was used, in which there is one input layer, one hidden layer and one output layer. The number of neurons in the input layer was 50, 100 or 150 depending upon the category; and this number was set dynamically on the run time. Output layer had a fixed number of 62 neurons as there were 62 patterns in all. Though, there are various optimization techniques available that can be used to optimize hidden layer number of neurons; but here they tested two methodologies; one is to take the square root of number of input neurons; and the other to fix the number of neurons to 50 in the hidden layer. To train the network for patterns, epochs (number of iterations, network modified its weights) were set to 100,000; because this number showed the optimized results and was got by trial and error process. It has been concluded that classification of patterns effects a lot on the OCR systems positively. More the classification, more accurate results can be produced. The value of epoch and learning rates are inversely related to each other.

Nikola Dojcinovic et al. [5] provide an outline of OCR system based on neural network. Novel approach to character extraction trough MSER feature extractor is presented, making process of character extraction invariant to affine transformation and quick illumination change. Non-character objects were removed, in order to reduce problem of character recognition to classification. Neural network was used for character recognition. Neural network was trained with backpropagation algorithm for character recognition.

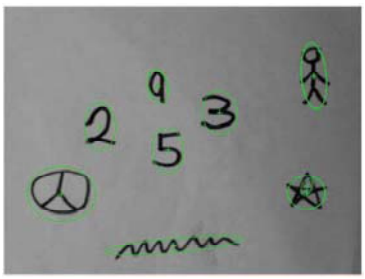

Fig. 3(a)

Fig. 3. (a) is scanned documents with detected MSER features circumscribed and (b) is extracted digits after affine rectification and outlier removal.

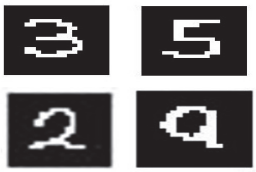

Fig. 3(b)

System testing was performed on 20 document images similar to one shown in Fig. 3, each containing numeral characters and other non-character objects. Major influence on performance of the system has character segmentation. All characters correctly segmented and correctly characterized were recognized with ultimate accuracy. Test showed that overall system accuracy does not differ significantly under change of space angle and distance, therefore proving the invariance of the system to affine transformation of document scans. Accuracy over 97.5% was obtained on set of 20 scanned documents, each with 4 characters and 4 non character objects.

M. M. Farhad et al. [6] proposed a new method of optical character recognition and its validity was checked for different seeking angles. Despite the shapes of the characters the curvature properties of the characters has been used here and the recognition rate was analyzed. A smaller seeking interval makes the input larger and makes the recognition process accurate but at the same time increase the calculations needed. Thus reduces the speed of the algorithm. This paper shows better result than other conventional methods in case of skewness and noise in input image.

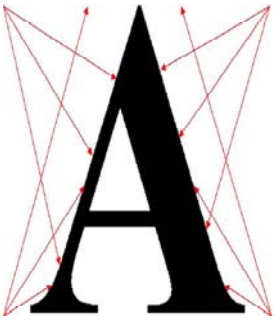

Fig. 4: Extraction of sought point of Character 'A'.

Each character has a specific shape and each character has a different distance of the text pixels (black pixels) from the four corners of the resized image. In order to determine a number of features that is actually suitable to recognize any characters and make it differentiable from others the features are

determined by varying the angles of seeking. In this algorithm they has used the seeking angles 15°, 30°, 45°, 60°, 75° (difference is 15°) and for the angles 10°, 20°, 30° ,40° ,50° ,60° ,70° ,80° (difference is 10°). If there is no text pixels found at any seeking angle, then the point of image border at that seeking angle is taken as sought point. Then the angular deviation between these featured points is determined which are the feature for the recognition process. Fig. 4 shows the process of finding the sought points.

Sushruth Shastry et al. [7] introduced this algorithm, was the simple fact that English alphabets are fixed glyphs and they shall not be changed ever. Due to this fact, usage of artificial neural networks and vector based data training provide almost accurate results, but these are performing a lot of redundant work. And also, most of the OCR processes today involves images from high resolution scanners and cameras. OCR technologies now can make use of this advancement in technology and consider techniques which were abandoned (the latest OCR innovation without shape training dates back to the year 2000 [8]) due to the lack of present day imaging technology. This algorithm has the advantage of speed, power, memory and area, since it does not include any training or learning mechanisms and also because of lack of image database which some OCR techniques require [9]. Also, this algorithm is the first, in being a multiple font OCR and also a no training type OCR.

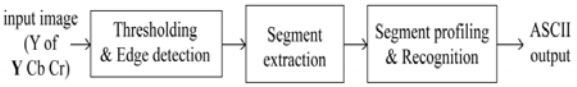

Fig. 5: Main block diagram of proposed algorithm

The algorithm is represented in a block diagram as shown in Figure 5. First, an image is acquired through any of the standard image acquisition techniques. The input image is assumed to be in the *Y Cb Cr* color format. The algorithm works on the *Y* part of the input image, which using an appropriate thresholding algorithm, the gray level image is then thresholded to obtain a binary image which quantizes alphabets and background to black and white colors respectively. The obtained binary image is then passed on to a specific edge detection process. The edge detection algorithm is performed such that only the right sided edges of each alphabet are obtained and the other edges are eliminated. After edge detection, the image is then segmented and feature extraction is performed. The next step is to profile stored line segments. Profiling of segments is the process of categorizing them into different types of segments such as short, long, line or curve, etc. When the complete alphabet is processed, the feature vector which contains all high bits represents the recognized alphabet which that feature vector belongs to. Finally, ASCII equivalent of the recognized alphabet is the output.

3. CONCLUSION

In this paper, we have reviewed several methods which adopt neural network approach for Optical character recognition. It has been concluded that classification of patterns effects a lot on the OCR systems positively. More the classification, more accurate results can be produced.

There is a lot of work that still needs to be done. Many other OCR techniques and algorithm will be included in this research and extensive tests need to be done with larger number of images.

REFERENCES

[1] Line Eikvil, “Optical character recognition”, December 1993.

[2] Sandeep Tiwari, Shivangi Mishra, Priyank Bhatiya, “OCR using matlab”, IJARECE, 2013.

[3] Nisha Sharma, Tushar Patnaik, Bhupendra Kumar, “Recognition for Handwritten English Letters: A Review”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[4] M Usman Raza, Ata Ullah, Khawaja MoyeezUllah Ghori, Sajjad Haider, “Text extraction using artificial neural network”.

[5] Nikola Dojcinovic, Igor Mihajlovic, Jugoslav Jokovic, Vera Markovic and Bratislav Milovanovic, “ Neural network based optical character recognition”, 11th symposium on neural network applications in electrical engineering. NEUREL-2012

[6] M. M. Farhad, S M Nafiul Hossain, Ahmed Shehab Khan, Atiqul Islam, “An Efficient Optical Character Recognition Algorithm using Artificial Neural Network by Curvature Properties of Characters”, 3rd International Conference on informatics, electronics & vision 2014.

[7] Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, “” i ” - A novel algorithm for Optical Character Recognition (OCR)”. IEEE, 2013.

[8] Tin Kam Ho and Nagy, G., “OCR with no shape training”, in Proc. Of Pattern Recognition 15th International Conference, Barcelona, 2000, pp. 27 - 30 vol.4.

[9] Mori, S., et al., “Historical review of OCR research and development”, in Proc. of the IEEE, 1992, pp.1029–1058.