Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

Analysis and Implementation Optical Character Recognition using Template Matching

Asst. Prof. Ravi J Khimani[1], Asst. Prof. Nishant S Sanghani[2], Asst. Prof. Pooja P Vasani[3]

[1]*Computer Science & Engineering, SLTIET, Rajkot*
[2]*Computer Science & Engineering, SLTIET, Rajkot*
[3]*Computer Engineering, AITS, Rajkot*

Abstract — Optical Character Recognition is one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. It deals with recognizing optically processed characters. OCR by using Template Matching is a system prototype that useful to recognize the character or alphabet by comparing two images of the alphabet. This system prototype has its own scopes which are using Template Matching as the algorithm that applied to recognize the characters, characters to be tested are alphabet (A – Z), grey-scale images were used with Times New Roman font type and recognizing the alphabet by comparing between two images. The processes are starting from the acquisition process, filtering process, threshold the image, clustering the image of alphabet and lastly recognize the alphabet.

Keywords - OCR, Character Recognition, Filtering Process, Acquisition Process, Clustering image, Alphabet

I. INTRODUCTION

Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents [2].

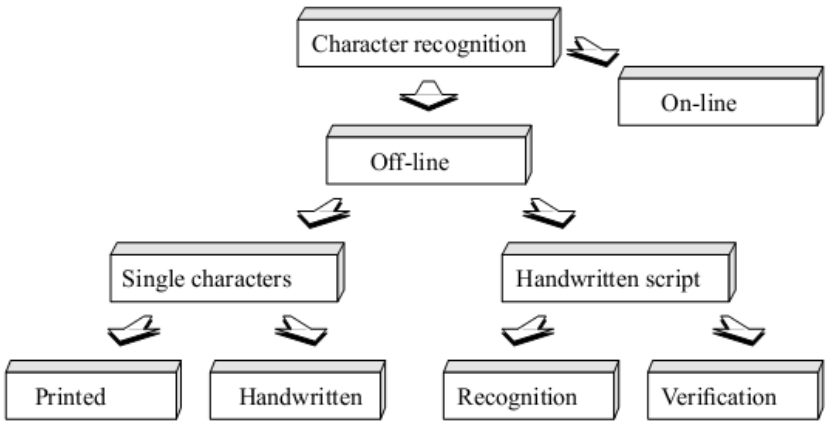

Figure 1. Types of character recognition

The more constrained the input is, the better will the performance of the OCR system be. However, when it comes to totally unconstrained handwriting, OCR machines are still a long way from reading as well as humans. However, the computer reads fast and technical advances are continually bringing the technology closer to its ideal [1].

II. APPLICATION OF OCR

Three main application areas are commonly distinguished; data entry, text entry and process automation.

2.1. Data entry

This area covers technologies for entering large amounts of restricted data. The systems are characterized by reading only an extremely limited set of printed characters, usually numerals and a few special symbols. They are designed to read data like account numbers, customer's identification, article numbers, amounts of money etc. The paper

formats are constrained with a limited number of fixed lines to read per document. These systems are specially designed for their applications and prices are therefore high [8]

2.2. Text entry

The second branch of reading machines is that of page readers for text entry, mainly used in office automation. Here the restrictions on paper format and character set are exchanged for constraints concerning font and printing quality. The reading machines are used to enter large amounts of text, often in a word processing environment. However, under controlled conditions the single character error and reject rates are about 0.01% and 0.1% respectively. The reading speed is typically in the order of a few hundred characters per second [5].

2.3. Process Automation

Within this area of application the main concern is not to read what is printed, but rather to control some particular process. This is actually the technology of automatic address reading for mail sorting. Hence, the goal is to direct each letter into the appropriate bin regardless of whether each character was correctly recognized or not. The general approach is to read all the information available and use the postcode as a redundancy check [4].

III. COMPONENTS OF OPTICAL CHARACTER RECOGNITION

A typical OCR system consists of several components. In figure 1.2 a common setup is illustrated. The first step in the process is to digitize the Analog document using an optical scanner. When the regions containing text are located, each symbol is extracted through a segmentation process. The extracted symbols may then be pre-processed, eliminating noise, to facilitate the extraction of features in the next step.

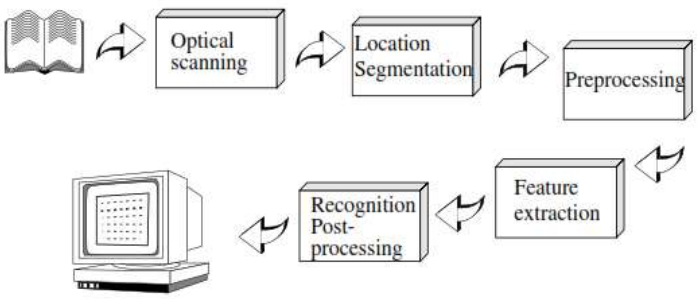

Figure 2. Components of OCR

The identity of each symbol is found by comparing the extracted features with descriptions of the symbol classes obtained through a previous learning phase. Finally contextual information is used to reconstruct the words and numbers of the original text. In the next sections these steps and some of the methods involved are described in more detail [3].

3.1. Optical scanning

Through the scanning process a digital image of the original document is captured. optical scanners generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing OCR, it is common practice to convert the multilevel image into a bi-level image of black and white. Often this process, known as Thresholding, is performed on the scanner to save memory space and computational effort. Thresholding has best methods which can vary threshold over local properties of documents like contrast and brightness. However, such methods usually depend upon a multilevel scanning of the document which requires more memory and computational capacity. Therefore such techniques are seldom used in connection with OCR systems, although they result in better images [6][7].

3.2. Location and segmentation

Segmentation is a process that determines the constituents of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. It is applied to text, segmentation is the isolation of characters or words from the images, logos, etc. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts. it has four groups like

extraction from touching and fragmented characters, distinguishing noise from text, mistaking graphics for text and mistaking text for graphics [6][7].

3.3. Pre-processing

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for Thresholding, the characters may be smeared or broken, can be eliminated by using a pre-processor to smooth the digitized characters. Two methods, Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. In addition to smoothing, pre-processing usually includes normalization. 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. [1]

3.4. 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. The techniques for extraction of such features are often divided into three main groups, where the features are found from: The distribution of points, Transformations and series expansions, Structural analysis [9].

3.5. 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. Another is pattern characteristics derived from the physical structure of the character which are not as easily quantised. In these cases the relationship between the characteristics may be of importance when deciding on class membership, known as Structural Method [10].

3.6. Post Processing

Post Processing includes Grouping and Error Detection and Correction. In first, The result of plain symbol recognition on a document is a set of individual symbols. 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. And in second, 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. There are two main approaches, where the first utilizes the possibility of sequences of characters appearing together. This may be done by the use of rules defining the syntax of the word, by saying for instance that after a period there should usually be a capital letter [11].

IV. IMPLEMENTATION OF OCR WITH TEMPLATE MATCHING

In template matching algorithm, we first define templates of alphabets and numbers in our context. Then we take image to read data. In which, first detect the strings line by line and then using the matching metric and recalculation of image size, we found match from the collection of alphabets, numbers and symbols. If match is not found, we simply repeat step until match is found up to the maximum metric.

4.1. Implementation steps

The template-matching algorithm implements the following steps [7]:

- Firstly, the character image from the detected string is selected.
- After that, the image to the size of the first template is rescaled.
- After rescale the image to the size of the first template (original) image, the matching metric is computed.
- Then the highest match found is stored. If the image is not match repeat again the third step.
- The index of the best match is stored as the recognized character.

4.2. Flowchart for template matching

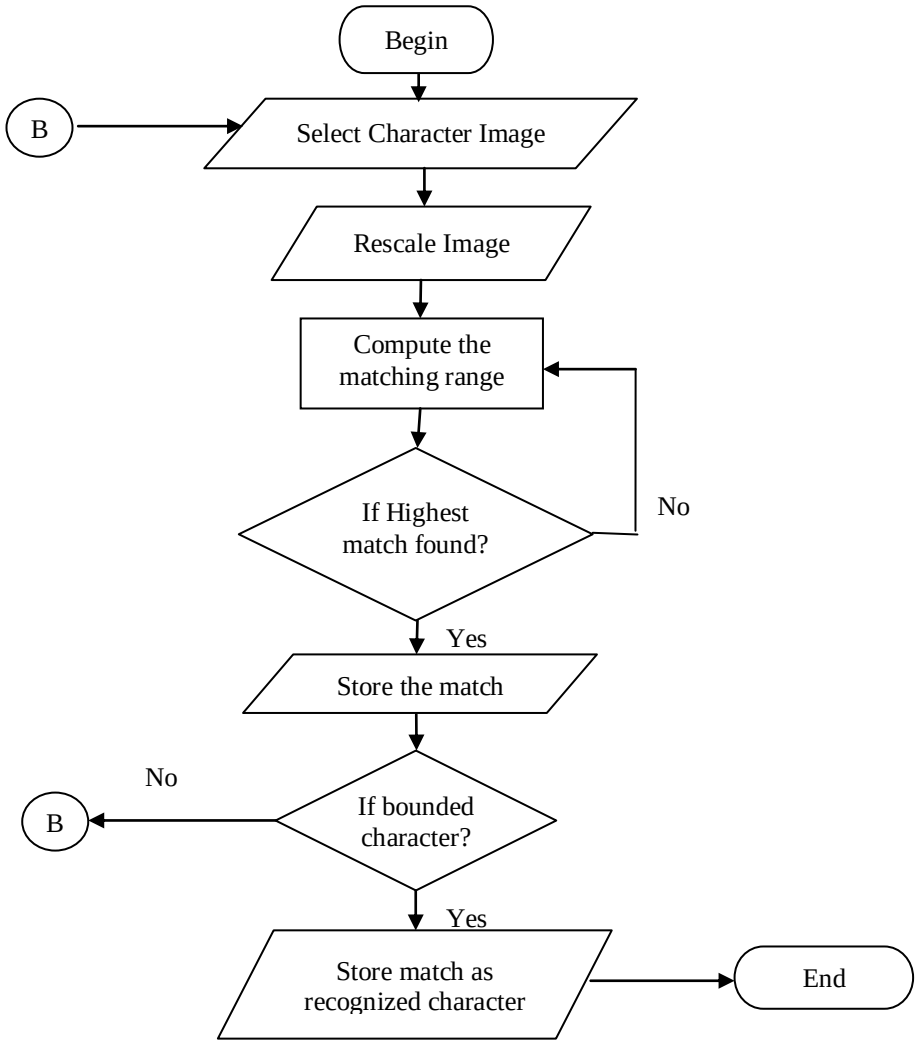

Figure 3. Flowchart for implementation of Template Matching for OCR

V. RESULTS

5.1. Experiment one

In first experiment, a gray scale image is given, all the text are clearly visible, so each character of whole string is recognized properly and generates output without any error.

Figure 4. Experiment one

5.2. Experiment two

In second experiment, the gray scale image given with large font size, OCR recognizes all the character and digits but the additional spaces are padded in between in resultant text.

Figure 5. Experiment two

VI. CONCLUSION

As an overall view of the system prototype, it could be conclude that this system prototype has been developed by using the technique that has mentioned and elaborated which is the Template Matching approach to recognize the character image. Besides, the interface of the system prototype looks user-friendly and makes the user of this system prototype easier to use it. As a result, the recognition process of this system become smoothly because of the steps that used in this system while recognizing the character. Even though this system prototype could gives several advantages to the users, but this system prototype are still facing a number of limitations.

REFERENCES

[1] Line Eikvil, "Optical Character Recognition", 1993.
[2] Shunji Mori, Ching Y. Suen, Kazuhiko Yamamoto, "Historical Review of OCR Research and Development, Proceeding of IEEE, Vol. 80(7), pp. 1029-1058, 1992.
[3] Quin Chen, "Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises", University of Ottawa, 2003.
[4] Thomas Natschlager, "Optical Character Recognition", Institute of Theoretical Computer Science.
[5] Aparna Vara Lakshmi Vemuri, T.V.Sai Krishna, Atul Negi, "Dataset Generation for OCR" [OCR_Datasheet_Generation.pdf].
[6] Kartar Singh Siddharth, Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Science and Engineering, Vol. 3(6), pp. 2332-2345, 2011.
[7] Nadira Muda, Nik Kamariah Nik Ismail, Siti Azami Abu Bakar, Jasni Mohamad Zain "Optical Character Recognition By Using Template Matching (Alphabet)".
[8] Jesse Hansen, "A Matlab Project in Optical Character Recognition (OCR)".
[9] Oivind Due Trier, Anil K. Jain, Troffin Taxt, "Feature Extraction Methods for Character Recognition – A Survey", Pattern Recognition, Vol. 29(4), pp. 641-662, 1996.
[10] Jairo Rocha, Theo Pavlidis, "Character Recognition Without Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17(9), pp. 903-909, 1995.
[11] Kirill Safronov, Dr.-Ing. Igor Tchouchenkov, Dr.-Ing Heniz Worn, "Optical Character Recognition Using Optimization Algorithm", Workshop on Computer Science and Information Technology, pp. 1-5, 2007.