Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-5, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Optical Character Recognition System for English Script in offline Processing

Lovely[1], Satbir Singh[2], Saurabh Mahajan[3]
[1]M.Tech. Research Scholar [2]Assistant Professor, Dept. of ECE, GNDU, RC, Gurdaspur, Pb.
[3]Assistant Professor, Dept. of ECE, BECT, Gurdaspur, Pb., India

Abstract*: In this paper we have explored an off line system for the recognition of the text characters from an input image. Recognition is the psychological feature of a computer to receive and execute intelligible offline input image from different sources such as paper documents, photographs, touch screens, and other devices. A complete Optical character recognition system also handles formatting, performs correct segmentation into characters and determines the most plausible words. Character recognition can be online or offline. Online character recognition has real time related information while offline character recognition system control in not a real time.The image of the text may be sensed "off-line" from a piece of paper by optical scanning. Character recognition is a process which contacts a symbolic meaning with objects (letter, symbols & numbers) drawn on an image. In a very first stage of typical OCR systems, optical scanner digitized the input image. After that position 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 restore the original symbol strings, and context may then be applied to detect and correct errors. Optical character recognition is one of the most successful applications of automatic pattern recognition.*

Key Words: *Optical character recognition, image processing toolbox, feature extraction, MATLAB.*

Introduction

In Today's world character recognition plays important role to make everything automotive. Optical Character Recognition is the process of converting offline text document or printed documented into device readable form. Now a day's fast growing technology, digitization of the papers or documents are significant for future use which gives scope for the researches to perform Optical Character Recognition [1]. Optical character recognition is a technique, which is used to convert offline scanned image, into editable text format. It is an ability of detecting, segmenting and identifying characters from image[6]. OCR has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Optical character recognition is divided into: online and offline recognition. The difference originates from the type of input data that is available for recognition

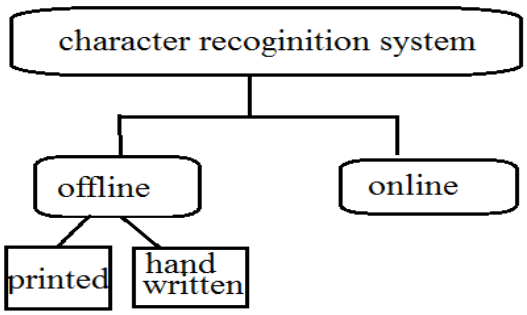


Figure 1: Types of character recognition system.

In online character recognition, characters are recognized at real time as soon as it is written. Online systems perform better than offline recognition as they have timing information and since they avoid the initial search step of locating the character [7]. Online systems obtain the position of the pen as a function of time directly from the interface. This is usually done through pen-based interfaces where the writer writes with a special pen on an electronic tablet.

In offline character recognition can be further classified to printed characters and handwritten character recognition. In offline character recognition, the typewritten, handwritten character is typically scanned in form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm [8]. Offline character recognition become more challenging due to shape of characters, great variation of character symbol and document quality. Therefore offline character recognition is considered as a more challenging task then its online counterpart.

OCR SYSTEM

OCR system consists of following phases of recognition process such as Optical scanning of Input Image, preprocessing, Segmentation, Feature Extraction, Classification and Recognition.

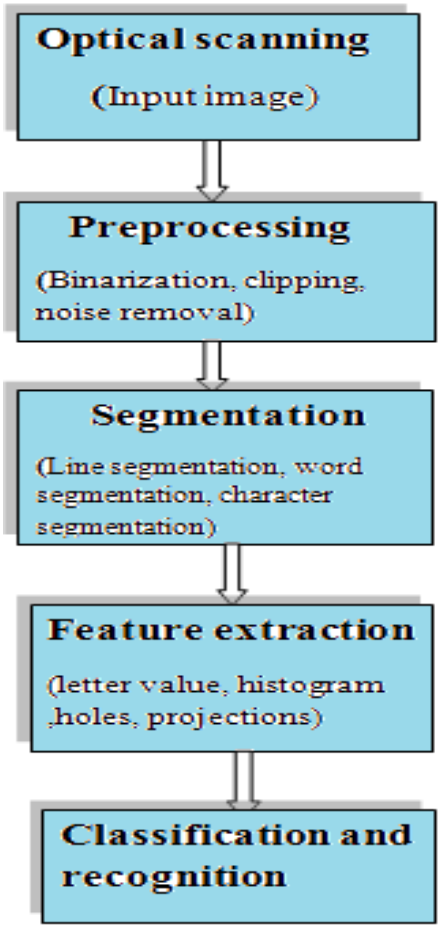


Figure 2: Block diagram of OCR

The output of first phase becomes the input of next phase the assignment of preprocessing relates to the elimination of noise and discrepancy in offline input text image or document. Segmentation is partition of the individual characters of an input offline image. Feature extraction is used to extract the most relevant information which is used to classify the objects in the offline image [9]. The classification is the process of identifying each character of input offline image and assigning to it the correct character class. Several areas where OCR used including mail sorting, bank processing, document reading and postal address recognition require offline text recognition systems, pattern recognition

2.1 Optical scanning

A digital image of the original document is captured through the opening 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, it is 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[5]. 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. Two categories of thresholding exist: Global and Adaptive. Global thresholding picks one threshold value for the entire document image, often based on an estimation of the background level from the intensity histogram of the image. Adaptive thresholding is a method used for images in which different regions of the image may require different threshold values.

2.2 Pre-processing

A sequence of operations is to be performed in pre-processing stages. The main objective of the preprocessing is to organize the information so that the subsequent character recognition task becomes simpler. The preprocessing phase aims to extract the relevant textual parts and prepares them for segmentation and recognition. The main objectives of preprocessing are i) noise reduction, (ii) normalization of data and (iii) compression in the amount of information to be retained. In pre-processing scanned document is converted to binary image and various other techniques to remove noise, to make it ready and appropriate for feature extraction are applied [8]. These techniques include segmentation to isolated individual characters, skeletonization, contour making, normalization, filtration etc. It essentially enhances the image rendering it suitable for segmentation. The various techniques performed on the image during pre-processing stage are which includes thresholding, binarizing, filtering, edge detection, gap filling, and segmentation.

2.3 Segmentation

Segmentation is an important stage, because the extent one can reach in separation of lines, words, characters directly affects the recognition rate of the script. It is done by partition from the individual

characters of an image. Segmentation of offline text document characters into different zones (upper, middle and lower zone) and characters is more difficult than that of printed documents that are in standard form. This is mainly because of variability in paragraph, words of line and characters of a word, skew, slant, size and curved. Sometimes components of two adjacent characters may be touched or overlapped and this situation creates difficulties in the segmentation task [9]. Touching or overlapping problem occurs frequently because of modified characters in upper-zone and lower-zone. Segmentation is an important stage. The process of segmentation includes isolating line, word, Individual characters from input image.

Line Segmentation → Word Segmentation → Character Recognition

Line segmentation- In a printed script, the text lines are almost of same height, provided that the script is written in a specific font size. If the script is composed by a type-machine, surely the font size will be uniform everywhere. The input image to separate the all text lines, line segmentation is used.
Word segmentation- Word segmentation is providing the gap between words of a selected line.
Character segmentation- It offers the spacing between the characters of each word, called Character Recognition [4]. After the line segmentation, consider each and every line which is segmented before going through the process of character segmentation. Each line is segmented in its individual characters (isolated) for further operation [5].

2.4. Feature Extraction

This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images. It is the important step in character recognition, however other steps also need to be optimized because these steps are closely related to each other as outputs of earlier step is inputted to later step. Feature extraction is an essential part of pattern recognition systems with a direct effect on the result. In offline input image recognition, feature extraction involves image analysis and processing. One of the most commonly used filters for image processing are Gaussian filters and derivatives of Gaussians. Gaussian filter is frequently used as a low-pass filter for noise suppression and Gaussian derivatives are used to detect and localize edges along with determining their orientation. Feature extraction involves representing offline input image text by a set of discriminant features [2]. The feature extraction step selects and prepares data which is used by a classifier to achieve the recognition task. The feature representation is based on extraction of certain types of information from the image. Features can be broadly classified into two categories: structural features and statistical features. Structural features are involved of structural elements like loop, line, crossing point, curve, end point and stroke etc. Statistical features are computed by some statistical operations on image pattern and these include features like zoning, projection, profiling, histogram and distance etc.

2.5 Classification

The classification is the process of identifying each character and assigning to it the right character class. In this section two different approaches for classification in character recognition are discussed. First decision-theoretic recognition is treated. These methods are used when the picture of the character can be numerically represented in a feature vector [3]. We may also have pattern characteristics consequent 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.

3. Results and Discussions

In this section all the above said process is discuss with the help of results in the following steps.

Step1: Original Input image

The Input Image which we have taken to perform the different processing of the OCR system as shown above. The original input image is having three Lines or three words.

Step2: Pre-Processing

In pre-processing scanned offline input image or document is converted to binary image and various other techniques to remove noise

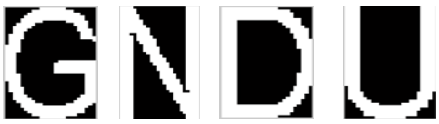
.

Step3: Segmentation

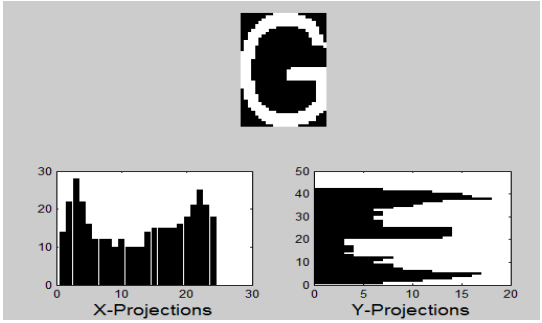
Line or Word Segmentation

In the original input image only three words are there which are representing the three lines and three words as shown above.

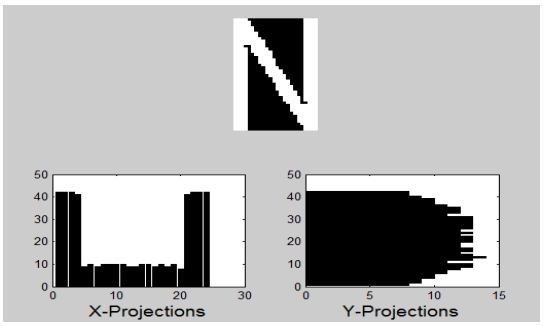
Characters Segmentation (with X&Y Projections)



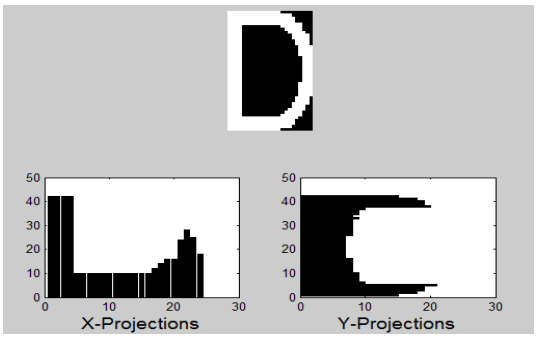
Each character from the word GNDU is segmented for further operations shown above.



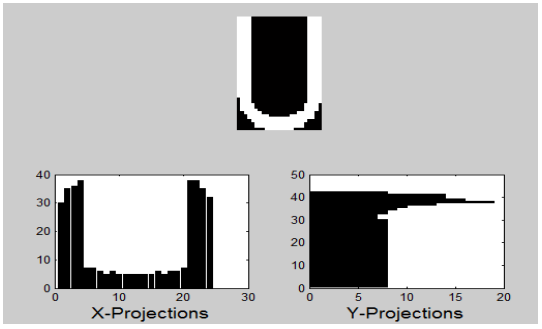
(a)



(b)



(c)



(d)

We have also represented the histogram X&Y projections of the word GNDU in the below graphs as a, b, c, d respectively for each character.

4. Conclusion

Optical character recognition involves pre-processing, segmentation, feature extraction, classification and recognition of Script. This paper is regime about OCR system for offline character recognition. The systems have the ability to yield excellent results. The first step is image acquisition which acquires the scanned image followed by noise filtering, smoothing and normalization of scanned image, rendering image suitable for segmentation where image is decomposed into sub images. Feature Extraction improves recognition rate and misclassification. The feature extraction step of optical character recognition is the most important. It

can be used with existing OCR methods; hence we develop a very good character recognition system which must achieve highest accuracy. Feature extraction plays a significant role in achieving high recognition accuracy. For feature extraction mostly use Gabor filter, but here rather than using Gabor filter for whole image we firstly divide image into dual parts then segment it. Also, not all the features of an image are useful for classification and therefore feature extraction helps in yielding only the significant features for feeding into a classifier.The results obtained by using our formulation establish by experiments are very encouraging and auspicious; therefore, this kind of formulation can improve the performance of the system.Finally brief survey of classification is studied for our future work.

5. References

[1] Ashwin S Ramteke, Milind E Rane, *"Offline Hand-written Devanagari Script Segmentation,"* Interna-tional Journal Of Scientific & Technology Research Vol. 1, Issue 4, MAY 2012.

[2] Line Eikvil, "Optical character recognition", December 1993.

[3] N. Arica and F. Yarman-Vural, "An Overview of Character Recognition Focused on Off-lineHandwriting", IEEE Transactions on Systems, Man,and Cybernetics, Part C: Applications and Reviews,Vol.31 no.2, pp. 216 - 233. 2001

[4] Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat,"Optical Character Recognition Implementation Using Pattern Matching "(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, pp. 2088-2090

[5] Gur, Eran, and ZeevZelavsky, "Retrieval of Rashi Semi-Cursive Handwriting via Fuzzy Logic." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[6] Veena Bansal and R. M. K. Sinha, "Integrating Knowledge Sources in Devanagari Text Recognition System," IEEE Transaction on system,man,and cy-bernetics,System and Humans, Vol. 30, No.4, July 2000.

[7] U. Bhattacharya and B. B. Chaudhuri, "*Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals,"* IEEETrans. Pattern Anal. Mach. Intell., Vol. 31, no.3, pp. 444–457, Mar. 2009.

[8] Pranob K Charles,V.Harish, M.Swathi, CH. Deepthi, "A Review on the Various Techniques used for Optical Character Recognition", *International Journal of Engineering Research and Applications (IJERA),* Vol. 2,Issue 1,Jan-Feb 2012.

[9] Sushruth Shastry, Gunasheela G, Thejus Dutt, Vinay D S and Sudhir Rao Rupanagudi, i - A novel algorithm for Optical Character Recognition (OCR), 978-1-4673-50907/13 IEEE, 2013.