

REVIEW ARTICLE



ISSN: 2321-7758

REVIEW ON VARIOUS CHARACTER RECOGNITION TECHNIQUES

SWATI MALIK¹, PREETI ANEJA²

¹Student, Department of CSE, Hindu College of Engineering, India

²Assist Prof, Department of CSE, Hindu College of Engineering, India

Article Received:07/05/2015

Article Revised on:15/05/2015

Article Accepted on:20/05/2015



ABSTRACT

Character Recognition (CR) has been studied from the past several decades, and is still a demanding research topic in the field of pattern recognition and image processing. Character recognition has long been a critical area of the Artificial Intelligence. Recognition is a trivial task for humans, but to make a computer program that does character recognition is extremely difficult. Recognizing patterns is just one of those things humans do well and computers don't. The reasons for this are the many sources of variability, abstraction and absence of hard-and-fast rules that define the appearance of a visual character. Hence rules need to be heuristically deduced from samples. The various character recognition techniques are optical character recognition, feature extraction, template matching & neural networks etc. work has been done on all these techniques but still improvement is required. This paper presents detailed review of existing works in character recognition based on analysis done in past few years.

Key Words-- Feature extraction, Classification, OCR, Character Recognition, Clustering, pattern Matching, Segmentation

©KY Publications

I INTRODUCTION

Machine simulation of human functions has been an extremely challenging research field since the advent of digital computers in the field of image processing and pattern recognition. The intensive research in the field of character recognition (CR) is not only because it helps in postal address, bank check amount reading but also because it automates processing of bulk amount of papers, transfers data into machines and web interface to paper document [6].

Character recognition is the process to classify the input character according to the predefined

character class. With the increasing interest of computer applications, modern society needs that the computer should read the text. The text may be in the form of scanned handwritten document or typed text in various fonts or a combination of both. The character recognition system helps in making the communication between a human and a computer easy [3]. It is still challenging and interesting task to design a system which gives high recognition accuracy, without considering quality of the input document and different character font style variation [4].

Pattern recognition has three main steps: observation, pattern segmentation, and pattern classification. Optical Character Recognition (OCR) systems is transforming large amount of documents, either printed alphabet or handwritten into machine encoded text without any transformation, noise, resolution variations and other factors.

In general, handwriting recognition is classified into two types as off-line and On-line character recognition. Off-line handwriting recognition involves automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. Off-line handwriting recognition is more difficult, as different people have different handwriting styles. But, in the on-line system, On-line character recognition deals with a data stream which comes from a transducer while the user is writing. The typical hardware to collect data is a digitizing tablet which is electromagnetic or pressure sensitive. When the user writes on the tablet, the successive movements of the pen are transformed to a series of electronic signal which is memorized and analyzed by the computer [1].

Optical Character Recognition (OCR) is a field of research in pattern recognition, artificial intelligence and machine vision, signal processing. Optical character recognition (OCR) is usually referred to as an off-line character recognition process to mean that the system scans and recognizes static images of the characters. It refers to the mechanical or electronic translation of images of handwritten character or printed text into machine code without any variation [1].

OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. The input of one step is the output of next step. The task of preprocessing relates to the removal of noise and variation in handwritten. Several area where OCR used including mail sorting, bank processing, document reading and postal address recognition require offline handwriting recognition systems, pattern recognition.

Digitization:- It is the process of converting a paper-based handwritten document into electronic format. Here, each document consists of only one character. The electronic conversion is accomplished by using a method where by a document is scanned and an

electronic representation of the original document as a image file format is produced. [1].

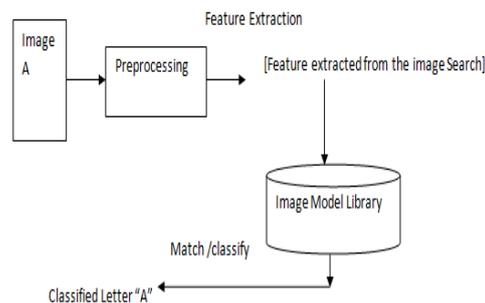


Fig 1 Phases of General Character Recognition System

Pre-processing In This phase, there is a series of operations performed on the scanned input image. It enhances the image rendering it suitable for segmentation the gray-level character image is normalized into a window sized. After noise reduction, a bitmap image was produced. Then, the bitmap image was transformed into a thinned image [1].

Segmentation It is one the most important process that decides the success of character recognition technique [5]. It is done by separation from the individual characters of an image. Segmentation of handwritten 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. Touching or overlapping problem occurs frequently because of modified characters in upper-zone and lower-zone. Segmentation is an important stage [1]. Segmented image is shown in fig 2.

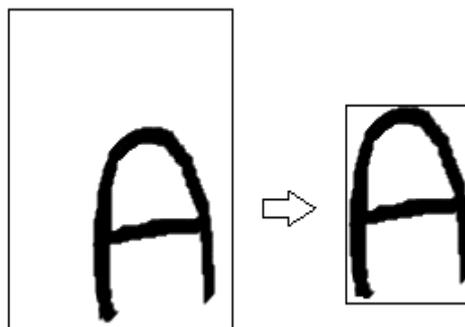


Fig. 2 Segmented Image [8].

Feature Extraction In this phase, features of individual characters are extracted. The performance of the character recognition system depends on the features that are extracted. The extracted features from input character should allow classification of a character in a unique way. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images [7]. In this diagonal features were used for intersection and open end points features, transition features, zoning features, directional features, parabola curve fitting-based features, and power curve fitting-based features in order to find the feature set for a given character [1].

Post processing Post-processing stage is the last stage of the proposed recognition system [6]. It prints the corresponding recognized characters in the structured text form.

II HISTORY OF CHARACTER RECOGNITION

The very first effort in the direction of CR was made by Turing who attempted to develop an aid for the visually handicapped.

The first character recognizer appeared in around 1940s. The early works were concentrated either upon machine-printed text or upon a small set of well-separated hand-written text or symbols. Machine-printed CR generally used template matching and for handwritten text, low-level image processing techniques were used on the binary image to extract feature vectors, which were then fed to statistical classifiers. A good survey of the CR techniques used until 1980s can be found in. The period from 1980 - 1990 witnessed a growth in CR system development due to rapid growth in information technology [6]

After 1990, image processing techniques and pattern recognition were combined using artificial intelligence. Along with powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, there came in efficient, modern use of methodologies such as neural networks (NNs), hidden Markov models (HMMs), fuzzy set reasoning, and natural language processing. The 1990's systems for the machine-printed off-line [6] and restricted vocabulary size, user-dependent on-line handwritten characters were satisfactory only for restricted applications.

III CHARACTER RECOGNITION TECHNIQUES

Some approaches take a holistic approach, recognizing entire words, while others focus more on recognizing individual characters. Holistic approaches incur more computational cost since there are more models, but have more expressive and discriminative power since the visual cues are gathered over large areas. Fig. 3 shows the classification of character recognition techniques. Basically the character recognition can be done using online and offline methods [3].

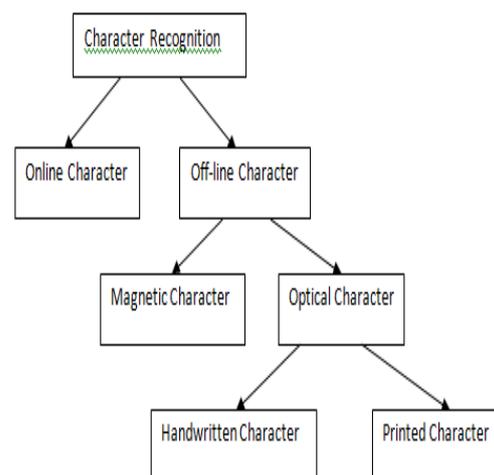


Fig. 3 Classification of Character Recognition Techniques

On-line recognition Online handwriting recognition has gained interest due to increase in usage of hand held devices. Nonparametric methods have recognition time proportionate to the training set size. These methods use all points per stroke for calculating the similarity measurement. The incorporation of keyboard being difficult in the hand held devices demands for alternatives, and in this respect, online method of giving input with stylus is gaining quite popularity. On-line recognition seems to be a simpler problem since more information is available. A few studies on converted or independent on-line and off-line data suggest superior recognition performance for on-line data [3]. The distinction between on-line and off-line recognition is not as rigid as it may seem. The challenges posed by the online character recognition system are to increase the recognition accuracy and to reduce the recognition time. Recently some methods have been proposed that extract temporal

information from static off-line data. This would allow for on-line treatment of off-line data. Also on-line systems have been designed using spatial representations of the on-line data, or combining on-line and off-line representations by supplementing each on-line data point with a pixel image of its local surroundings [3].

Off-line recognition Off-line recognition operates on pictures generated by an optical scanner. The data is two-dimensional and space-ordered which means that overlapping characters cannot be separated easily. Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand "printed" text.

Template Matching Optical Character Recognition by using Template Matching is a system prototype that useful to recognize the character or alphabet by comparing two images of the alphabet. Template matching is the process of finding the location of a sub image called a template inside an image. Once a number of corresponding templates is found their centers are used as corresponding points to determine the registration parameters. Template matching involves determining similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure [9].

Neural network An Artificial Neural Network as the backend is used for performing classification and Recognition tasks. In offline character recognition systems, the Neural Network has emerged as the fast and reliable tools for classification towards achieving high recognition. Neural network architectures can be classified into two major sets specifically; feed-forward and feedback (recurrent) networks and the majority common ANN used in the CR systems are the multilayer perceptron of the feed forward networks and the Kohonens Self Organizing Map (SOM) of the feedback networks, use Feed Forward Neural Network. In a feed forward neural network, nodes are organized into layers; each "stacked" on one another. The neural network

consists of an input layer of nodes, one or more hidden layers, and an output layer. Each node in the layer has one corresponding node in the next layer, thus creating the stacking effect. Back propagation is a learning rule for the training of multi-layer feed-forward neural network. Back propagation derives its name from the technique of propagating the error in the network backward from the output layer. To train a Back propagation neural network, it must be exposed to a training data set and the answers or correct interpretations of the set [9].

Feature Extraction The idea of the feature point extraction algorithm is to identify characters based on features that are somewhat similar to the features humans use to identify characters [3]. Programmers must manually determine the properties they feel are important. Some example properties might be Aspect Ratio, Percent of pixels above horizontal half point, Percent of pixels to right of vertical half point, Number of strokes, Average distance from image centre, Is reflected y axis, Is reflected x axis. Researchers have used many methods of feature extraction for handwritten characters [3]. Shadow code, fractal code, profiles, moment, template, structural (points, primitives), wavelet, directional feature etc., have been addressed in the literature as features.

IV APPLICATIONS OF CHARACTER RECOGNITION

Data entry.

This area covers technologies for entering large amounts of restricted data. Initially such document reading machines were used for banking applications. 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.

Text entry.

The second branch of reading machines is that of page readers for text entry, mainly used in office automation [2]. 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.

These page readers are in strong competition with direct key-input and electronic exchange of data. This area of application is therefore of diminishing importance.

In spite of the great number of algorithms that have been developed for character recognition, the problem is not yet solved satisfactory, especially not in the cases when there are no strict limitations on the handwriting or quality of print. Up to now, no recognition algorithm may compete with man in quality. However, as the OCR machine is able to read much faster, it is still attractive. In the future the area of recognition of constrained print is expected to decrease. Emphasis will then be on the recognition of unconstrained writing, like omni font and handwriting [2].

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 [2].

V CONCLUSION & FUTURE SCOPE

The character recognition methods have developed remarkably in the last few years. A variety of techniques have emerged, influenced by developments in related fields such as image recognition and face recognition. Neural Network helps the system to recognize the character even if the exact pattern is not available in the database. The template matching has high speed, but is not very effective when there are font discrepancy, font slant, font defilement, stroke connection and stroke breaking due to the environment or the instrument itself. From various studies we have seen that selection of relevant feature extraction and classification technique plays an important role in performance of character recognition rate. The accurate recognition is directly depending on the nature of the material to be read and by its quality. Each technique have own porn's and con's. So, in future there is lots of work to be done to remove drawbacks should give us good accuracy and increase performance.

REFERENCES

- [1] Er. Neetu Bhatia "Optical Character Recognition Techniques: A Review" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014, pp 1219-1223.
- [2] Line Eikvil" Optical Character Recognition" December 1993 pp 1-35.
- [3] Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal "Survey of Methods for Character Recognition", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012. pp 180-189.
- [4] Rohini B. Kharate, Dr.S.M.Jagade, Sushilkumar N. Holambe "A Brief Review and Survey of Segmentation for Character Recognition" International Journal of Engineering Sciences January 2013, pp 14-17.
- [5] Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 – 8887) Volume 12– No.2, November 2010 pp 8-15.
- [6] Richa Goswami and O.P. Sharma "A Review on Character Recognition Techniques" International Journal of Computer Applications (0975 – 8887) Volume 83 – No 7, December 2013 pp19-23.
- [7] Sukhpreet Singh "Optical Character Recognition Techniques: A Survey" Journal of Emerging Trends in Computing and Information Sciences Vol. 4, No. 6 June 2013 pp 545-550.
- [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, pp.659-662.
- [9] Vijay Laxmi Sahu, Babita Kubde "Offline Handwritten Character Recognition Techniques using Neural Network: A Review" International Journal of Science and Research (IJSR), India, pp 87-94.