International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

REVIEW ARTICLE

ISSN: 2321-7758

REVIEW ON: ENGLISH SCANNED DOCUMENTS

PARDEEP KAUR, POOJA CHOUDHARY

Article Received: 07/03/2015

Article Revised on15 /03/2015

Article Accepted on:18/03/2015

MECHANICAL
ENGINEERING
www.ijoer.in
ENGINEERS
MAKE A WORLD OF DIFFERENCE
International Journal of Engineering Research-Online

ABSTRACT

The character recognition is very important research topic in the field of pattern recognition and image processing. There is a large demand for optical character recognition on written document and scanned document. Handwritten character recognition is a complex problem due to the huge disparity of writing styles, dissimilar size of the characters. Many types of handwriting styles from diverse persons are considered in this effort. An image with elevated resolution will certainly take a lot longer time to calculate than a lower resolution illustration. In the realistic image acquisition systems and environment, shape distortion is common processes because different people's handwriting has dissimilar shape of characters. In this paper we present a review on various preprocessing and feature extraction parameters and classification techniques.

Keywords— Image Processing, Character Recognition, Neural Network.

INTRODUCTION

The character recognition is developing in near around 1940. The character recognition is very active research in the field of image processing. The character recognition are used the problems related that image processing, pattern recognition, cognitive science, and artificial intelligence. It is widely used to computer to receive and interpret the input source such as photographs, paper document, touch screen and other devices. There to improve the text accuracy and efficiency many methodologies are used for character recognition. There are different writing styles for different people like size, shape orientation, thickness format, and pen pressure. Basically character recognition are two types online character recognition and offline character recognition. There are mainly six different stages to solve the OCR problem. The character recognition is ability of a convert books and documents in electronic files to computerize a record keeping system in an office or to publish the text on website. It is basically used to improve the interface between man and machine printed text accuracy and efficiency in a lot of application.

Historical evolution of OCR

In 1935 Tauschek was also granted a US patent on his method. The Tauschek machine was a mechanical device that used templates and photo detectors. In near about 1940 the first

Character recognitions techniques developed. The RCA engineers in 1949 worked on the first primitive computer type OCR to blind people. This device can be converted printed character to machine language and machine language then spoke the lattes. But this technique was very expensive and was not used after testing. A good survey of OCR techniques used until 1980 can be found.

Methodology of OCR:

The machine printed OCR are used for handwritten text, low level image processing and template matching techniques are used on the binary images to extract feature vectors, which ware then fed to statistical classifier. In 1929 Gustav Tauschek obtains a patent on OCR in Germany.

There are different types of methodologies are used for OCR system. The OCR approaches are depending upon the OCR system and the methodology used. The literature review in the field of OCR is hierarchical process. It performs step by step each task. Like firstly perform preprocessing followed by segmentation, representation then training and recognitions and post processing. In several methods some time the OCR methods are combined or omitted and the feedback mechanisms is used to update the output of each stage.

Data Acquisition

The data acquisition is initial stage of OCR. In image acquisition a scanned image of a character data collected as an input image. The collected images has a specific format according to the user requirement such as jpeg, btm etc. This image is acquired by using electronic devices like digital camera, scanner or any other suitable digital devices.

Preprocessing

There are various steps are perform on image in preprocessing stage to extract noise and distortion free image. The mainly in preprocessing multilevel colored image is converted into bit level binary images and then to remove the noise of binary image. The various techniques are performing in image preprocessing are discussed below.

i. Noise Reduction: - The noise reduction is a process of noise removing in an image. The noises degrade the image quality. The noise can occur at different stages like capturing, during transmission, and compression. There are different types of filters and morphological operations are used for removing image noise. The median filter is basically used for removing the image noise.

ii. Filtering: - The filtering techniques are used for modifying or enhancing the image quality. The filters are mainly used to suppress either the high frequencies in the image for example smoothing the image, or the low frequencies for example enhancing or detecting edges in the image. The image can be filtered either in frequency or spatial domain.

iii. Morphological operation:- The morphological operation is used to connect unconnected pixels, broken strokes, decompose the join strokes, remove the isolated pixels and also provided smoothing pixel boundary. It works on geometrical structure and digital images. That can be originally developed for binary images. The morphological operation can be successfully to take away the noise from document images due to low quality of document and ink.

iv. Binarization:- The process of Binarization to convert the gray scale image to binary images is called Binarization or thersholding.There are two approaches for conversion of gray level images to binary images like global threshold and local or adaptive thresholding. Global threshold select single threshold value based on background level from intensity histogram of the image. Local or adaptive threshold is used different values for each pixel according to the local information. The main purpose of binariation is to identify the extent of object and also concentrate on shape analysis by using skealtionation.

v. Normalization:- This is most important part of pre processing stage. This is not affecting the identity of the word. In handwritten image from a scanned images include some steps , which are mainly begin with cleaning of image, line detection ,slant and slope removal, skew correction ,and character size normalizations. Generally the normalization provides a tremendous reduction in data size, thinning extraction the shape information of the character.

vi. Compression:-The compression is a space domain technique. In compression process two main techniques are used thresholding and thinning thresholding techniques. This technique is used to reduce the storage requirement and increases the speed of processing. Thinning is used to extract the shape or size information of the character.

vii. Segmentation: - The goal of segmentation is to simplify and change the representation of an image into more meaningful and easier to analyze. The segmentation process use to decompose large size image into sub images. It is basically used to line and curves etc in image. The segmentation is the process of assigning a label to every pixel in an image and increases the accuracy of an image. The segmentation is used to set of contours extracted from the image like edge detection. The thresholding is simplest method of segmentation. This method is used on a clip level to turn a gray scale image into binary image. There is also called balanced histogram thresholding.

Feature Extraction

The feature extraction method is analyzes a scanned document images and select a set of features that can be used for uniquely classifying the character. Features extraction is extract for each class that helps to differentiate it from other class and each stage of character is represented as feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of feature which increases the accuracy and recognition rate. According to the nature of scanned document with it high degree of variability and imprecision obtaining these features is a difficult task. Basically feature extraction methods are based on three types of feature.

i. Statistical Representation: - The Statistical representation of scanned document image by statistical distribution of points takes cause of style variations to some extent. This type of statistical representation does not allow the reconstruction of original image. It is providing high speed and low complexity. The statistical representation is used to reducing the dimension of the feature set.

ii. Zoning: - The frame contain the character is divided into m*n zones. For each zone feature are extracted to frame the feature vector. The densities of the point or some feature in different regions are analyzed local characteristics instead of global characteristics.

iii. Crossing and Distances: - The statistical feature is the number of crossing of contour by a line segment in a specified direction. The crossing count the number of transitions from background to foreground image pixels along horizontal and vertical line though the document image and distances calculated the distances of the first document image pixel detected from the upper and lower boundaries of the image along vertical lines and from the left and right boundaries along horizontal line. The features imply that a horizontal threshold is established above, beneath and through the center of normalized script. The calculate value of the feature is decided by the number of times document images crosses the threshold. The ascending and descending portions of the script are used.

iv. Projection: - The projection histogram counts the number of pixels in each column and row of a character image. The projection histogram can be separate character such as M and N. This representation creates a 1-D signal from a 2-D signal image. This can used to represent the character image

v. Geometrical and topological Representation:- The structural feature are based on topological and geometrical properties of the document, such as cross points, loops, aspect ratio, branch points, strokes and their direction, inflection between two points, horizontal curves at top or bottom. There are various global and local property of character can be represented by structure features with high tolerance top distortions and style variation. This type of representation may also encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object.

vi. Global Transformation and series Expansions:- The linear combination coefficients provides a compact encoding know as transformation and serious expansion. A continuous signal generally contains more information they need to be represented for the purpose of classifications .This is may be true for discrete continuous signal. The signal is represented by linear combination of a sequence of simpler well defined function. The common transform and series Expansion method used in the CR field includes the following.

vii. Fourier Transforms:-The one of the most attractive properties of Fourier transformation is the capability to identify the position shifted document. As soon as observe the magnitude spectrum and ignores the phase Fourier transforms have been useful to in many ways.

viii. Gabor Transform:-This is a windowed Fourier transformation distinction. In this case, the window used is defined by a Gaussian function rather than a window of discrete size.

ix. Wavelets: - wavelet transformation techniques allow one to represents the signal at different level of resolution. The segments of documents image which may correspond to letters or words are represented by wavelets coefficients. These coefficients are then fed to a classifier for recognition.

x. Moments: - Moments such as central moments, from a compact represented of the original image that make the process of reorganization. The moments are observe object scale, translation, and rotation invariant. Then the original image can be completely reconstructed from the moment coefficients.

Classification

The classification stage is the decision making stage of recognition system. The quality of feature is depending upon the performance of the classifier. There are two types of classifier are used for recognition soft computing technique and classical technique.

i. **Template matching:** The template matching is simplest techniques of character recognitions are based prototypes matching against the character or word to be recognition. The template matching techniques determines the degree of similarity between two vectors like group of pixels, curvature, shapes etc in the feature space. There are three types of matching techniques are direct matching, Deformable templates, and elastic matching and Relaxation matching.

ii. **Statistical Techniques:** The statistical techniques are based on statistical decision functions and set of optimality criteria, which maximize the probability of the observed pattern given the model of a certain class. The major statistical methods applied in the OCR field are Hidden Markov Modeling (HMM), Fuzzy set Reasoning, Quadratic classifier, Nearest Neighbor (NN).

iii. **Structural Techniques:** This is recursive description of a complex pattern in term of simpler patterns based on the shape of the object. The structural techniques are used to describe and classify the character in the CR system.

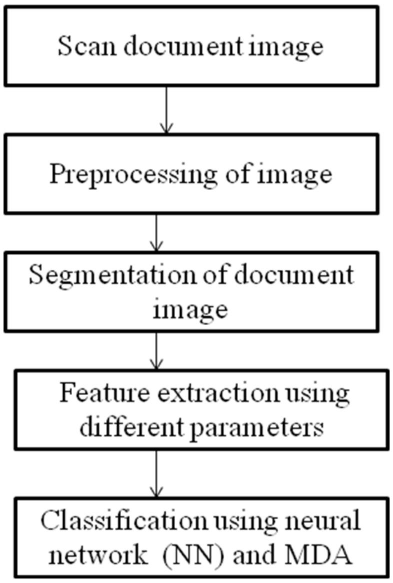
iv. **Neural network** (NN): The neural network is a computing architecture that consists of massively parallel interconnection of adaptive neural processors. Because of its parallel nature, it can perform computation at a higher rate compared to the classical techniques. Because of its adaptive nature it can adopt to change in the data and learn the characteristics of input signal. The neural network architecture can be classified as feed forward and feedback neural network. In OCR the mostly used multilayer preceptron of the feed forward network and the self organizing map (SOM) feedback network.

v. **Support vector machine classifier**: the support vector machine is a state of the art classification method introduced in 1992 by Boser, Guyon, and vapnik. The SVM classifier is widely used in bioinformatics due to its highly accurate, able to calculate and process the high dimensional data such as gene expression and edibility in modeling diverse source of data.SVM belong to the general category of kernel methods. it is primarily a two class classifier. These patterns called support vector finally define the classification function. Their number is minimized the margin. The support vector replaces the prototypes with the main difference between SVM and tradition template matching techniques.

Proposed Methodology

In the character recognition system first we take a scanned image as input . the apply preprocessing steps on that scanned image. After the preprocessing the next step is the segmentation of the image. The next step to segmentation is feature extraction; the output of the segmentation is the input of the feature extraction. Then different classifier is used for classification.

Md. Alamgir Badsha, Md. Akkas Ali, Dr. Kaushik Deb, Md. Nuruzzaman Bhuiyan [1] introduces“Handwritten bangle character Recognition using neural network” This paper introduce an off-line recognition system for Bangla handwritten characters using Back- propagation Feed-forward neural network. Character is recognized by analyzing its shape and evaluates its features that differentiate each character. Priya Sharma, Randhir Singh [2] present “survey and classification of character recognition system” this paper present a survey, and classification of different character recognition methods. Reetika Verma1, Mrs. Rupinder Kaur [3] present “Enhanced Character Recognition Using Surf Feature and Neural Network Technique” This paper, the planned result focus on applying Neural Network Algorithm model for character recognition. Gaurav Kumar, Pardeep Kumar Bhatia [4] introduce “Neural Network based Approach for Recognition of Text Images”


Scan document image
Preprocessing of image
Segmentation of document image
Feature extraction using different parameters
Classification using neural network (NN) and MDA

Literature Survey

In this paper present the preprocessing is done through Otsu’s method and feature extracted Through Fourier descriptor and classifier is NN is used. Shalin A. Chopra1, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi4, Prof. Gandhali S. Gurjar [5] presents “Optical Character Recognition” in this paper ,competent and less expensive approach to construct OCR for reading any manuscript that has fix font size and handwritten style.

Er.Puneet kaur and Er.Balwinder Singh [6] presents “recognition of signboard images of Gurumukhi” In this paper proposed recognition of inaccessible handwritten numerals of Gurumukhi script. Ashok Kumar, Pardeep Kumar Bhatia [7] introduce “Offline Handwritten Character Recognition Using Improved Back- Propagation Algorithm”

This paper gives idea on the improved neural network technique to recognize the offline handwritten characters. Rajiv Kumar Nath, Mayuri Rastogi [8] “Improving Various Off-line Techniques used for Handwritten Character Recognition: a Review” In this paper describe various preprocessing and feature extraction parameters and different classification techniques.

Conclusion

Recognition move toward heavily depends on the nature of the facts to be recognized. In this paper, character recognition systems for handwritten English script are converse in detail. Many segmentation methods and different Classifiers with different features are also discussed. We consider that our survey will be supportive for researchers in this field There are a lot of factors that influence the performance of OCR system .we use the neural network and MDA classifiers for classification. Other kinds of preprocessing and feature extraction model may be tested for a better recognition rate in the future research in OCR System.

REFERENCES

[1]. Md. Alamgir Badsha, Md. Akkas Ali, Dr. Kaushik Deb, Md. Nuruzzaman Bhuiyan "Handwritten bangle character Recognition using neural network" © 2012, IJARCSSE.

[2]. Priya Sharma, Randhir Singh "survey and classification of character Recognition System" International Journal of Engineering Trends and Technology Volume 4 Issue 3- 2013.

[3]. Reetika Verma1, Mrs. Rupinder Kaur"Enhanced Character Recognition Using Surf Feature and Neural Network Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5565-5570.

[4]. Gaurav Kumar, Pardeep Kumar Bhatia "Neural Network based Approach For Recognition of Text Images" International Journal of Computer Applications (0975 – 8887) Volume 62– No.14, January 2013.

[5]. Shalin A. Chopra1, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar "Optical Character Recognition" International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014.

[6]. Er.Puneet kaur and Er.Balwinder Singh "recognition of signboard images of Gurumukhi" Journal of Global Research in Computer Science Volume 3, No. 6, June 2012.

[7]. Ashok Kumar, Pardeep Kumar Bhatia "Offline Handwritten Character Recognition Using Improved Back- Propagation Algorithm" International Conference On Emerging Trends in Engineering and Management, ICETEM 2013

[8]. Rajiv Kumar Nath, Mayuri Rastogi "Improving Various Off-line Techniques used For Handwritten Character Recognition: a Review" International Journal of Computer Applications (0975 – 8887 Volume 49– No.18, July 2012 .

[9]. Anita Pal, Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network" ,International Journal of Computer Science & Communication Vol. 1, No. 2, pp. 141-144, July-December 2010.

[10]. Rashad Al-Jawfi,"Handwriting Arabic Character Recognition LeNet Using Neural Network", The International Arab Journal ofInformation Technology, Vol. 6, No. 3, July 2009.

[11]. Nafiz Arica and Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-Line Handwriting", in IEEETrans. Sys.Man. Cybernetics,Vol. 31, No. 2, pp.216-238, 2001.

[12]. Srinivasa Kumar DeviReddy, Settipalli Appa Rai, "Hand written character Recognition using back propagation network", Journal of Theoretical and Applied Information Technology, 2005-2009.

[13]. R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey,"IEEE. Transactions on Pattern Analysis andMachine Intelligence, vol. 22, no. 1, pp. 63-84, 2000.

[14]. U. Bhattacharya, and B. B. Chaudhuri, "Handwritten numeral databases of Indianscripts and multistage recognition of mixed numerals," IEEE Transaction on Pattern analysis and machine intelligence, vol.31, No.3, pp.444-457, 2009.

[15]. Anil.K.Jain and Torfinn Taxt, "Feature extraction methods for character recognition-A Survey," Pattern Recognition, vol. 29, no. 4, pp. 641-662, 1996.

[16]. N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", 9th International Conference on Information Technology (ICIT'06), ICIT, pp. 133-136.

[17]. F. Bortolozzi, A. S. Brito, Luiz S. Oliveira and M. Morita, "Recent Advances inHandwritten Recognition", Document Analysis, Umapada Pal, Swapan K. Parui, Bidyut B. Chaudhuri, pp 1-30.