RESEARCH ARTICLE OPEN ACCESS

English Scanned Document Character Recognition Using NN and MDA

Ms. Pardeep Kaur [1], Ms. Pooja Choudhary [2]

ABSTRACT

In this paper use neural network for English scanned document character recognition to increases the performance or accuracy of character. Most of the traditional system is not extensible enough. In neural network is very good ability to recognize more character sets than initially defined. Neural network method makes hundred percent perfect systems to identify and verification the characters as compared to conventional scanning process in devise. The character recognitions technique is one of the most widely used for authentication of person as well as documents.

Keywords: **-** English Character recognition, pre-processing, segmentation, NN, feature extraction.

I. INTRODUCTION

Optical character recognitions has attracted research in recent times and received extensive attention in academic and production fields. The optical character recognitions are important area in image processing and pattern recognition. In India there are multi languages and multi scripts are used, the eighteen officials scripts and accepted and have hundred regional languages. Today many researchers have been done to scanned English documents for character recognition for using various methods. The OCR is used to developing algorithms for reading text on the image taken by camera in reading registration plates, reading scanned books and scanned documents etc.

These algorithms for machine vision and artificial intelligence for example neural network vectors machine fuzzy classifiers etc. The OCR is mostly used machine encoding text and that text can be easily edited searched and can be processed in many other ways according to requirements. It is also used small size for storage in comparison to scanned documents. Computerized processing to recognize individual character is required to convert scanned document into machine language encoded form. There are two types of character recognition online and offline optical character recognition. The online optical character recognition is real time optical recognition of characters. The online system obtains the position of pen as a function of time directly interfaces. Online scanned character recognition the documents are capture and store in digital form via different means. The special pen is used for conjunction with an electronic surface. It is generally accepted that the online method of recognition handwritten text has achieved better result than its offline counterpart. Offline character recognition the typewritten and 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. The offline character recognition is a more difficult and challenging task as there is no control over the medium and instruments used.

II. PROBLEM FORMULATION

Optical character recognition deal with the problem of recognition optically processed characters. Optical recognition is performed offline after the writing or printing has been completed as opposed to on line recognition where the computer recognizes the character as they are drawn. The performance of character recognition is depends upon the quality of scanned documents. The pre-processing steps are used to removing low frequency background noise, normalization the intensity of individual scanned documents. Several filter are use for reduces certain image details enable an easier or faster evaluation. The propose solution focus on applying Multilinear Discriminant analysis algorithm and neural network model for character recognition. The character to be recognized is in the form of vector elements. The main elements are in the form of 0 and 1 or -0 and -1.

For existing algorithm there are several factors are depends of character recognitions including:-

- Need for better result
- Need for Increase performance and accuracy for OCR
- Existing OCR is less accurate and need for more enhancements

In documents scanning steps a scanner is used to scan the documents. The quality of scanned documents depends upon the scanner the scanner is high speed and color quality is proper the accuracy of the recognitions speed is very high. In the character recognition process include several complex algorithms are used. Some algorithms are previously loaded templates and dictionary which are cross checked with the character in the documents and corresponding machine editable ASCII character. The verifying of scanned documents is done either randomly or chronologically by human intervention.

III. METHODOLOGY

The optical recognition using neural network is basically in the field of research. To gain better knowledge, techniques and solution regarding the procedures we studied the various re-search papers on previously OCR system. All these study helped us with classifying our target goals. The character recognition system involves many steps to completely recognize and help to produce machine encoded text. The computer recognizes the character in the documents through a revolutionizing technique called character recognition. There are various steps are used for character recognition as: scanned document Image, image acquisition , Image cropping steps, scanned image pre-processing , segmentation, Feature extraction steps and NN classification .

Image Acquisition

In Image acquisition, the optical character recognition system acquires a scanned document image as an input image. This image is acquired with the help of scanner, digital camera or any other suitable digital input device.

Pre-processing

The pre-processing of image means applying a number of procedures for image to improve the accuracy of OCR like thresholding, filtering, resizing, Thinning, smoothing, and Skeletonization etc. So that successive algorithm to final classification can be made simple and more accurate

RGB to Gray conversion

The scan image is true color (RGB) and this has to be converted into a binary image based on a threshold value of documents. In this conversion firstly checking the dimension of test image by size ()if the image is RGB image, then it is converted into gray image by rgb2gray ()

```
If size (testImg, 3) ==3
 TestImg=rgb2gray (testImg)
```

To get a binary image, this RGB format image has to be converted gray-scale format, and then by using the threshold value found by Ostu's method to used gray-scale image is converted to binary image. The value of pixel lies "between" 0 to 1 or "between" 0 to 255 depending upon its class.

Noise Reduction

The noise introduced by the optical scanning device or the writing methods or writing instruments, causes discount line segments, gapes in between lines, filled loops and bumps etc.The distortion including local variation rounding of corners, erosion, dilations also problem for documents. The median filter is a process that replaces the value of a pixel by the median of gray levels in the neighborhood of that pixel.

Thinning

The thinning is a morphological operation process that is used to remove selected foreground pixels from the binary images and thin the image to single pixel width level. There are various standard functions are now available in MATLAB for thinning operation.

Smoothing

The objective of smoothing is to smooth shape of broken and noisy input scanned document image. The low pass filter is used for smoothing the image. Some pixels are added in the image so that a smooth shape may be obtained.

Skeletonization

Skeletonization is used for thinning. It refers to the process of reducing the width of lines like object from many pixels and can be remove irregularities in letters and in turn, makes the recognition algorithms simpler because they have to operate on a character stroke, which is only one pixel wide. It also decreases the memory space required for storing the information about the input scanned documents and no doubt this process decreases the processing time.

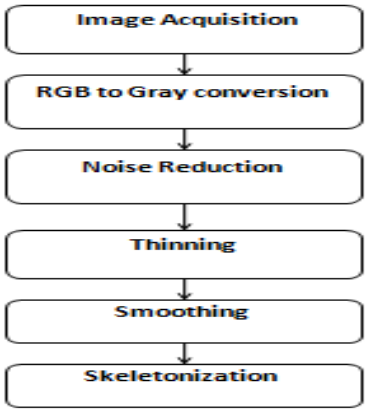

Fig. 1 Image pre-processing steps

Segmentation

In the process of segmentation the image of sequence of character is decompose large into sub images. The main goal of segmentation is to simplify and change the representations of an image into more meaningful and easily analyze. The segmentation is mostly used to line and curves etc. The segmentation is process of assigning a label to every pixel to increases the accuracy if recognition .The thresholding is simplest method of character segmentation. This method is used to convert the gray scale image into binary image. This is called balanced histogram thresholding.

IV. FEATURES EXTRACTION

The feature extraction is play very important role in pattern recognition and image processing because each character has its own different feature for each image. The feature extraction is provided the needed information of the pattern so that the task of classifying the pattern is made easy by formal procedure. The main goal of feature extraction is used to obtain the most relevant information from the original data set and represent the information in a lower dimensionality space. Some time the input data to an algorithm is too large and also may be redundant then the input data will be transformed into a reduced representation set of feature .The term feature extraction is transforms the input data into the set of feature. To find the feature of character image is very carefully. The feature set is use to extract the needed information from the input data in order to perform the task. In purpose system there are following feature extracted for scanned character documents:

Zoning: The character scanned image is divided into n*m zones. The densities of the points or some feature in different regions are analyzed and represented.

Crossing and distances: The crossing and distance is used for line segment in a specific direction and count the number of transitions from background to foreground pixels along vertical and horizontal lines through the character image and distances calculate .the distance of the first image pixel detected from the upper and lower boundaries of the image along vertical lines and from the left and right boundaries along horizontal lines.

Image cropping : The scanned image size is very high and high resolution. So the size of the input image must be decreases. The reduction is done very carefully that the aspect ratio remains same.

Orientation: The orientation is the angle in between ranging -90 to 90 degrees between the major axis and x-axis of the character that has the same second moments as the region.

Area perimeter: The ratio of area perimeters are obtained by dividing the number of non-zero pixels in a character to the length of the smoothest boundary.

Binariation of image: After gray scale conversion they obtained matrix calculation is very complicated because the elements in the matrix cover from 0 to 255 and then we make a

processing of binariation on image. The original gray image is 0 to 255 are converted into binary image 0 to 1. After the preprocessing is to binaries' the scanned document image is converted into binary image (black and white) having pixel value 0and 1. The scanning image is true color and this has to be converted into a binary image based on threshold value. The Ostu's method is used for this work for the purpose of selecting the threshold and binnarizing the gray scale image. The resulting image has 0 as background pixels of the image and 1 as foreground pixels of the image.

Ostu's method

In image processing the Ostu's method is used to automatically perform clustering based image thresholding or the reduction of a gray level image to a binary image. In Ostu's method we exhaustively search for the threshold that minimize the intra class variance and define a weighted sum of variances of two classes.

$$\sigma^2_{\omega}(t) = \omega_1(t)\,\sigma_1^2(t) + \omega_2(t)\,\sigma_2^2(t)$$

Where the weights ω_i are probabilities of two classes separated by a threshold t and σ_i^2 variances of these classes.

The Ostu's show that minimizing the intra class variance and is the same as maximizing inter class variance

$$\sigma_b^2(t) = \sigma^2\omega^2(t) = \omega_1(t)\,\omega_2(t)\,[\mu_1(t) - \mu_2(t)]^2$$

Which is expressed in terms of class probabilities ω_i and class means μ_i and the class probability $\omega_1(t)$ is computed from the histogram t. While the classes mean $\mu_{(t)}$ is:

$$\mu_1(t) = [\Sigma^t{}_0\rho(i)\,X(i)] / \omega_1$$

Where x (i) is the value at the center of the i^{th} histogram. Similarly you can compute omega-2(t) and $\boldsymbol{\mu_2}$ on the right hand side of the histogram for bines greater than t and the class probabilities and class can be compute iteratively.

Multilinear Discriminant analysis MDA

A MDA is an information processing paradigm that is inspired by the information process system. The novel structures of the information processing system are main elements of MDA. It composed a large number of highly inter connected processing elements working in union to solve specific problem. A MDA is specific application such as character recognition or data classification through learning process system. The MDA is used multilevel inter-related subspace can collaborate to discriminate different classes. The MDA algorithm can avoid the curse of dimensionality and solve the small sample size problems. It is helpful to decreasing the computational cost in the learning stage.

Where

$$\mathbf{Y_i = X_i x_1 U_1 \cdots x_{k-1} U_{k-1} x_{k+1} U_{k+1} \cdots x_n U_n}$$

Recognizes using neural network

The recognition of scanned documents is very complex problem. In scan documents image character has different size orientation thickness format and dimensions. The neural networks play very important role for character recognition. The recognize capability of neural network to generalize and insensitive the missing data would be very beneficial in scanned documents. In this paper we can use recognize for English scanned document using Feed Forward Multi- Layer Perceptron network with one hidden layer has been used. For training scanned document back propagation algorithm has been implemented.

Multilayer perceptron network algorithm:

Multilayer perceptron network with the BP algorithms have been applied to various type of problem. We used to recognize purpose for in this paper two layer perceptron on hidden layer and one output layer has been used.

In MLPN with back propagation training network algorithms the calculation and procedure as follow:

$$Fj(x) = 1/(1 + e^{-net}) \text{ and } net = \Delta wIjoi$$

Where the Dpk and Opk are desired and actual values of the output unit k and training pair p. Updating the weight is achieved by using following formulas:

$$W_{ij}(n+l) = W_{ij}(n) + \Delta W_{ij}(n)$$

$$\Delta W_{ij}(n) = \eta\delta X_j + \alpha\,(W_{ij}(N) - W_{ij}(n-1))$$

Where □ is the learning rate $W_{ij}(n)$ is the momentum, Wij(n) is weight from hidden layer node i from an input to node *j* at nth iteration Xi is either the output of unit i or is an input and x j is an error term for unit j.

$$\Delta j = O_j\,(l_j - O_j)\,(D_j - O_j)$$

The unit j is an internal hidden unit.

Then

$$\Delta j = O_j\,(l_j - O_j)\,\Sigma\delta_{kj}W_{kj}$$

Edge Detection Algorithm

The edge detection in the binaries image is done using sobel technique. After locating the edge the image is dilated and the holes present in the image are filled by using sobel technique. This operation performs in the last stages to produce the pre-processed image suitable for segmentation and improve the accuracy of optical character recognition. There are number of research have been used a Gaussian smoothed step edge as the simplest extension of the ideal step edge model foe modeling the effects of edge blur in practical application. Then one-dimensional image f which has exactly one edge placed at x=0 may be.

$$f(x) = \frac{Ir - Il}{2}\left(\text{erf}\left(\frac{x}{\sqrt{2}\sigma}\right) + 1\right) + Ii$$

The edge the intensity value of left side is:

$$Ii = \lim_{n \to -\infty} f(x)$$

And the right of the edge it is:

$$Ii = \lim_{n \to \infty} f(x)$$

The blur scale parameter σ is called the edge of image.

V. EXPRIMENTAL RESULTS

Character recognition using NN and MDA

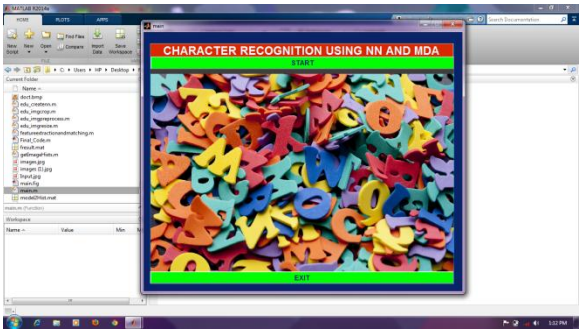

Fig. 2 recognition using NN and MDA

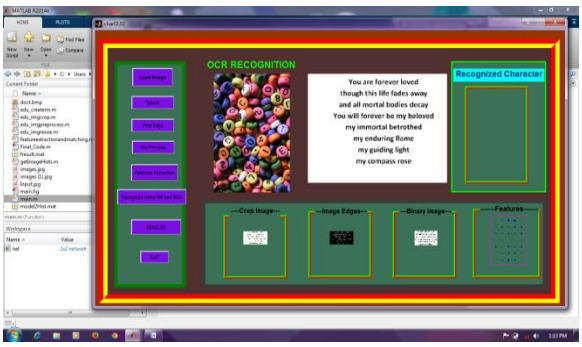

Fig 3 Load the scanned document image

1. Load an input scanned document image.
2. Select character from the input image.
3. Find edge using edge detection algorithm for input image.
4. The pre-processing can be done in next stage first we remove noise then convert it into binary image.
5. At last the feature extraction will be done by using pattern matching and the pattern match with the data base.
6. Finally we character recognize by using NN and MDA.

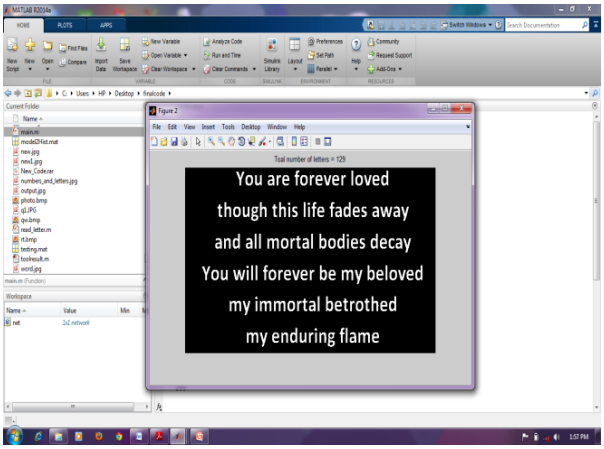

Fig .4 Inputs images and total number of character

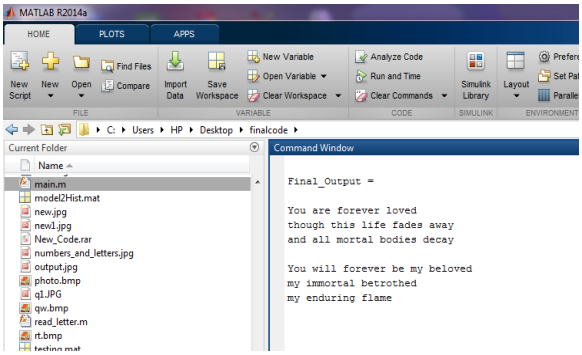

Fig. 5 Output of recognize image

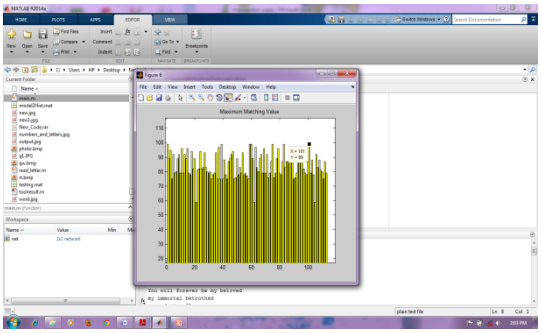

Fig. 6 maximum matching value

V. CONCLUSION

In this paper a system for recognizing English scanned document has been develop. This is very beneficial for various concepts involved and boost further advances in the area. The accuracy of recognition is directly depending on the nature of the material to be read and by its quality. The NN are mostly used to character recognition due to their high noise tolerance. In optical recognition the feature extraction step is very important. There are different types of factors affect the performance of OCR.There are different types of pre-processing and feature extraction techniques are used for increase the accuracy of OCR.

REFERENCES

[1] Mohammad Imrul Jubair & Prianka Banik, "An Approach to Extract Features from Document Image for Character Recognition" Global Journal of Computer Science and Technology Graphics & Vision, Volume 13 Issue 2 Version 1.0 Year 2013

[2] Sonam Jain, Harwinder Singh Sohal " A Novel Approach for Word Segmentation in Correlation based OCR System" International Journal of Computer Applications (0975 – 8887) Volume 99– No.18, August 2014

[3] Richa Goswami , O.P. Sharma " A Review on Character Recognition Techniques" International Journal of Computer Applications (0975 – 8887) Volume 83 – No 7, December 2013

[4] Thomas M. Breuel, Adnan Ul-Hasan, Mayce Al Azawi and Faisal Shafait† "High-Performance OCR for Printed English and Fraktur using LSTM Networks" 2013 12th International Conference on Document Analysis and Recognition

[5] Ivan Kastelan, Sandra Kukolj, Vukota Pekovic, Vladimir Marinkovic, Zoran Marceta, "Extraction of Text on TV Screen using Optical Character Recognition"10th Jubilee International Symposium on Intelligent Systems and Informatics • September 20-22, 2012

[6] Neha Sahu,R. K. Rathy PhD.,Indu Kashyap "Survey and Analysis of Devnagari Character Recognition Techniques using Neural Networks"International Journal of Computer Applications (0975 – 888)Volume 47– No.15, June 2012

[7] Parveen Kumar ,Nitin Sharma ,Arun Rana "Handwritten Character Recognition using Different Kernel based SVM Classifier and MLP Neural Network (A COMPARISON)"International Journal of Computer Applications (0975 – 8887) Volume 53– No.11, September 2012

[8] Anoop Rekha " Offline Handwritten Gurmukhi Character and Numeral Recognition using Different Feature Sets and Classifiers - A Survey "International Journal of Engineering Research and Applications (IJERA) "Vol. 2, Issue 3, May-Jun 2012

[9] Prof. S.P.Kosbatwar, Prof.S.K.Pathan "Pattern Association for character recognition by Back-Propagation algorithm

using Neural Network approach" International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.1, February 2012

[10] Anita Pal1 & Dayashankar Singh "Handwritten English Character Recognition Using Neural Network" International Journal of Computer Science & CommunicationVol. 1, No. 2, July-December 2010

[11] Kai Wang, Jianming Jin, Qingren Wang "High Performance Chinese/English Mixed OCR with Character Level Language Identification" 2009 10th International Conference on Document Analysis and Recognition

[12] Md. Abul Hasnat, S.M. Murtoza Habib, Mumit Khan "Segmentation Free Bangla OCR using HMM: Training and Recognition" 2nd International Conference on Electrical Engineering (ICEE),Khulna, Bangladesh, 2002

[13] Sobia T. Javed, Sarmad Hussain, Ameera Maqbool, Samia Asloob, Sehrish Jamil and Huma Moin "Segmentation Free Nastalique Urdu OCR" World Academy of Science, Engineering and Technology 70 2010