# English Scanned Document Character Recognition and Matched and Missed Matched Analysis using NN and MDA

Pardeep Kaur
Department of CSE
CTIEMT, Jalandhar, (PB), India

Pooja Choudhary
Department of CSE
CTIEMT, Jalandhar, (PB), India

Varsha Sahni
Department of CSE
CTIEMT, Jalandhar, (PB), India

## ABSTRACT
In this paper, the optical character recognition is used to recognize the scanned English documents by using neural network and MDA. The human mind easily read any interrupted scanned documents but it is difficult to machine. So the optical character recognition are solved this problem. The output images are not editable by capturing camera or scanned document but with the help of optical character recognition this problem easily solved. The OCR process consists of three major sub processes like pre processing, segmentation and recognition. The neural networks are playing very important role for character recognition its helps to provide high accuracy for the character.

## Keywords
English Character recognition, pre-processing scanned documents, segmentation, NN, feature extraction.

## 1. INTRODUCTION
Optical character recognitions have interested area for researchers. Recently, many organizations have need to received extensive documents attention in academic and production fields. The optical character recognitions are vast field in image processing and pattern recognition. In India, there are multi languages and multi scripts are used in different location, the eighteen officials scripts and accepted and have hundred regional languages. Today many researchers have been done optimize character to scanned English documents for character recognition 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 [1]. These algorithms are based on machine vision and artificial intelligence. For example neural network vectors machine fuzzy classifiers etc [2]. However, OCR is used machine encoding text and text can be easily edited, updated, and modified. OCR can be processed in many other ways according to requirements. It is also used small size for storage in comparison to scanned documents. The neural network is basically used in the fields of character recognition [3]. There are various phases of OCR involves to completely recognize and produce machine encoded text. The computer recognizes the scanned character in the documents through revolutionizing techniques called optical character recognition [8], [9]. The main phases of optical character recognition as: pre-processing phase, segmentation phase, feature extraction, and classification phase. In the OCR, there has been used following techniques

### 1.1 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. Generally, The Ostu's method is used for segmentation process. The gray level information it does not gives better segmentation results. The Ostu's method was proposed which works on both gray level thresholds of each pixel as well as its spatial correlation information within the neighborhood. The Otsu's method can obtain satisfactory segmentation result when it is applied to the noisy image.

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

Where the weights $\omega_i$ are probabilities of two classes separated by a threshold t and $\sigma_i^2$ 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 $\omega_i$ and class means $\mu_i$ 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.

### 1.2 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 [4]. 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.


**Fig. 1 edge detection using sobel technique**

### 1.3 Multilinear Discriminant analysis (MDA) and Linear Discriminant analysis (LDA)

The linear Discriminant Analysis is method used for data classification and dimensionality reduction. The LDA does not change the location but only tries to provide more class reparability and draw decision between given class. In Discriminant analysis two scatter matrices called within class and between class matrices [5], [6]. LDA classical algorithm has been successfully applied and extended to various biometric signal recognition problems. The recent advancements in multilinear algebra led to a number of multilinear extensions of the LDA, Multilinear Discriminant analysis proposed for the recognition of biometric signals using their natural tonsorial representation [7]. The MDA Check Multilinear projection and maps the input data from one space to another space. MDA is an information processing paradigm that is inspired by the information process system [13]. 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, $\mathbf{Y_i}=\mathbf{X_i}x_1U_1\cdots x_{k-1}U_{k-1}x_{k+1}U_{k+1}\cdots x_nU_n$

### 1.4 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. The neural network algorithms have been applied to various type of problem. In neural network the computing architecture is consists of massively parallel interconnection of adaptive neural processor. The neural network is parallel in nature so it can perform computations at a higher rate compared to other classical techniques. The neural network architectures can be classified as feed forward and feedback word. The information processing using neural network in paradigm is inspired by the biological nervous system such as the brain, process information. Many reports of scanned documents recognition in English have been published but till high recognition accuracy and minimum tranni9ng time of scanned English character using neural network. In this paper efforts have been made to develop scanned English documents character recognition with high recognition accuracy and minimum training and classification time.

$$\mathrm{E}p = \frac{1}{2}\Sigma \mathrm{o}(top - ypo)2 \quad (1)$$

$$\mathrm{G} = \frac{\partial E}{\partial wij} = \frac{\partial}{\partial wij}\Sigma pEp = \Sigma p\frac{\partial Ep}{\partial wij}$$

Chan rule

$$\frac{\partial E}{\partial \mathrm{woi}} = \frac{\partial \mathrm{E}}{\partial yo} * \frac{\partial \mathrm{yo}}{\partial woi} = \frac{\partial E}{\partial yo} = (to - yo) \quad (2)$$

$$y_o = \Sigma_j w_{oj} x_j \frac{\partial \mathrm{yo}}{\partial \mathrm{woi}} = \frac{\partial}{\partial \mathrm{woi}}\Sigma \mathrm{j}wojxj = \mathrm{x}i \quad (3)$$

Using the equation (2) and Equation (3);

$$\frac{\partial E}{\partial woi} = -(to - yo)\, xi$$

When applying the correction in a direction we get the following.

$\Delta w_{oi} = \eta(t_o - y_o)x_i$

This is a rate of learning.

## 2. METHODOLOGY

In character recognition, we have taken 40 records of scanned image in database. These images have been scanned through the scanner HP 1510.These images are taken the book of History of ICSE board that is published in 2013. Now each scanned image is stored in the data base for the character recognition. These scanned images are considered to PSNR, MSE, and matching time to recognize the each character forms the documents.

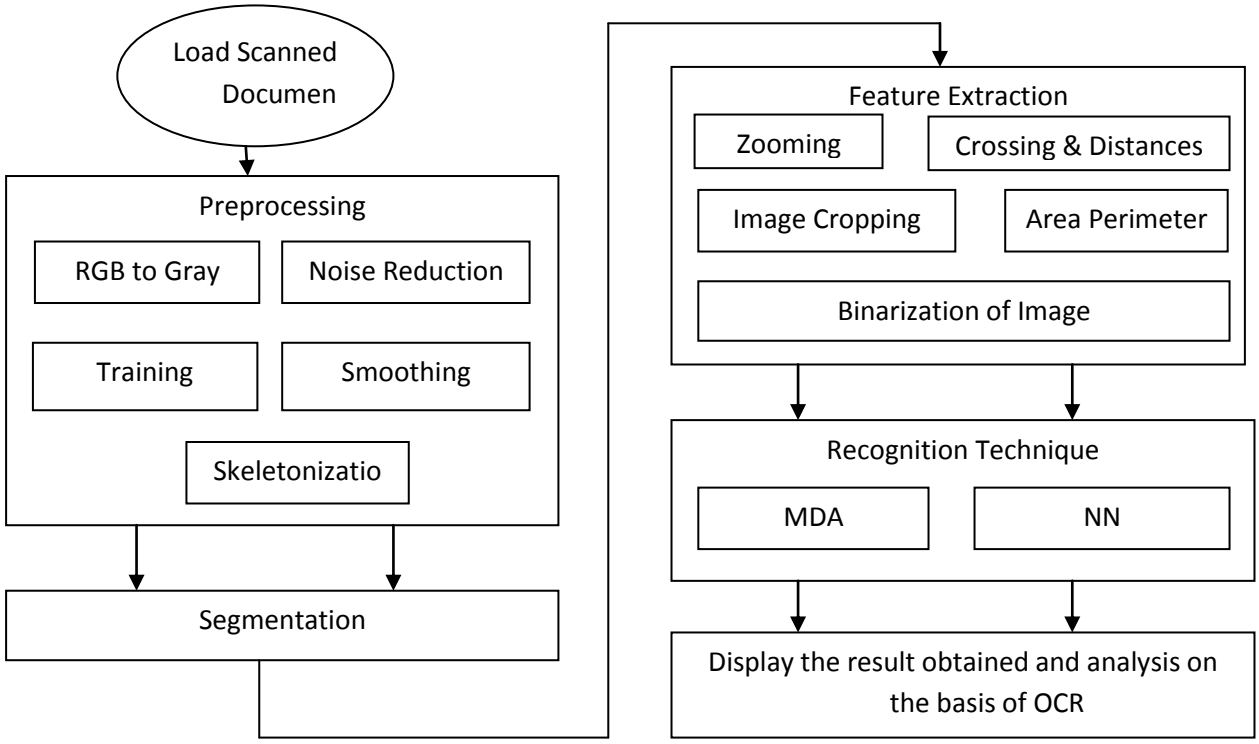
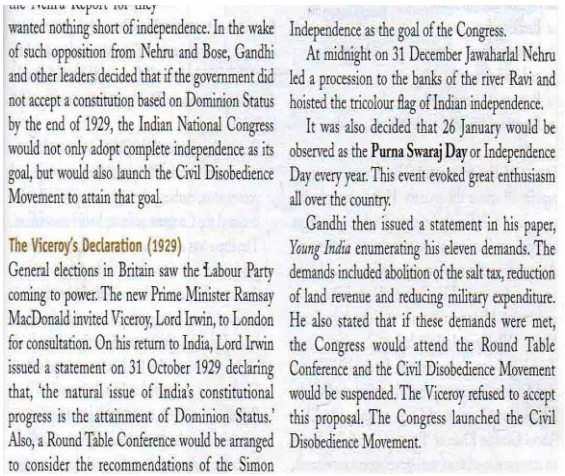



**Fig. 2 Character recognition Using MDA and NN**

**Pre-processing Steps**

1. To load the scanned input document image
2. Then select the character from the input image.
3. Find out the edge using edge detection algorithm for input image using sobel techniques.
4. The pre-processing can be done in next stage first we remove noise then convert gray scale image to binary image.
5. In the 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.



wanted nothing short of independence. In the wake of such opposition from Nehru and Bose, Gandhi and other leaders decided that if the government did not accept a constitution based on Dominion Status by the end of 1929, the Indian National Congress would not only adopt complete independence as its goal, but would also launch the Civil Disobedience Movement to attain that goal.

**The Viceroy's Declaration (1929)**

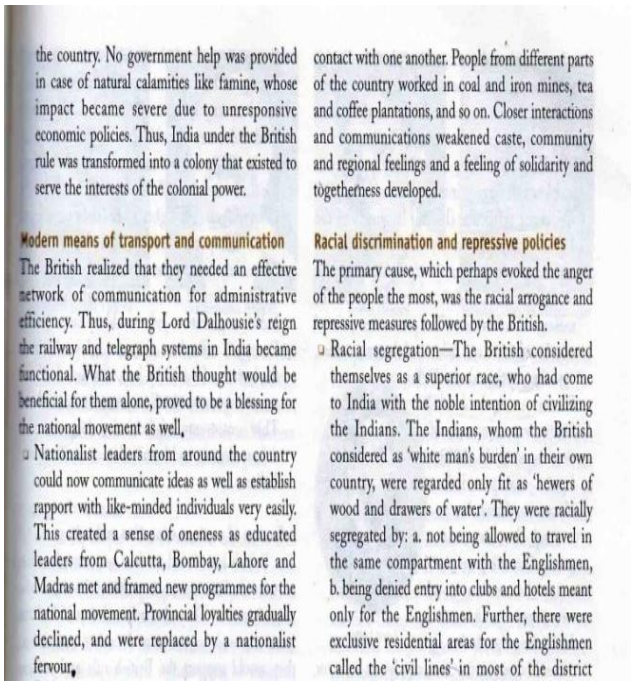
General elections in Britain saw the Labour Party coming to power. The new Prime Minister Ramsay MacDonald invited Viceroy, Lord Irwin, to London for consultation. On his return to India, Lord Irwin issued a statement on 31 October 1929 declaring that, 'the natural issue of India's constitutional progress is the attainment of Dominion Status.' Also, a Round Table Conference would be arranged to consider the recommendations of the Simon

Independence as the goal of the Congress.

At midnight on 31 December Jawaharlal Nehru led a procession to the banks of the river Ravi and hoisted the tricolour flag of Indian independence.

It was also decided that 26 January would be observed as the **Purna Swaraj Day** or Independence Day every year. This event evoked great enthusiasm all over the country.

Gandhi then issued a statement in his paper, *Young India* enumerating his eleven demands. The demands included abolition of the salt tax, reduction of land revenue and reducing military expenditure. He also stated that if these demands were met, the Congress would attend the Round Table Conference and the Civil Disobedience Movement would be suspended. The Viceroy refused to accept this proposal. The Congress launched the Civil Disobedience Movement.



the country. No government help was provided in case of natural calamities like famine, whose impact became severe due to unresponsive economic policies. Thus, India under the British rule was transformed into a colony that existed to serve the interests of the colonial power.

**Modern means of transport and communication**

The British realized that they needed an effective network of communication for administrative efficiency. Thus, during Lord Dalhousie's reign the railway and telegraph systems in India became functional. What the British thought would be beneficial for them alone, proved to be a blessing for the national movement as well.

- Nationalist leaders from around the country could now communicate ideas as well as establish rapport with like-minded individuals very easily. This created a sense of oneness as educated leaders from Calcutta, Bombay, Lahore and Madras met and framed new programmes for the national movement. Provincial loyalties gradually declined, and were replaced by a nationalist fervour.

contact with one another. People from different parts of the country worked in coal and iron mines, tea and coffee plantations, and so on. Closer interactions and communications weakened caste, community and regional feelings and a feeling of solidarity and togetherness developed.

**Racial discrimination and repressive policies**

The primary cause, which perhaps evoked the anger of the people the most, was the racial arrogance and repressive measures followed by the British.

- Racial segregation—The British considered themselves as a superior race, who had come to India with the noble intention of civilizing the Indians. The Indians, whom the British considered as 'white man's burden' in their own country, were regarded only fit as 'hewers of wood and drawers of water'. They were racially segregated by: a. not being allowed to travel in the same compartment with the Englishmen, b. being denied entry into clubs and hotels meant only for the Englishmen. Further, there were exclusive residential areas for the Englishmen called the 'civil lines' in most of the district

**Fig. 3 input scanned image from Database**

# 3. RESULT DISCUSSION

## 3.1 MSE

In Fig. 4, we have calculated the MSE between the previous method and the proposed method. It is calculated to reduce the mean square error and measure the PSNR value for the English character recognition from the scanned documents. Our method is better as compared to the previous method on the basis to computed results.

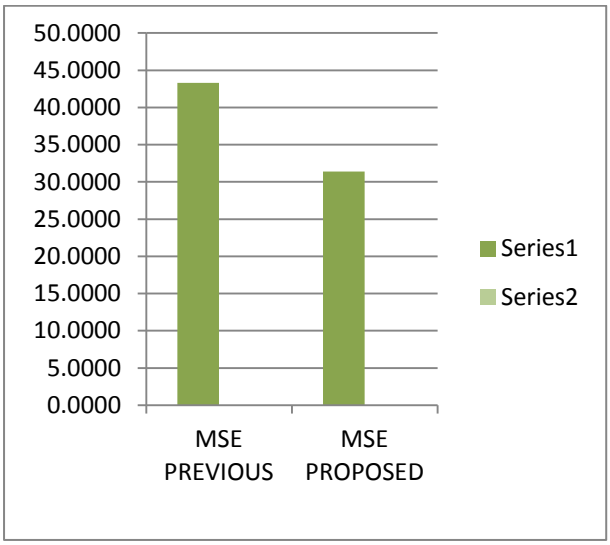



**Fig. 4 Comparison of MSE for OCR**

## 3.2 PSNR

In Fig. 5, we have calculated the psnr value for the quality of the signals to improve the English character recognition using the previous method and the proposed method. Our method is better to calculate the PSNR value as compared to the previous method.

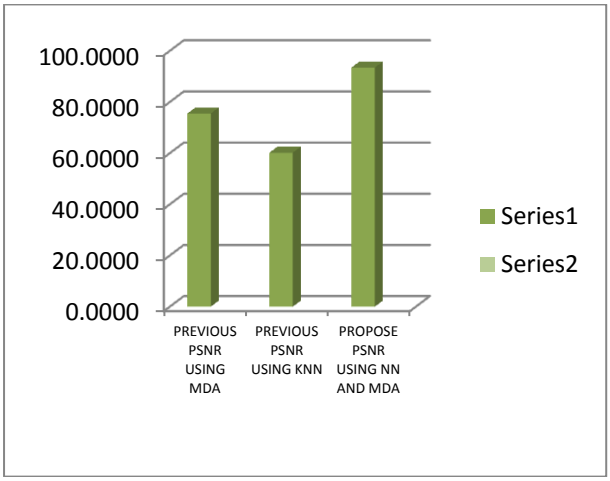



**Fig. 5 Comparison of PSNR for OCR**

## 3.3 Matching Time

matching time measures the time to matching the characters from the scanned documents. Our proposed method gives the less matching time as compared to the previous method. In Fig. 6, we have shown the matching time comparison as follows:

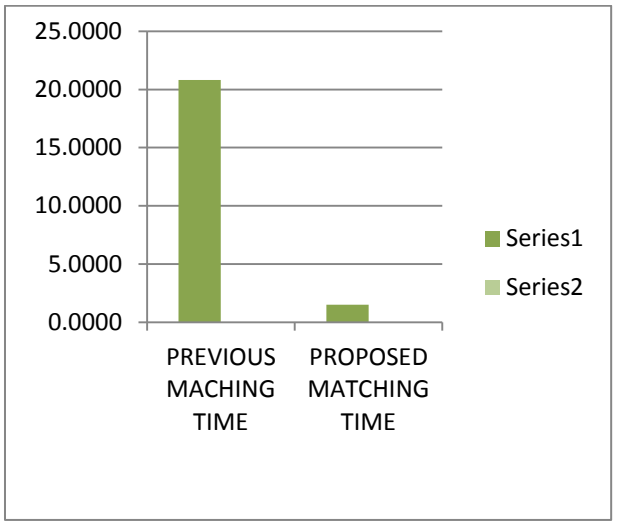



**Fig. 6 Comparison of Matching Time for OCR**

## 3.4 Matched and Missed Matched Character

In fig. 7, we have found the matched character and missed matched character out of total character in the line of the scanned document using the previous method and the proposed method. Our method has given the better matching character results as compared to the previous method as shown below:

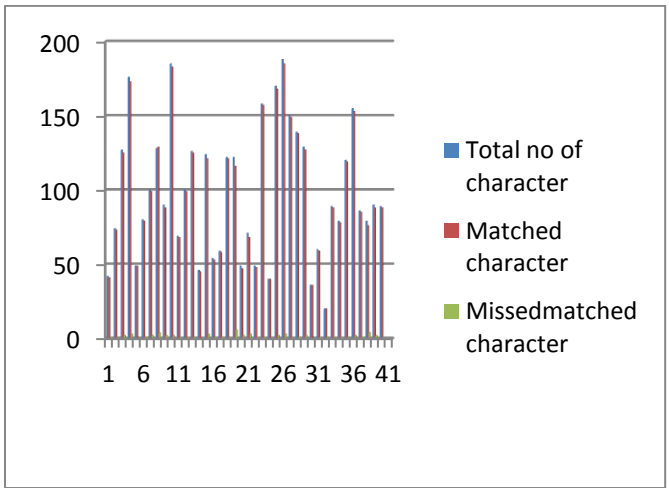



**Fig. 7 Comparison of Matched and Missed Matched character for OCR**

In table 1, we have shown the matched and missed matched character in the scanned document in the optical character recognition. Our proposed method has shown more matching character and very less missed matched character.

**Table 1 Matched and Missed Matched character for OCR**

| S.no | Total no of character | Matched character | Missedmatched character | character mismatched name |
|---|---|---|---|---|
| 1 | 42 | 41 | 1 | i |
| 2 | 74 | 73 | 1 | i |
| 3 | 127 | 125 | 2 | I,r |
| 4 | 176 | 173 | 3 | I,b,n |
| 5 | 49 | 49 | 0 | |
| 6 | 80 | 79 | 1 | i |
| 7 | 100 | 99 | 2 | I,t |
| 8 | 128 | 129 | 4 | I,r,t,n |
| 9 | 90 | 88 | 2 | I,p |
| 10 | 185 | 183 | 2 | i |
| 11 | 69 | 68 | 1 | i |
| 12 | 100 | 99 | 1 | i |
| 13 | 126 | 125 | 1 | i |
| 14 | 46 | 45 | 0 | |
| 15 | 124 | 121 | 3 | I,r,t,n |
| 16 | 54 | 53 | 1 | |
| 17 | 59 | 58 | 1 | i |
| 18 | 122 | 121 | 1 | i |
| 19 | 122 | 116 | 6 | I,r,t,n,m,b |
| 20 | 49 | 47 | 2 | I,r |
| 21 | 71 | 68 | 3 | I,r,n |
| 22 | 49 | 48 | 1 | i |
| 23 | 158 | 157 | 1 | n |
| 24 | 40 | 40 | 0 | |
| 25 | 170 | 168 | 2 | I,n |
| 26 | 188 | 185 | 3 | I,n,r |
| 27 | 150 | 149 | 1 | i |
| 28 | 139 | 138 | 1 | i |
| 29 | 129 | 127 | 2 | I,n |
| 30 | 36 | 36 | 0 | |
| 31 | 60 | 59 | 1 | i |
| 32 | 20 | 20 | 1 | i |
| 33 | 89 | 88 | 1 | i |
| 34 | 79 | 78 | 1 | n |
| 35 | 120 | 119 | 1 | i |
| 36 | 155 | 153 | 2 | I,n |
| 37 | 86 | 85 | 1 | i |
| 38 | 79 | 76 | 4 | I,n,r,t |
| 39 | 90 | 88 | 2 | I,n |
| 40 | 89 | 88 | 1 | r |

## 4. CONCLUSION

In this research paper, the character recognition is crucial problem to improve the character from the scanned documents. To provide good opportunity, we have needed to improve the English character recognition from the state-of-art techniques and the proposed method. In this paper our proposed method neural network and MDA method is well suited for the whole scanned documents to provide the matching time, matched character and missed matched character from the whole scanned document for used. In further research work, English character recognition may be possible from the digital electronic devices such running for live application devices.

## 5. REFERENCES

[1] Pardeep kaur and Pooja Choudhary, "Review on: English Scanned Documents", International Journal Engineering Research, Vol. 3, Issue.2, 2015.

[2] 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

[3] Kauleshwar prsad , devvrat C.Nigam,Ashmika Lakhotiya and Dheeren Umre B.I.T Durg, India "Chracter recognition using Matlab neural network tool" International journal of u and e service and technology vol 6 no1 Feb. 2013.

[4] S.K. Thilagavathy and De R. Indra Gandhi "recognition of Distorted character using Edge Detection Algorithm" International Journal of Innovative Research in computer and communication Engineering vol 1 issue 4 June 2013.

[5] Ayatullah Faruk Mohllah, Nabamita Majumder, Subhadip Basu and Mita Nasipuri "Design of an Optical Chracter Recognition System for Camera based

Handheled Devices" IJCSI International Journal of computer science issue vol 8 issue 4 no 1 july 2011.

[6] Aamir khan Hasan Farooq "Principal Component Analysis Linear Discriminant Analysis Feature Extractor for pattern Recognition" IJCSI international journal of computer science issue vol 8 issue 6 no 2 nov.2011.

[7] Yusuf Perwej and Ashish Chaturvedi " Machine Recognition of Hand written Chracter using Neural networks" Internation journal of computer application vol 14, no 2 Jan. 2011.

[8] Vivek Shrivastava and Navdeep Sharma "Artificial Neural Network based optical character recognition" Signal & Image Processing : An International Journal (SIPIJ) Vol.3, No.5, October 2012.

[9] Md Fazlul Kader1 and Kaushik Deb2 "Neural Network based English Alphanumeric recognition" International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.4, August 2012

[10] 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

[11] 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

[12] 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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] Reetika Verma1 ,Rupinder Kaur2 "Efficient Technique for Chracter Recognition using neural network and surf Feature Extraction" Reetika Verma et al, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014

[21] Rapanjot Kaur, Gagangeet Singh Aujla "Review on: Enhanced Offline Signature Recognition Using Neural Network and SVM