ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

Recognition and Classification of Broken Characters using Feed Forward Neural Network to Enhance an OCR Solution

Manas Yetirajam, Manas Ranjan Nayak, Subhagata Chattopadhyay

***Abstract*— The major problem faced by an Optical Character Recognizer (OCR) can be attributed to the presence of a large number of broken characters in a digital document image. Recognition of such documents accurately, that contain abundant broken characters still remains a challenge to most of the OCR solutions. Multi Layer Feed Forward Neural Network Classifier (MLFNC) can be used to enhance the efficiency of an OCR. MLFNC tries to improvise the recognition by classifying broken characters into a different group. This gives an opportunity to process broken characters in a more effective manner separately. Here, a new method has been proposed which uses feed forward neural network to classify broken characters prior any processing is done by an OCR with a considerable accuracy. MLFNC is a simple network with a very small time complexity due to which, there is a least effect on the time complexity of the solution provided by OCR.**

***Index Terms*— Binarization, Broken Characters, Character Classification, Feature Extraction, Feed forward Neural Network;**

I. Introduction

Recognition performed by an OCR depends on the quality of the printed characters. When the solution provided by it is used for processing old historic texts, the accuracy drops drastically [1]. The major problem for this drop in accuracy is the presence of a large number of distortions such as (i) broken characters (ii) touching characters (iii) fax documents and (iv) heavy print in such a document [2]. This paper however limits itself to only broken characters. This problem can be dealt by classifying all broken characters in a class and processing.

Manuscript received Oct 10, 2012.

***Manas Yetirajam**, Dept. of Computer Science and Technology, National Institute of Science and Technology, Berhampur, India, Mob: +919439156325.*

***Manas Ranjan Nayak**, Dept. of Computer Science and Technology, National Institute of Science and Technology, Berhampur, India, Mob: +919861773088.*

***Subhagata Chattopadhyay**, Principal of Bankura Unnayani Institute of Engineering, India.*

In order to solve this problem more emphasis has to be given to the preprocessing phase, such as segmentation and document binarization. While dealing with broken characters, various options are generally available: enhancing and restoring the quality of character [3][4] or recovering the broken character proposed in [5]. There has been a lot of work in this regard with a variety of approaches yielding satisfactory results [6]-[9]. This work proposes an approach to recover broken characters by isolating damaged characters in a class. By this, broken characters are separated from a document and hence these characters can be easily replaced with their equivalent unbroken form.

The quality of old historic document is not so good; sharpness of the characters is often blurred and the color is faded .This type of documents cannot be directly segmented for recognition purpose [10][11]. Thus preprocessing of these documents is very important to facilitate the task of segmentation. In this work, *image blurring* and *image binarization* has been used to deal with these imperfections in the document. After preprocessing, the outcome is a binary image with a black background and a white colored character. Based on the average coverage of the character a grid dimension is calculated and the entire document is divided into the number of grids with same dimensions. Based on the connected components, the image in a grid is segmented. If a grid contains more than one segment then the character is known to be broken, but if a grid contains only one segment the character can be either broken or regular.

(a) (b)

Fig 1: Devanagri (ka) regular and broken character having only one segment.

All Rights Reserved © 2012 IJARCET

Segmentation of above characters in fig 1(a) produces only one segment but fig 1(b) is a broken character. Thus, by simply having a count of the number of segments, we cannot come to a conclusion whether a character is regular or broken.

Here, a Multi Layer Feed Forward Neural Network Classifier (MLFNC) has been used to solve this problem which is caused due to segmentation. An MLFNC is an artificial neural network in which, the information moves in only one direction, i.e. forward, from the input nodes, through the hidden nodes (if any) to the output nodes. There are no cycles or loops in the network. It is thus a widely used soft computing technique in various research domains [12]-[20]. Fig 2 shows a FFNN with 4:3:1 structure.

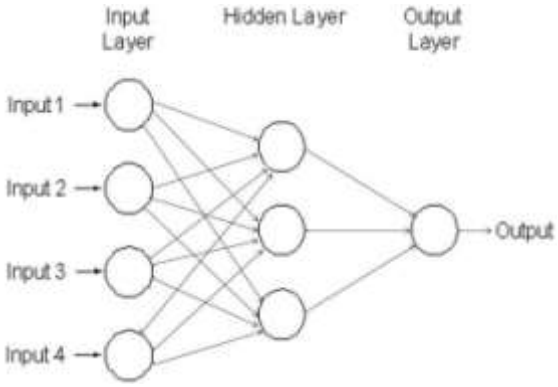


Fig 2: Feed Forward Neural Network topology.

II. Methodology

2.1 Preprocessing

Preprocessing aims to diminish the effect of spurious noise and to blur the false contours that may be present in a digital document. A low pass filter (*Gaussian blur*) has been used for this purpose to reduce the high frequency noise component of the image (refer to Fig.3).

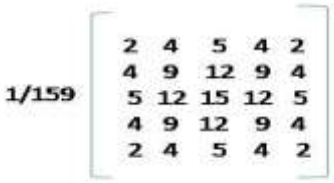


Fig 3: Gaussian Blur used for preprocessing

2.2 Binarization

Binary thresholding has been carried out by computing the total variance of the entire image by considering all the pixels.

The total variance computed is used as an initial threshold.

$$v = \sum_{i \in Q} \frac{(m - x(i))^2}{|Q-1|} \quad \ldots.(1)$$

Where ‘v’ denotes total variance

m: mean of the pixels in the image

x (i) : each pixel from the set of pixels Q

Result of binarization is a binary image document with a black background and a white colored character in it. Then the document is divided into grids of same size based on the average area of coverage of a character in a document.

Document binarization provides a low level representation to the character in each grid of document. Fig 4 shows a character in a grid 4(a) and its low level representation 4(b) to facilitate feature extraction.

(a) (b)

Fig 4: Low level representation of a character in a grid

2.3 Feature Extraction

Each grid is separately processed to extract features of a character in a grid. Following are the features which have been extracted for each character.

- Height
- Width
- Aspect Ratio
- Diagonal coverage Distance
- No. of pixels
- Foreground Background ratio (Fb-ratio)
- Curvature ratio
- Distance from center
- X-variance
- Y-variance

These features are fed to the MLFNC as an input for further processing. It has been found that by an increase in the number of features the accuracy of recognition also increases, this is discussed in section 3.

2.4 Feed Forward Neural Network

All Rights Reserved © 2012 IJARCET

ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

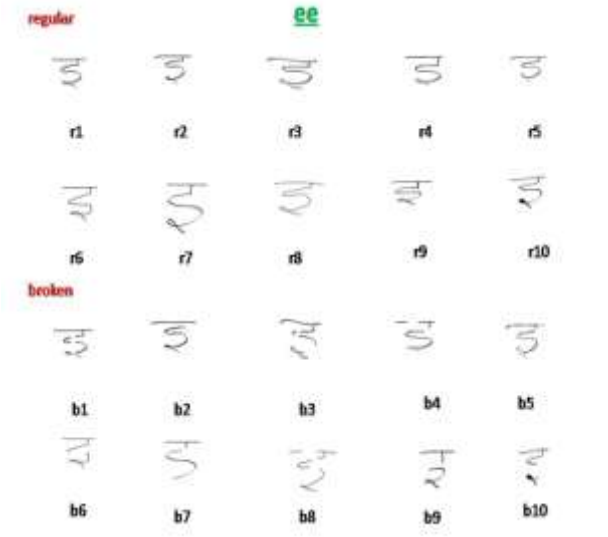
In this work, a Multi Feed-forward Neural Network Classifier (MLFNC) is designed with a network architecture as n:(n/2):1

Where 'n': number of input nodes for the input layer;
'n'=6,10
Results for two values of 'n' were observed and compared which is discussed in section 3. The transfer function used in the hidden layer is Gaussian followed by Sigmoidal. The transfer function used for output layer is also Sigmoidal to keep the values between 0 and 1.

III. Results and Discussion

Features for 3 different alphabets sets each having 10 regular (r*i*) and 10 broken (b*i*), where i=1 to 10 cases ("aa", "ee", "cha") where extracted and fed to the MLFNC.

(a)



(b)

(c)

Fig 5: Set of alphabets used as data set

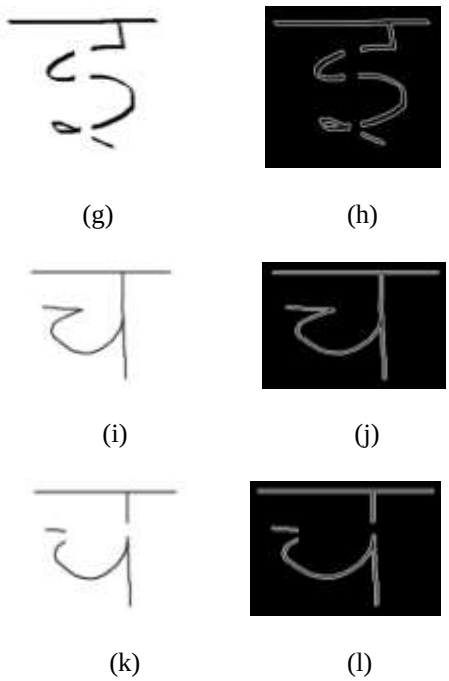
Result of binarization and feature extraction for characters from each of 3 sets ("aa", "ee", "cha") , Fig 5(a)r1, Fig 5(a)b1, Fig 5(b)r1, Fig 5(b)b1, Fig 5(c)r1, Fig 5(c)b1 is shown in Fig 6 and Table 1, Table 2, Table 3. Ten features mentioned in section 2.3 were fed to MLFNC. Well defined 10 regular characters and 10 broken characters features were also used to train the network for each of 3 alphabet sets("aa", "ee", "cha"), and the corresponding training mean was computed shown in Table 4.

(a) (b)

(c) (d)

(e) (f)

All Rights Reserved © 2012 IJARCET



(g) (h)

(i) (j)

(k) (l)

Fig 6: Showing the result of binarization on characters.

Table 1: Extracted Features for Fig 6(a), Fig 6(c)

Features	Values	
	Fig 6(a) (regular)	Fig 6(c) (broken)
Height	312	313
Width	309	309
Aspect Ratio	1.009708	1.012944
Diagonal Distance	439.118435	439.829512
No. of pixels	11938	9320
Fbratio	0.046287	0.035773
Curvature ratio	0.904472	1.828034
Distance from center	150.695056	287.031541
X-variance	9387.790956	10170.6200893
y-variance	7069.394817	6806.026313

Table 2: Extracted Features for Fig 6(e), Fig 6(g)

Features	Values	
	Fig 6(e) (regular)	Fig 6(g) (broken)
Height	251	252
Width	188	188
Aspect Ratio	1.335106	1.340425
Diagonal Distance	313.600063	314.401017
No. of pixels	6881	6185
Fbratio	0.026166	0.023457
Curvature ratio	0.755329	0.980654
Distance from center	216.462467	324.388017
X-variance	3808.358083	3844.920201
y-variance	2964.367087	3192.966981

Table 3: Extracted Features for Fig 6(i), Fig 6(k)

Features	Values	
	Fig 6(i) (regular)	Fig 6(k) (broken)
Height	261	261
Width	201	201
Aspect Ratio	1.298507	1.298507
Diagonal Distance	329.426774	329.426774
No. of pixels	5317	4723
Fbratio	0.020099	0.017814
Curvature ratio	0.772287	0.962383
Distance from center	208.655218	299.061172
X-variance	3761.568896	4166.567972
y-variance	4013.556928	4222.622120

During testing, difference between testing value and training mean (regular and broken) was computed. Minimum difference corresponds to regular character and maximum difference to broken character. Qualitative results of testing are given in Table 5 and Table 6.

Table 4: Training Mean

Character	regular	broken
aa	0.567798	0.575717
ee	0.560683	0.561857
cha	0.561054	0.568972

Table 5: Testing Result for Regular Characters

Fig	Char->	aa	ee	cha
	Actual	Obtained	Obtained	Obtained
r1	REGULAR	REGULAR	BROKEN	REGULAR
r2	REGULAR	BROKEN	BROKEN	REGULAR
r3	REGULAR	REGULAR	REGULAR	BROKEN
r4	REGULAR	REGULAR	REGULAR	BROKEN
r5	REGULAR	REGULAR	REGULAR	REGULAR
r6	REGULAR	REGULAR	REGULAR	REGULAR
r7	REGULAR	REGULAR	REGULAR	REGULAR
r8	REGULAR	REGULAR	BROKEN	BROKEN
r9	REGULAR	REGULAR	BROKEN	BROKEN
r10	REGULAR	BROKEN	REGULAR	REGULAR

Table 6: Testing Result for Broken Characters

Fig	Char->	aa	ee	cha
	Actual	Obtained	Obtained	Obtained
b1	BROKEN	REGULAR	BROKEN	BROKEN
b2	BROKEN	BROKEN	BROKEN	BROKEN
b3	BROKEN	BROKEN	REGULAR	BROKEN
b4	BROKEN	BROKEN	BROKEN	BROKEN
b5	BROKEN	REGULAR	BROKEN	BROKEN
b6	BROKEN	REGULAR	BROKEN	REGULAR
b7	BROKEN	REGULAR	BROKEN	REGULAR
b8	BROKEN	BROKEN	BROKEN	BROKEN
b9	BROKEN	BROKEN	BROKEN	BROKEN
b10	BROKEN	BROKEN	REGULAR	REGULAR

All Rights Reserved © 2012 IJARCET

Table 7: Accuracy (%) of Testing Results

Number of Features ->	6	10
aa	50	70
ee	35	70
cha	50	65

From Table 7 it can be seen that by increasing number of features percentage of accuracy increases drastically.

IV. Conclusion

The results obtained by the designed MLFNC shows that in most of the cases, broken and regular characters are classified correctly with an accuracy of 68.33 % .Thus it facilitates the work of an OCR by providing a chance to process broken characters separately. In this work it has been observed that accuracy of classification has changed from 45% to 68.33% by changing number of features from 6 to 10. Since MLFNC is a simple network with a very small time complexity, when compared to Back Propagation Neural Network (BPN) or any other complex neural networks having a high degree of accuracy and time complexity, it tries to improvise the recognition process pursued by an OCR for old historic document, efficiently with very less time complexity.

References

[1]. S. Chaivatna, T. Supachai, "*Recognizing Broken Thai Characters Based on Set-Partitions and N-grams Graphs*", Journal of Pattern Recognition Research 1 (2012) 26-41.

[2]. M. K. Shukla, Dr. H. Banka, "*A Study of Different Kinds of Degradation in Printed Bangla Script*", International Journal of Advanced Computer *Engineering and Architecture Vol.2 143-151*.

[3]. A.Whichello, H. Yan, "*Linking broken character borders with variable sized masks to improve recognition*", Pattern Recognition Elsevier. 29 (8) (1996) 1429-1435.

[4]. B.Allier, N. Bali, H. Emptoz, "*Automatic accurate broken character restoration for patrimonial documents*", IJDAR 8 (4) (2006) 246--261.

[5]. K. Lee, H. Byun, and Y. Lee, "*Robust Reconstruction of Damaged Character Images on the Form Documents*", *Lecture Notes in Computer Science 1389*, Springer Verlag, pp. 149-162, 1998.

[6]. A.H. Pilevar , M.T. Pilevar , " *Broken and Touching Characters Recognition in Persian Text Documents* ", world Applied Sciences Journal 13 (6) pp. 1459-1464,2011.

[7]. D. Yu, H. Yan ," *Reconstruction of broken handwritten digits based on structural morphological features*", *The Journal of Pattern Recognition Society*,vol.34 , pp. 235-254, 2001.

[8]. C. Sumetphong , S. Tangwongsan , " *Recognizing Broken Thai Characters Based on Set-Partitions and N-grams Graphs*," *Journal Of Pattern Recognition Research,* vol.7, pp. 26-41,2012.

[9]. L.L.Sulem, M. Sigelle, "*Recognition of degraded Characters using dynamic Bayesian networks*", *Journal of pattern Recognition*, vol. 41, pp. 3092- 3103, 2008.

[10]. P. Nucharee, P. Wichian, P. Ubolrat and S. Narita, "*Broken Characters Identification for Thai Character Recognition Systems* ", conference digest2003 464-179.

[11]. L. L. Sulem, M. Sigelle, *"Recognition of degraded characters using dynamic Bayesian networks*", Pattern Recognition Elsevier 41 (2008) 3092 – 3103.

[12]. S. Chattopadhyay, "Neurofuzzy Models to Automate the Grading of Old-age Depression". *Expert Systems: the Journal of Knowledge Engineering* (2012); DOI: 10.1111/exsy.12000 (in press)

[13]. U.R. Acharya, V.S. Sree, S. Chattopadhyay, J.S. Suri "Automated Diagnosis of Normal and Alcoholic EEG Signals", *International Journal of Neural Systems* (in press)

[14]. U. R. Acharya, F. Molinary, V.S. Sree, S. Chattopadhyay, K-H. Ng, J.S. Suri "Automated Diagnosis of Epileptic EEG using Entropies". *Biomedical Signal Processing and Control (2012)*; 7: 401-407

[15]. T. Dash, T. Nayak, S. Chattopadhyay "Offline Handwritten Signature Verification using Associative Memory Net". *International Journal of Advanced Research in Computer Engineering & Technology* (2012): 1(4): 370-374

[16]. S. S. Behera, S. Bhanja Choudhuri, S. Chattopadhyay "A Comparative Study on Neural Net Classifier Optimizations". *International Journal of Advanced Engineering and Technology*, (2012) 4(2): 179-187

[17]. T. Dash, S. Chattopadhyay, T. Nayak "Handwritten Signature Verification using Adaptive Resonance Theory Type-2 (ART-2) Net". *Journal of Global Research in Computer Science* (2012) 3(8): 21-25.

[18]. S. Chattopadhyay, P. Kaur, F. Rabhi, U. R. Acharya "Neural Network Approaches to Grade Adult Depression", *Journal of Medical Systems* (2012), 36(5): 2803-2815

[19]. A. Dasari, N.B. Hui, S. Chattopadhyay "A Neuro-fuzzy System for Modeling the Depression Data", *International Journal of Computer Applications* (2012) 53(6): 1-6.

[20]. K. Misra, S. Chattopadhyay, D. Kanhar "A Hybrid Expert Tool for the Diagnosis of Depression", *of Medical Imaging and Health Informatics* (2012); accepted on 01/10/12, in press.

Manas Yetirajam and **Manas Ranjan Nayak** have completed their B.Tech, in the Dept. of Computer Science and Engineering at NIST, India. Soft computing, Pattern Recognition, and Image Processing are their key research interests. They have published one research paper together.

Dr. Subhagata Chattopadhyay earned his MBBS and DGO from Medical College & Hospital, Kolkata India, a M.Sc. degree in Bioinformatics from Sikkim Manipal University and a PhD degree in Information Technology from Indian Institute of Technology, Kharagpur India. He pursued a prestigious postdoctoral fellowship in the University of New South Wales (UNSW), Sydney Australia, where he worked on a large international project on e-Health systems, supported by the World Health Organizations (WHO), Geneva Switzerland. Pattern recognition, Artificial intelligence and Soft Computing remain his key research fields. He has published over ninety research papers in various refereed international journals, books and conferences of repute. His biography has been selected as a distinguished medical professional in Marquis Who's Who (Biomedical researcher) in 2006-2007. Currently Prof. Chattopadhyay is the Principal of Bankura Unnayani Institute of Engineering, India.

All Rights Reserved © 2012 IJARCET