Automatic Recognition of Handwritten Bengali Broken Characters (BBC): Simulating Human Pattern Matching

Manas Ranjan Nayak
Dept. of CSE
National Institute
of Science and Technology
Berhampur, Odisha,
India

Saswat Nayak
Dept. of CSE
National Institute
of Science and Technology
Berhampur, Odisha,
India

Yetirajam Manas
Dept. of CSE
National Institute
of Science and Technology
Berhampur, Odisha,
India

Sangeeta Bhanja Chaudhuri
Dept. of CSE
National Institute
of Science and Technology
Berhampur, Odisha,
India

Subhagata Chattopadhyay
Dept. of CSE
Camellia
Institute of Engg.
Kolkata-129,
West Bengal,
India

ABSTRACT
This paper presents an automatic detection of handwritten Bengali Broken Characters (BBC) using a feed forward neural network (FFNN). It simulates the Human Visual System (HVS) the way human eye matches the patterns of the broken characters to a meaningful character and identifies it. Here the challenge is to detect and retrieve handwritten character which has been distorted up to 90%. The database consists of fifty bangle characters, each with twenty samples. Each character is presented as an image, which has been preprocessed, segmented and the features are then extracted. A new method has been proposed in this paper. It uses FFNN to calculate the mismatch for the recognition of a character, where it is observed that the distorted characters show very low mismatch with the original characters. For example, characters up to 70% distortions are found to be retrieved effectively.

Keywords
Bengali Broken Character (BBC); Pattern matching; Feed-Forward Neural Network (FFNN); Handwriting Recognition

General Terms
Image processing; Neural network; Feature extraction; Distorted Bengali characters

1. INTRODUCTION
Pattern recognition is a large domain involving complex research domains [1-10]. *Recognition* of handwritten characters is studied since decades and much work has been done in this field. Optical Character recognition is an application of *pattern recognition* and is a multidisciplinary field which is a combination of i) Image Processing and ii) Artificial Intelligence techniques. When it comes to recognizing *broken* handwritten characters, it is challenging using normal recognition methods [11]. These broken characters are found in the old text and historical documents. Such manuscripts are needed to be taken care of; otherwise, there is a fair chance of losing precious information for e.g., important facts and figures. Hence, the objective is to preserve those for further research and queries.

Character image preprocessing is done to extract feature characteristics from the images which can be used for accurate recognition. The importance of preprocessing lies in its ability to remove some of the problems which might occur due to the factors such as i) Quality of the paper, ii) scan resolution, iii) type of printed documents, iv) Scanner quality and so on [12]. *Character Recognition* is a popular application of *Pattern Recognition*. Though much research work has been done on characters, there is still much scope to work on *Broken Character recognition* in Bengali language, i.e., Bengali Broken Characters (BBC). Only a few studies have been reported related to broken character research. The reason behind the fact is the missing information in the broken characters due to ink fading, inadequate scanning, tired printer or copier cartridges, misadjusted impact printers worn ribbons, faxed document, dot matrix text and many more issues [13].

2. LITERATURE REVIEW
The literature review shows that, though a few research work have been proposed for recognizing the degraded character in a scanned document image, there has been a great deal of versatility in the techniques used. There has been a proposed method for Hidden Markov Model (HMM) to the recognition of the degraded and touching texts [14]. A technique for segmentation of the touching character using projection profile and topographic feature extracted from the gray scale image has also been used [15]. A method for double differential function has been constructed to segment the touching character [16]. Others [17] have proposed a very useful method decision tree for resolving ambiguity in segmentation touching character. A useful method for recursive segmentation for touching character has been constructed [18]. Hong [19] attempted to visual inter-word constraint available in a text image to split word image into pieces for segmentation degraded roman script character. Lu [20] proposed to measure different technique for mapping vertical projection profile on the second projection called peak to valley ratio.

Some work has also been reported to recognize the degraded character in Indian language script document. A very useful principal of water overflow from a reservoir is used to segment

the touching character in Oriya script [21]. The structure properties for segmentation of the touching character in middle and upper zone of printed Gurumukhi script has also been implemented [22]. A method to segment the touching character in upper zone of Gurumukhi script has also been suggested [23, 24]. A technique for segmentation of the conjucticton (one kind of touching patterns) in Devnagari script using the structure properties of the script has also been implemented [25]. Apart from these, other relevant works have been enlisted in Table 1.

Table-1 Relevant literature review

Author (Yr.)	Aim	Technique	Findings (accuracy)
Manas et al., (2012) [26]	OCR of broken Hindi characters (up to 90%)	Multilayer feed-forward neural network	~64% accuracy
Sahu et al., (2012) [27]	Fingerprint identification	Modified tree matching	High
Sumetphong & Tangwongsan (2012) [28]	Recognition of broken Thai Characters	Set partitioning Technique.	96%.
Abubacker & Gandhi (2011) [29]	Recognition of distorted broken characters	Shape and Line tracing method	Completely recognized
Pilevar & Pilevar (2011) [30]	Recognition of Broken and Touching Persian characters	Eleven Connected chain Method and template matching	93%.
Sulem & Sigelle (2008) [31]	Recognition of degraded characters	Dynamic Bayesian network	96%
Yu & Yan (2001) [32]	Recognition and reconstructing broken handwritten digits	Optimized nearest neighbor .	98.1%. 99.6% (After reconstruction)

3. METHODOLOGY

The objective of this work is to (i) preprocess the image and extract the feature of the character and then (ii) develop a Feed-forward Neural Network (FFNN) classifier, which is constructed with twenty five features for the recognition of Broken Bengali Characters (BBC). It is important to note that due to the fast processing and higher adaption, the FFNN has widely been used as a pattern recognizer (classifier) [33-40]. To accomplish the objective, following steps are adopted (see Fig.1).

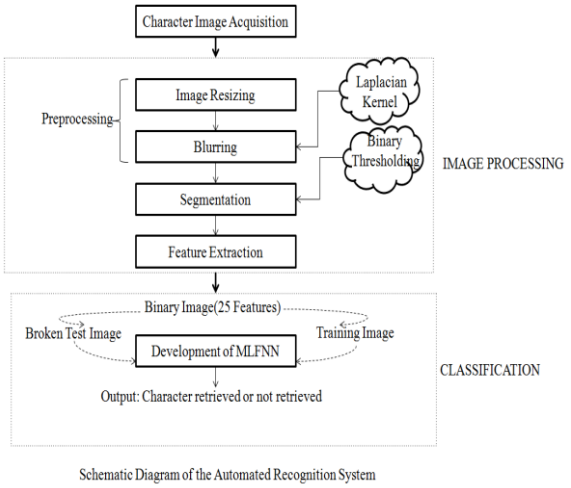


Fig.1. The flow diagram of the methodology adopted.

3.1 Character Image Acquisition

The character image (a sample set of 1000 isolated basic handwritten characters) has been collected from ISI image database of handwritten characters .These images are in JPEG format. There are 50 characters of 20 samples each. The characters are of different size, shape, variation and thickness etc.

3.2 Preprocessing

The character images are normalized to 54 × 54 pixels. Then blurring operation is performed on the images to soften the edges of the image. Laplacian Kernel (of size-5 × 5) has been used for this purpose. It removes the noise too.

3.3 Segmentation

The same kernel which is used for blurring and noise removal is also used here for character segmentation using Otsu's Thresholding method. One segmented image is shown in Fig.2, which is an Applet view.

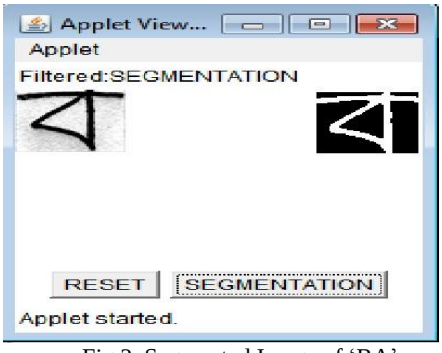


Fig.2. Segmented Image of 'BA'.

The following algorithm is applied on the image:

For each image- 50 × 50 pixels, steps i and ii are performed

i. Mean Variance is calculated and is being used as the threshold.

ii. Pixels having greater value than the threshold are considered as the foreground i.e. '1', else it is considered as background i.e. '0'.

3.4 Feature Extraction

The following algorithm is applied on the segmented image to extract the feature vectors:
The matrix is divided into 10x10 pixels.

i. For each matrix consisting of zeros and ones respectively, the number of 1's are searched.
ii. If the number of 1's exceeds 10, then the matrix is assigned with a value of 1.
iii. Steps i and ii are repeated respectively and a 5 × 5 matrix is obtained.

After processing of character image, a 5 × 5 matrix consisting of 25 binary features are obtained. These 25 features serve as the input for the FFNN.

3.5 Classification

In this work, a Multi layered Feed-forward Neural Network Classifier (FFNN) are designed (using Dev C++), which consists of 25 input nodes for the input layer,12 hidden nodes for single hidden layer and one output node. The generic topology is shown in Fig.3.

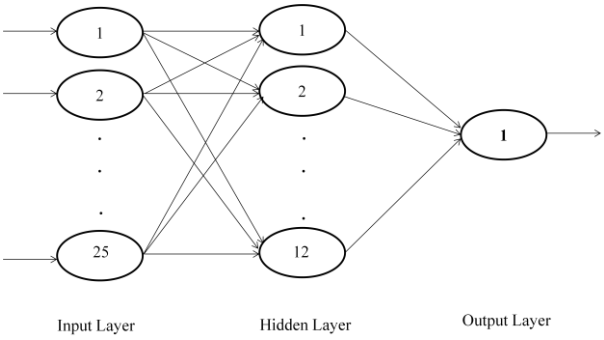


Fig 3. An MLFNC Topology (generic)

The transfer function used in the hidden layer is Gaussian Function followed by Sigmoid Transfer Function. The Gaussian would help accommodating maximum information while the later would facilitate the normalization of the output values between 0 and 1. The network is trained with 10 well defined isolated characters and tested with 9 broken characters, for each single character 'A' , 'AA', 'TA', 'KA' and 'BA', respectively. These five characters are considered, because there is a resemblance between 'A', 'AA' and 'TA' and 'KA' and 'BA' respectively. Authors wanted to see whether the classifier is able to distinguish these similar looking broken characters.

4. RESULTS AND DISCUSSIONS

The characters obtained from the ISI database are of different shapes, sizes and thicknesses as shown in Fig.4. These data has been used for training of network.

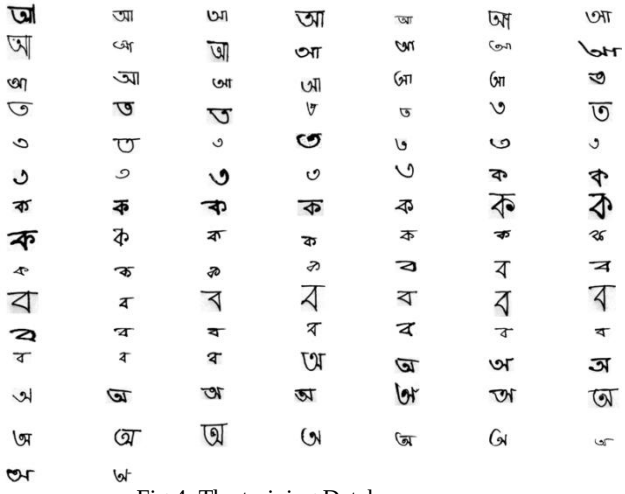


Fig.4. The training Database.

For testing purpose the character data has been broken or distorted up to 90% by considering the pixel ratio. A character "A" distorted up to 90% is shown in Fig 5.

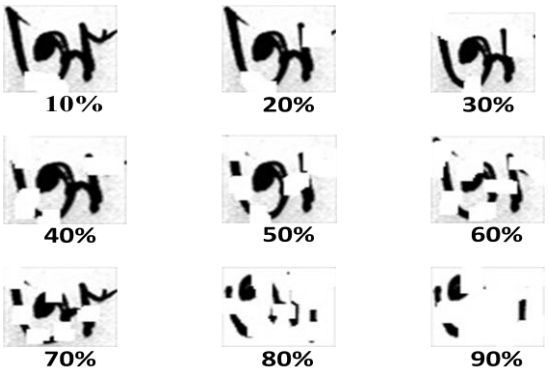


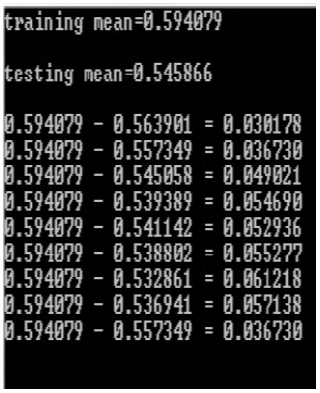
Fig.5. A sample of the Testing data

The objective of the network training is to calculate the mismatch between the well defined characters and the distorted, broken characters. To achieve this and retrieve the distorted characters respectively, the following algorithm has been implemented.

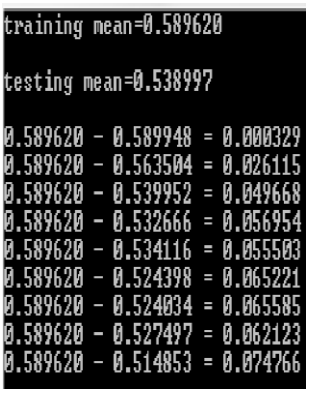
i. The outputs for each pattern of the well defined training data for every character are calculated using feed-forward algorithm.
ii. For every single character, mean is calculated for all the patterns.
iii. The output of all the patterns for the distorted characters (distorted with different %) are calculated.
iv. The mismatch is measured as a difference of the mean of the character with the output of its distorted character.
v. The Gaussian function has been used in the hidden layer to obtain the threshold up to which a distorted character can be retrieved.

The training averages for Bengali character A, AA, TA, KA and BA are calculated as 0.594079, 0.589620, 0.539009, 0.568363 and 0.538192, respectively as shown in Fig.6. The testing averages are 0.545866, 0.538997, 0.522724, 0.519983 and 0.515112, respectively. The results show that there is minimum mismatch between the averages of training and

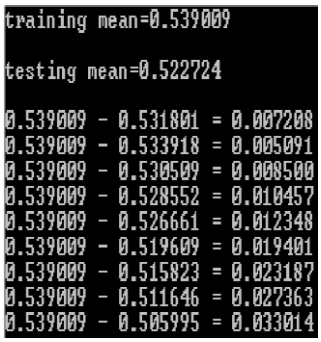
testing, respectively. These are 0.048213, 0.050623, 0.016285, 0.04838, and 0.02308, respectively.



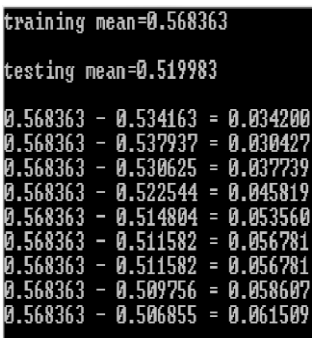
A



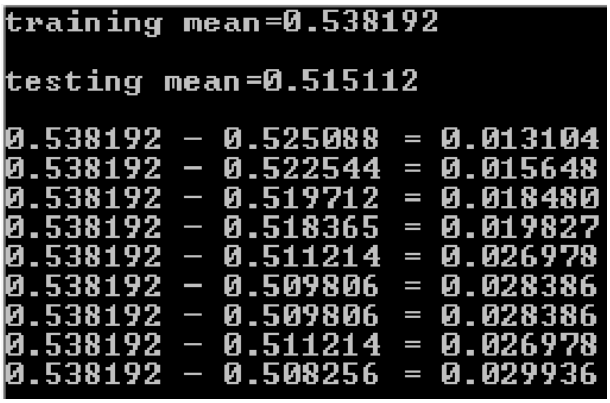
AA



TA



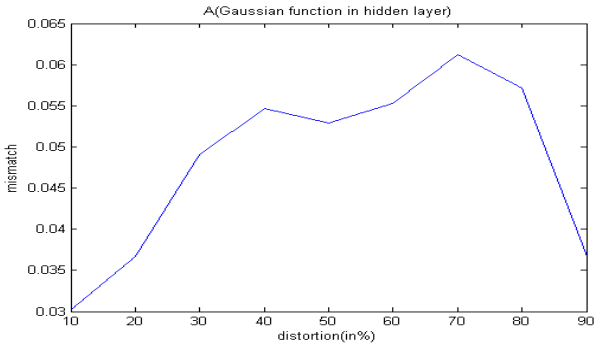
KA



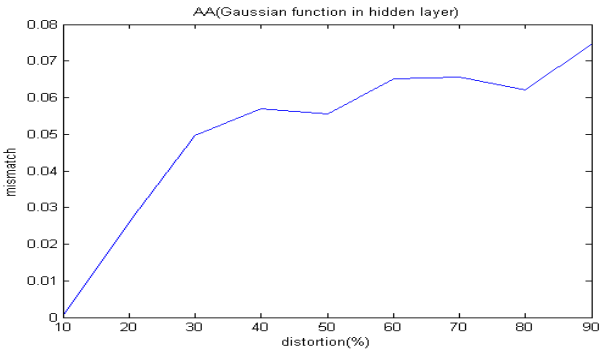
BA(Gaussian)

Fig. 6. Results of Training and Testing

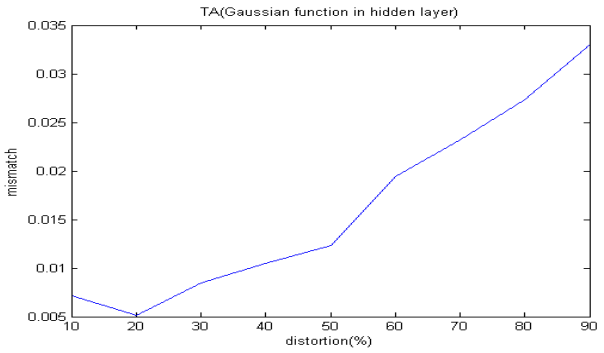
The mismatch shown in Fig. 6 is calculated as the Training Mean minus the output of the distorted character. The experiment shows that the mismatch increases as the image distortion increases. Up to 70% the mismatch increases but after that the mismatch decreases which is abnormal as shown in Fig.7.



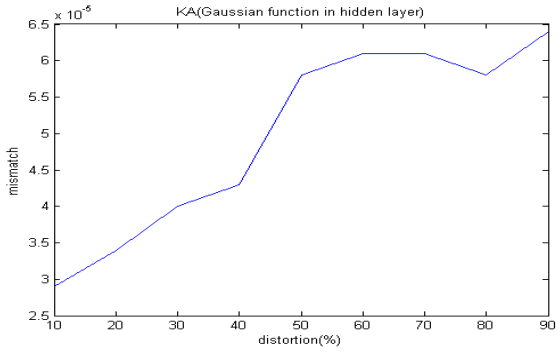
A



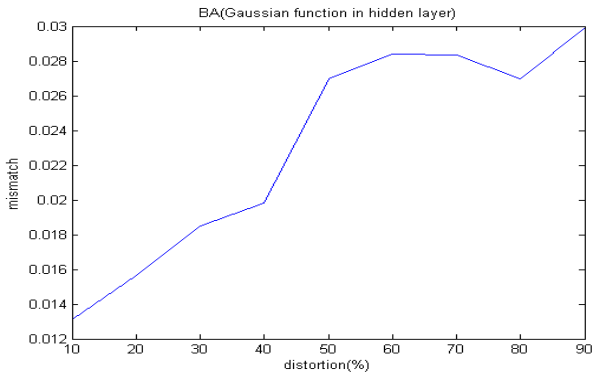
AA



TA



KA



BA

Fig.7 Plot of mismatch versus distortion during Testing.

This 70% mismatch is hence considered as a threshold for the distortion. Any character, which is distorted more than 70%, could not be retrieved. The characters with distortions less than and equal to 70% are considered as the test case and this cases we have used for the basis of output calculation.

5. CONCLUSIONS

The proposed character recognition system is tested for five Bengali Broken Characters (BBC) and it is found that the characters distorted up to 70% could be retrieved successfully. Experimental results seem to be encouraging considering the size of the training data and the varying font size and font style. The system is able to handle training data of smaller sample size which is important to make it efficient. The application of this system is that, it may be used to build digital library by retrieving characters from ancient historical documents.
The scope of the work is not limited to a particular script and it can be extended to any other script. Our future work is to test with more number of characters. The main aim is to develop a full-fledged, Automated Recognition System for the old documents whose contents are harder to retrieve.

6. REFERENCES

[1] S. Chattopadhyay "A Prototype Depression Screening Tool for Rural Healthcare: A Step towards e-Health Informatics", Journal of Medical Imaging and Health Informatics Vol. 2, No. 3, pp 244-249, 2012.

[2] S. Satapathy, S. Chattopadhyay "Observation-Prevention of Cardiac Risk Factors: an Indian Study", Journal of Medical Imaging and Health Informatics Vol. 2, No. 2, pp 102-113, 2012.

[3] S. Chattopadhyay, R.M. Davis., D.D. Menezes, G. Singh, U.R. Acharya, T. Tamura "Application of Bayesian Classifier for the Diagnosis of Dental Pain", Journal of Medical Systems Vol. 36, pp. 1425-1439, 2012

[4] S. Chattopadhyay, F. A. Rabhi, U.R.Acharya, R. Joshi, R. Gajendran "An Approach to Model Right Iliac Fossa Pain using Pain-only-parameters for Screening of Acute Appendicitis", Journal of Medical Systems Vol. 36, pp. 1491-1502, 2012.

[5] M. Yetirajam, M.R. Nayak, S. Chattopadhyay "Design of Fuzzy Logic Controller to Enhance the Operation of Cricket Bowling Machine", International Journal of Computer Technology and Applications, Vol. 3, No. 5, pp. 1662-1666, 2012.

[6] 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 Vol. 3, No. 8, pp. 21-25, 2012.

[7] S. Chattopadhyay S., R. Saurabh, L. Land, U.R. Acharya "Studying Infant Mortality Rate: A Data Mining Approach", Health and Technology Vol. 1, No. 1, pp. 25-34, 2011.

[8] S. Chattopadhyay, F. Daneshgar "A Study on Suicidal Risks in Psychiatric Adults". International Journal of Biomedical Engineering and Technology Vol. 5, No. 4, pp. 390-408, 2011.

[9] P. Ray, S. Chattopadhyay "Fuzzy Awareness Model for Disaster Situations". Intelligent Decision Technologies: an international journal [Special Issue on Intelligent Decision Making in Dynamic Environments: Methods, Architectures and Applications] Vol. 3, No. 1, pp. 75-82, 2009.

[10] S. Chattopadhyay 'A Study on Suicidal Risk Analysis'. In proceedings of 9th IEEE Intl. Conf. on e-Health Networking, Applications and Service (HEALTHCOM 2007), pp. 74-79, Taipei, Taiwan, 2007.

[11] J. Rocha, T. Pravlidis,"A Shape analysis model with application to a character recognition system," IEEE Trans. on Pattern Analysis and Machine Intelligence , vol. 16, pp. 393-404, 1994.

[12] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[13] M. K. Shukla, Dr. H. Banka, "A Study of Different Kinds of Degradation in Printed Bengali Script", International Journal of Advanced Computer Engineering and Architecture Vol.2, pp. 143-151, 2012.

[14] C. B. Bose and S-S Kuo, "Connected and degraded text recognition using Hidden Markov Model", Pattern Recognition, Vol. 27, No. 10, pp. 1345-1363, 1994.

[15] S-W Lee, D-J. Lee, and H-S Park, "A New Methodology for Gray-Scale Character Segmentation and Recognition", IEEE Transactions on PAMI, Vol. 18, No. 10, pp. 1045-1050, 1996.

[16] S. Kahan, T. Pavlidis, and H. S. Baird, "On the recognition of printed characters of any font and size", IEEE Transactions on PAMI, Vol. 9, No. 2, pp. 274 – 288, 1987.

[17] S. Tsujimoto and H. Asada, "Resolving Ambiguity in Segmenting Touching Characters", Ist Int. Conference on Document Analysis and Recognition, Saint-Malo, France, pp. 701-709, Oct. 1991.

[18] R. G. Casey and G. Nagy, "Recursive Segmentation and Classification of Composite Character Patterns", Proc. 6th Int. Conf. on Pattern Recognition, Munich, Germany, pp. 1023-1026, 1982.

[19] T. Hong, "Degraded text recognition using visual and linguistic context", Ph.D. Thesis, Computer Science Department of SUNY at Buffalo, 1995.

[20] Y. Lu, "Machine printed character segmentation – An overview", Pattern Recognition, Vol. 28, No. 1, pp. 67-80, 1995.

[21] B.B. Chaudhuri, U. Pal, and M. Mitra, "Automatic Recognition of Printed Oriya Script", ICDAR, pp.795-799, 2001.

[22] M. K. Jindal, G. S. Lehal, and R. K. Sharma, "Segmentation problems and solutions in printed Degraded Gurmukhi Script", International Journal of Signal Processing, Vol. 2, No. 4, pp. 258-267, 2005.

[23] G. S. Lehal and C. Singh, "Text segmentation of machine printed Gurmukhi script", Document Recognition and Retrieval VIII, Proceedings SPIE, USA, Vol. 4307, pp. 223-231, 2001.

[24] G. S. Lehal and C. Singh, "A technique for segmentation of Gurmukhi script", Computer Analysis of Images and Patterns, Proceedings CAIP 2001, Warsaw, Poland, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2127, pp. 191-200, 2001.

[25] V. Bansal and R.M.K. Sinha, "Segmentation of touching and fused Devanagari characters", Pattern Recognition, Vol. 35, No. 4, pp. 875-893, 2002.

[26] Yetirajam M., Nayak M.R., Chattopadhyay S. "Recognition and Classification of Broken Characters using Feed Forward Neural Network to Enhance an OCR Solution", International Journal of Advanced Research in Computer Engineering & Technology, Vol. 1, No. 8, pp. 11-15, 2012.

[27] Sahu S.K, Sahani S., Jena P.K,, Chattopadhyay S. "Fingerprint Identification System using Tree based Matching", International Journal of Computer Applications Vol. 53, No. 10, pp. 11-16.

[28] 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.

[29] R.I. Gandhi, N.F. Abubacker, "An Extended method for recognition of broken typewritten characters special reference to tamil scripts," in Proc. IEEE Conf., Open Systems (ICOS), pp.214-219, 2011.

[30] A.H. Pilevar, M.T. Pilevar, "Broken and Touching Characters Recognition in Persian Text Documents", world Applied Sciences Journal Vol. 13, No. 6, pp. 1459-1464, 2011.

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

[32] 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.

[33] 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.

[34] T. Dash, T. Nayak, S. Chattopadhyay "Offline Handwritten Signature Verification using Associative Memory Net". International Journal of Advanced Research in Computer Engineering & Technology, Vol. 1, No. 4, pp. 370-374, 2012.

[35] Dash T., Nayak T., Chattopadhyay S. "Handwritten Signature Verification (Offline) using Neural Network Approaches: A Comparative Study", *International Journal of Computer Applications* Vol. 57, No. 7, pp. 33-41.

[36] Dash T., Nayak T., Chattopadhyay S. "Offline Verification of Hand Written Signature Using Adaptive Resonance Theory Net (Type-1)". In the proceedings of the *4th International Conference on Electronic Computer Technology (ICECT-2012 Vol-2) Kanyakumari, India* (6-8 April'12). Editor: Yuan Li, pp. 205-210.

[37] S. S. Behera, S. Bhanja Choudhuri, S. Chattopadhyay "A Comparative Study on Neural Net Classifier Optimizations". International Journal of Advanced Engineering and Technology, Vol. 4, No. 2, pp. 179-187, 2012.

[38] S. Chattopadhyay, P. Kaur, F. A. Rabhi, U.R. Acharya "Neural Network Approaches to Grade Adult Depression", Journal of Medical Systems, Vol. 36, No. 5, pp. 2803-2815, 2012.

[39] A. Dasari, N.B. Hui, S. Chattopadhyay "A Neuro-fuzzy System for Modeling the Depression Data", International Journal of Computer Applications, Vol. 53, No. 6, pp. 1-6, 2012.

[40] S. S. Behera, S. Chattopadhyay "A Comparative Study of Back Propagation and Simulated Annealing Algorithms for Neural Net Classifier Optimization", International Conference on Modelling Optimization and Computing (ICMOC-2012) Nagercoli, Tamilnadu India Vol. 38, pp. 448-455, (10-11 April 2012).

[41] K. Krishna, A. Goyal, S. Chattopadhyay "Non-correlated Character Recognition using Hopfield Network: A Study". In the proceedings of International Conference on Computer and Computational Intelligence (ICCCI-2011) (Ed. Yi Xie) pp. 385-389 Bangkok, Thailand (2-4th December) 2011.

[42] S. Chattopadhyay, P. Kaur, F.A. Rabh, U.R. Acharya "An Automated System to Diagnose the Severity of Adult Depression". In the proceedings of 2nd International Conference on Emerging Applications of Information Technology (CSI EAIT-2011), 19-20th February Kolkata India, (Editors - Jana D., and Pal P.), pp. 121-124. Publishers: IEEE Computer Society and Conference Publishing Services IEEE Xplore.