Genetically enhanced Meetei Mayek Unicode recognition using Neural Network

[1]Wahengbam Kanan Kumar,[2]Chetan Das Gupta,[3]Jyoti Bordoloi,[4]Yogesh Kumar Sharma

Abstract— This paper exhibits a new computing approach for implementing and simulating recognition of genetically enhanced handwritten Meetei Mayek using neural network. Around 2550 samples i.e. 75 samples for each of the 34 letters were collected from people belonging to different age group. Each of the letters needs filtering for removal of noise, pre-processing must follow before applying them to neural network. Filtering is done by using Adaptive Noise filter, while segmentation is done by using GAFCM algorithm, now each and every letter is pre-processed properly. GAFCM segmentation was proven better for MRI image processing.The same technique is applied to the letters to achieve higher accuracy. Recognition is done using Multilayer Feed forward neural network with back propagation (BP) learning. The machine learns the different patterns present in the database and the output fetched every time is tabulated which are discussed on the following segments of the paper.

Index Terms— Meetei Mayek Script, GAFCM, Neural Networks, back propagation training, recognition

—————————— ◆ ——————————

1 INTRODUCTION

Pattern recognition is a method of giving name to a given input value. Pattern recognition is a classification, which assigns each input to one of a given set of classes. It performs the task of matching all the possible inputs taking into account their statistical variation. A common instance of pattern recognition algorithm is searching for patterns of some sort in textual data. Applications may include vehicle number plate recognition, bank cheque processing, automatic reading of area code and address from letter, etc.

The Meetei Mayek is an abugida that was used for the Meitei language; it is one of the official languages of the Indian state of Manipur and is primarily spoken in the valley region. It is the first language of the ethnic group Meitei. This script contains 18 original letters known as Eeyek Eepee, 9 additional known as Lom Eeyek,8 letters with short ending known as Lonsum Eeyek,8 vowel signs known as Cheitap Eeyek,3 punctuation marks known as Khudam Eeyek and 10 numerical characters known as Cheishing Eeyek.

———————————————————

- [1]**Wahengbam Kanan Kumar**,M.Tech scholar, Maharishi Markandeshwar University, Ambala, India, E-mail: wahengbam.kanankumar@gmail.com
- [2]**Chetan Das Gupta**,M.Tech scholar, Maharishi Markandeshwar University, Ambala, India, E-mail: chetandasgupta@gmail.com
- [3]**Jyoti Bordoloi**,M.Tech scholar, Maharishi Markandeshwar University, Ambala, India, E-mail: er.jyoti101@gmail.com@gmail.com
- [4]**Yogesh Kumar Sharma**,M.Tech scholar, Maharishi Markandeshwar University, Ambala, India, E-mail: er.yogeshsharma007@gmail.com

research work have been found to be presented paving the way for implementing better Optical character recognition technique. All the original figures of the Manipuri alphabets are derived from human anatomy and named accordingly. Very few research work have been found to be presented paving the way for implementing better Optical character recognition technique. A survey of Indian script recognition can be found in the literature [6]. The 34 letters that is being considered for the paper is shown in the figure below.

Neural Networks are widely applied to pattern recognition areas [7],[8], [10]. Neural Networks can be trained and then tested on handwritten digits. This paper shows the feed forward neural network along with back propagation learning approach for the handwritten digit recognition.

Fig.1The 34 Meitei Mayek characters used in the paper.

All the original figures of the Manipuri alphabets are derived from human anatomy and named accordingly. Very few

2. GAFCM Segmentation [1]:

GAFCM Algorithm:

i) ENCODING: Each chromosome represents a solution which is a sequence of cluster centroids. In N-dimensional space, each centers are mapped to N consecutive genes in the chromosome. Each of the genes is an integer representing an intensity value. 2 clusters are used in this paper so that binary image is obtained.

ii) POPULATION INITIALISATION: Set chromosomes as vector containing the centroids of the clusters. In the implementation, population is set as 10 and the number of generations as 40.

iii) EVALUATION OF FITNESS FUNCTION: Using the roulette selection procedure, it is ensured that chromosomes with more fitness values have better chance to get selected. The fitness function is set as the inverse of the objective function used in FCM algorithm, i.e. Fitness function = $1/J_m$.

iv) CROSSOVER: The Crossover step recombines the bits (genes) of the two selected strings or chromosomes. Out of the two Crossover techniques i.e. Single and two points. Single point crossover operator is selected for the project.

MUTATION: The next operation mutation is performed in a bit-by-bit basis. Let $p_m = 0.01$, i.e. we expect on an average 1% bit mutation. There exist altogether 10 chromosomes*5 bits/chromosomes = 50 bits in the whole population. 1% bit mutation thus means 50*0.01=0.5 bit mutation. Since every bit has an equal chance of mutation, we generate a random number in [0, 1] and if the generated number is $p_m < 0.01$, we select the chromosomes for mutation. So for each chromosome, the feasibility of the chromosome for mutation is checked. In order to determine the bit position of mutation, a random number in [0,n-1] is generated, where n is the word- length of the chromosomes. If random number generated is p, the pth bit of the selected chromosomes will be mutated. It ensures that the algorithm moves towards the global minima instead of getting stuck in local minima

3 NEURAL NETWORK

Neural networks are computationally efficient models inspired by the working of the neurons present in the human brain. Similarly a system of interconnected neurons can compute values by feeding information through the network. In a neural network for recognising handwriting, a set of input neurons will be activated by the pixels of an input image representing a letter or digit. Activations are then passed on, weighted and transformed by using some function determined by the network designer, to other neurons, etc., until an output neuron is activated which determines the character that was read. It consist of sets of adaptive weights that are tuned by a learning algorithm, it is also capable of approximating non-linear functions of their inputs. The adaptive weights are connection strengths between neurons that are activated during training and prediction. Back propagation is a systematic method of training multilayer

artificial neural networks. It is among the widely used method for supervised learning with a wide range of practical range application. In this paper the multilayer feed forward neural network with back propagation learning is used to recognise Meetei Mayek characters.
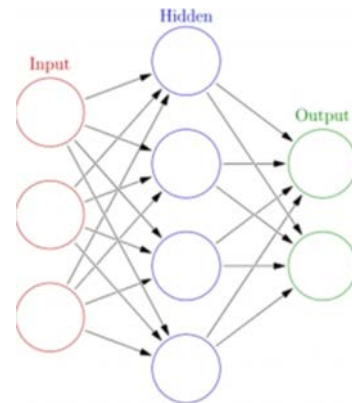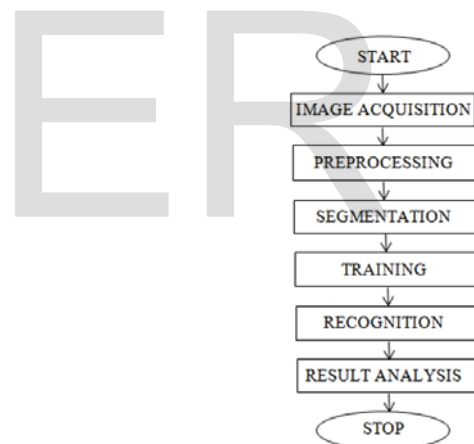

Fig. 2 Multilayer neural net- work structure

4 THE PROPOSED RECOGNITION MODEL

A flowchart showing the basic implementations is shown below:

a. **Image acquisition:** At the very beginning we have collected 2550 samples from person falling under different age group on a clear A4 paper by using a scanner. The picture below shows a sample of the handwritten Meetei Mayek characters
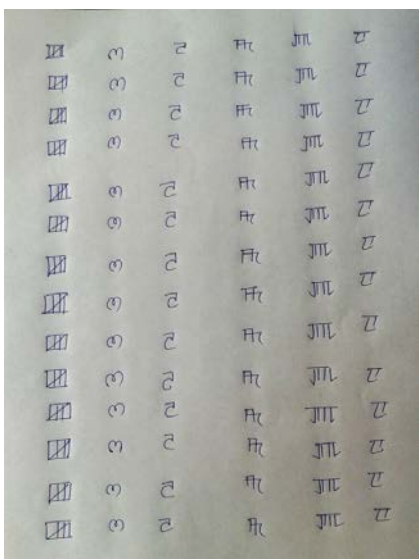
Figure 3. Scanned sample of Meetei mayek

b. **Pre-processing:** Pre-processing refers to the various operations performed onto the image so that the result is noise free, improved and is easy for a machine to learn. The pre-processing steps implemented in the paper is described as under:

- Image is changed to grayscale image.
- Thresholding is done to the image
- 2-D adaptive noise-removal filtration is performed on the input image.
- Line segmentation
- Character segmentation
- 140X140 pixels are used for all the letters so they have same no. of pixels.

c. **Segmentation:** It is the process of partitioning a digital image into multiple segments. The goal of which is to simplify or change the representation of an image into something that is more meaningful and easier to analyse. It is typically used to locate objects and boundaries (lines, curves, etc.) in images. Image segmentation is the process of assigning label to every pixel in an image such that pixels with the same label share certain visual characteristics.

Genetic Algorithm incorporating Fuzzy C Means (GAFCM) algorithm [1] is used to perform segmentation of the characters. The step-by-step procedure for GAFCM segmentation is shown on the section II of the paper.

Figure 4. Sample of GAFCM segmented image

d. **Neural network architecture:** The segmented images are fed into the feed forward neural network architecture with back propagation learning for the recognition process. The network consists of three layers to recognise each of the 34 characters that are being considered in the paper. All the simulation has been performed in the MATLAB environment by using back propagation neural network with Gradient descent with momentum and adaptive learning rate.

The output vector is composed of 34 neurons as it has to hold 34 elements. The digit 0 is to be represented by a 1 in the place where output is to be observed whereas 0 is seen in the rest of the neurons in the output vector. Training involves two passes, forward pass and backward pass.

Forward pass: The input signals propagate from the network input to the output.

Reverse pass: The calculated error signals propagates backwards through the network where they are used to adjust the weights. The output of one layer in weighted manner will be the input to the next layer. Traingdx function present in the MATLAB library is used for training the neural network, weights and bias values are adjusted according to the gradient descent and learning rate. Performance goal is 0.10000 and momentum constant is kept at 0.9. Sum squared error is used to measure the performance.

5. Simulation result:

The experimental results and observations shown in this section are retrieved from MATLAB environment. Out of the total 2550 samples taken 1700 were used for training, while 840 samples used for testing the performance of the technique mentioned above.

Table I shows the performance of the network, while Table II and Table III shows the recognition matrix.

Since pronunciation of the letters are difficult and long we have made some notations which is shown below:

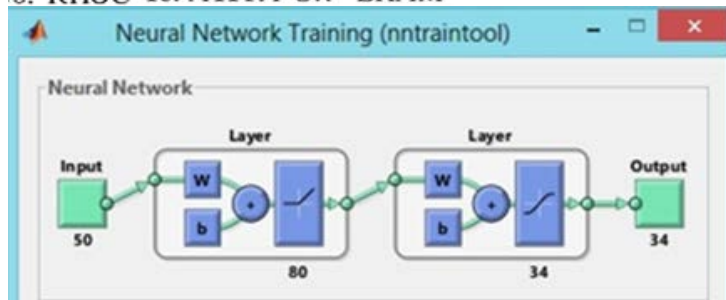|			
1. KOK	10. NGOU	19. GOK	28. KOK LONSUM
2. SAM	11. THOU	20. JHAM	29. LAI LONSUM
3. LAI	12. WAI	21. RAI	30. MIT LONSUM
4. MIT	13. YANG	22. BA	31. PA LONSUM
5. PA	14. HUK	23. JIL	32. NA LONSUM
6. NA	15. UN	24. DIL	33. TIL LONSUM
7. CHIL	16. EE	25. GHOU	34. NGOU LONSUM
8. TIL	17. PHAM	26. DHOU	
9. KHOU	18. ATIYA	27. BHAM	

Fig. 5 Neural network from the MATLAB environment;
Notation of the characters in numeral form

DIGITS	NO. OF TEST SAMPLES	TRUE RECOGNITION	FALSE RECOGNITION	PRECISION (%)
1	25	20	5	80
2	25	18	7	72
3	25	21	4	84
4	25	21	4	84
5	25	19	6	76
6	25	14	11	56
7	25	11	14	44
8	25	17	8	64
9	25	13	12	52
10	25	23	2	92
11	25	8	17	32
12	25	9	16	36
13	25	7	18	28
14	25	25	0	100
15	25	17	8	68
16	25	10	15	40
17	25	15	10	60
18	25	13	12	52
19	25	16	9	64
20	25	20	5	80
21	25	16	9	64
22	25	16	9	64
23	25	19	6	76
24	25	19	6	76
25	25	7	18	28
26	25	6	19	24
27	25	11	14	44
28	25	8	17	32
29	25	11	14	44
30	25	11	14	44
31	25	7	18	28
32	25	16	9	64
33	250	23	2	92
34	250	13	12	52

TABLE I: Performance evaluation

Where,

True recognition = Number of correct recognition

False recognition = Number of false recognition

Precision = Accuracy in percentage

1,2,……34 = 34 Meetei Mayek characters

|	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	20	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
2	0	18	0	1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
3	0	0	21	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
4	1	0	0	21	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
5	2	0	0	1	19	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
6	0	1	3	0		14	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0
7	1	0	1	1	2	0	11	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	1	2	0	0	0
8	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	1	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
9	0	0	0	3	0	0	0	0	13	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	2	0	1	0	2	0	0	0
10	1	0	0	0	1	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	0	1	0	0	0	0	0	0	2	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	1	9
12	0	0	0	2	1	0	1	2	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	1	2	5	0	0	0	1	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	8	0	4	11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	3	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	10	5	1	0	4	0	2	0	0	0	1	0	0	1	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	15	1	0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	4	13	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
19	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	2	16	1	0	0	1	1	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	5	0	0	0	0	0	0	0	0	0	0
21	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	1	0	1	0	0	0	0	0	2	0	0	1
22	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	1	0	17	0	2	0	0	0	0	0	0	0	1	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	4	19	0	0	0	0	0	0	0	0	0	0

24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	4	0	19	0	0	0	0	0	0	0	0	0	0
25	0	0	0	1	0	0	2	2	8	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	2	0	2	1	0	0	0	0	1	0	0	0	0	0	0	0	3	6	3	0	0	0	0	0	1	0	
27	0	0	0	0	0	1	0	1	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3	0	11	0	1	0	0	0	1	0	
28	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	8	0	1	2	6	2	1
29	0	0	0	0	0	0	0	0	6	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	2	0	2	0	11	0	0	0	0	0	
30	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	5	0	1	11	0	1	0
31	0	1	0	2	11	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7	0	0
32	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	0	0	0	0	0	1	1	16	1	1
33	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0
34	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	0	13

Table II. Recognition matrix for the first 17 characters.

Where, the characters on the left most column is compared with the images contained in the database. This table is known as Recognition Matrix as it indicates the number of times a particular character is recognised as itself or confused with the characters. In this matrix due to space limitation only 17 character has been shown for comparison, while the remaining 17 is shown in the next page.

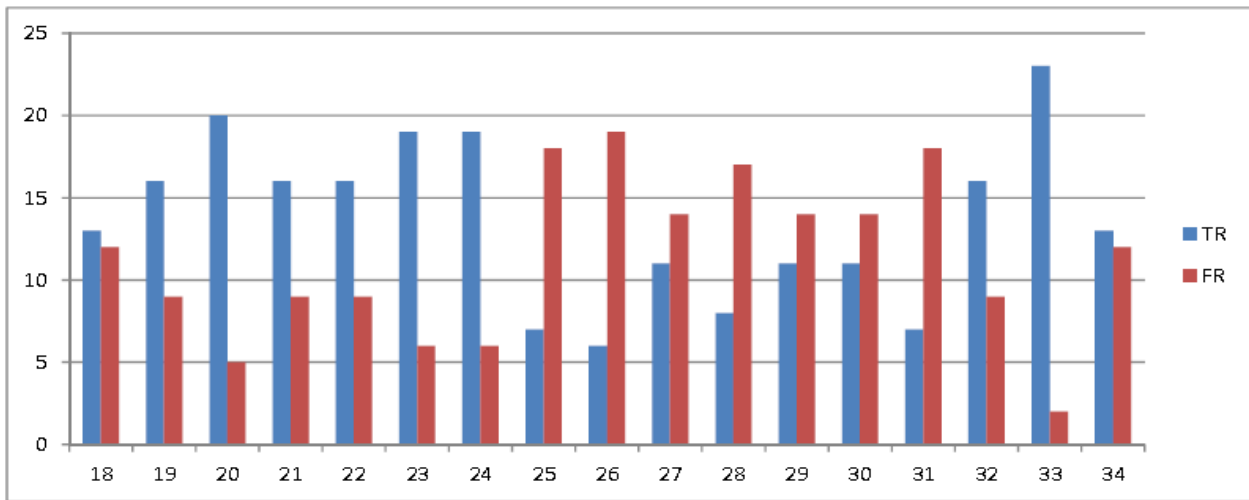(Bar chart — y-axis: 0 to 30; x-axis: characters 1 to 17; legend: ■ TR, ■ FR)
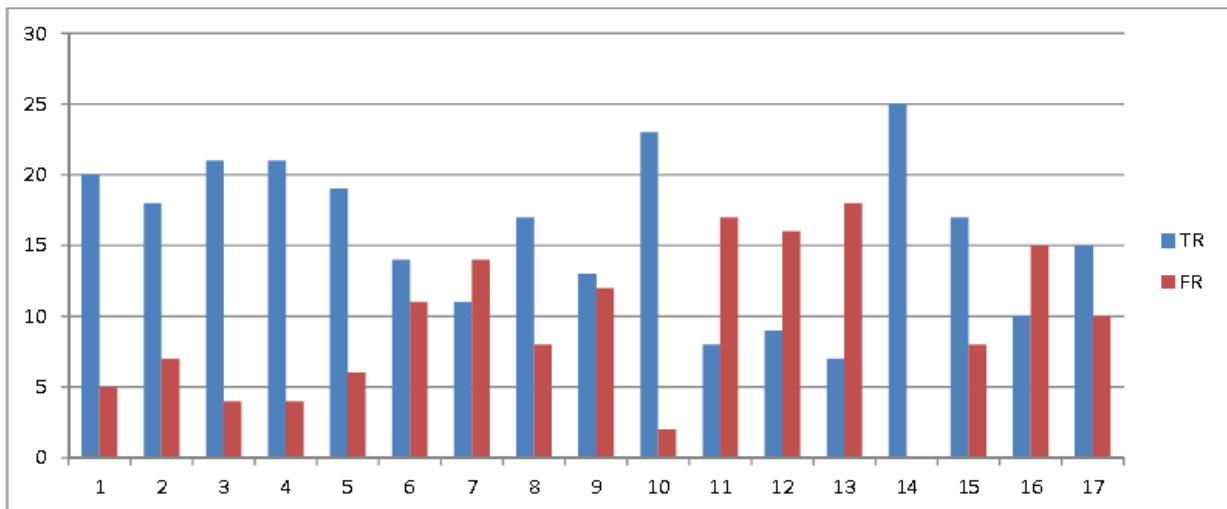
Figure 6. Graph showing True Recognition and False Recognition

6.   Conclusion

Through this paper a new method has been proposed that can segment and recognise handwritten Meetei Mayek script by using neural network approach. The overall accuracy that is achieved is about 54.47% which is good considering the fact that most of the characters look somewhat similar and the machine can be easily confused. This shows that GAFCM algorithm when combined with Artificial Neural Networks yields better pattern recognition result as compared that of ordinary neural network approach. Nevertheless accuracy for some of the samples is high if the characters are considered individually. The performance table also shows that that some characters have less accuracy and so in order to improve the overall accuracy better algorithm shall be implemented in the near future.

Reference

[1] Romesh Laishram, W.Kanan Kumar Singh, N.Ajit Kumar, Robindro.K, S.Jimriff, "MRI Brain Edge Detection Using GAFCM Segmentation and Canny Algorithm", International Journal of Advances in Electronics Engineering – IJAEE,volume 2 - Issue 3, ISSN:- 2278-215X, pp. 168-171,December 8,2012

[2] Rafael C. Gonzalez, Richard E.Woods, "Digital Image Processing", Pearson Education, Second Edition, ISBN 81-7758-168-6, 2005.

[3] Amit Konar, "Computational Intelligence Principles, Techniques and Applications", Springer edition, ISBN 3-540-20898-4 Springer Berlin Heidelberg, New York

[4] Wangkhemcha Chingtamlen, A short history of Kangleipak (Manipur) part-II, Kangleipak Historical & Cultural Research Centre, Sagolband Thangjam Leirak,Imphal,2007.

[5] Ng.Kangjia Mangang, Revival of a closed account, a brief history of kanglei script and the Birth of phoon (zero) in the world of arithmetic and astrology, Sanmahi Laining Amasung Punshiron Khupham (SalaiPunshipham), Lamshang,Imphal,2003

[6] U.Pal & B.B.Choudhuri,, "Indian script character recognition:a survey" *Pattern Recognition*, 37,2004, pp.1887 -1899

[7] Avani R.Vasant, G. R. Kulkarni, "Simulation and Modeling of Handwritten Gujarati Digits using Neural Network Approach, 2011 IEEE International Conference on Computational Intelligence and computing

[8] Arindam Basu, "Small-Signal Neural Models and Their Applications", IEEE transactions on biomedical circuits and systems,vol. 6, no. 1, February 2012.

[9] S N Sivanandam, S Sumathi, S N Deepa, "Introduction to Neural Networks using MATLAB 6.0",Tata McGraw-Hill, Eleventh reprint 2010,ISBN-13: 978-0-07-059112-7, ISBN-10: 0-07-059112-1

[10] Ankit Sharma, Dipti R Chaudhary "Character Recognition Using Neural Network", International Journal Of Engineering Trends And Technology (IJETTE). - Volume 4 Issue 4-April 2013.

[11] www.**unicode**.org/charts/PDF/UABC0.pdf