Recognition of Handwritten Character of Manipuri Script

Tangkeshwar Thokchom
Deptt. of Computer Science, Manipur University, Imphal, India
Email: tangke_th@hotmail.com
P.K.Bansal
Malout Institute of Management & Information Technology (MIMIT), Punjab, India
Email: pkbmimit@yahoo.com
Renu Vig
University Institute of Engineering & Technology, Punjab University, Punjab, India
Email: renuvig@hotmail.com
Seema Bawa
Computer Science & Engineering Department, Thapar University, Punjab, India
Email: sbawa@thapar.edu

***Abstract*— In this paper a backpropagation neural network based handwritten characters (Mapum Mayek) recognition system of Manipuri Script is investigated. This paper presents various steps involved in the recognition process. It begins with thresholding of gray level image into binarised image, then from the binarised image the character pattern is segmented using connected component analysis and from the resized character matrix, its probabilistic features and fuzzy features are extracted. Using these features the network is trained and recognition tests are performed. Experiments indicate that the proposed recognition system performs well with the combined features and is robust to the writing variations that exist between persons and for a single person at different instances, thus being promising for user independent character recognition.**

***Index Terms*— Binarisation; Segmentation; Feature extraction; Neural network; Fuzzy feature; Recognition**

I. Introduction

For humans, reading a character seems a trivial task, but this is not so for a computer. Writing has been the most natural mode of collecting, storing, and transmitting information through the centuries now serves not only for communication among humans but also serves for communication of humans and machines. Off-line handwritten charcter recognition refers to the process of recognizing the symbol that have been scanned from a surface such as a sheet of paper and are stored digitally in gray scale format. It means transforming a language represented in its spatial form of graphical marks into its symbolic representation [1].

The off-line recognition for English is dedicated to bank cheque processing, mail sorting, reading of commercial forms, etc, while the on-line recognition is mainly dedicated to pen computing industry and security domains such as signature verification and author authentication. A public domain off-line handwriting recognition system developed by the National Institute of Standards and Technology (NIST) recognizes the handprint written on Handwriting Sample Form as distributed on the CD-ROM, NIST Special Database 19[2]. A few reports have appeared for isolated hand-written characters and numerals of some Indian languages[3].

Manipuri, also called Meeteilon [4,5], Meiteiron and Meithei [6] in linguistic literature, is the official language of the State of Manipur, India and is primarily spoken in the valley region of the State. It is the mother tongue i.e., the first language of the ethnic group Meetei. However, apart from the Hindu Meiteis and the Meetei following the traditional religion of Sanamahi, Meitei Pangals i.e., Manipuri Muslims also speak Manipuri as their mother tongue.

Manipuri is a tonal language of Tibeto-Burman language family. This script contains Iyek Ipee/Mapung Iyek, which have 27 alphabets (18 original plus 9 letters called Lom Iyek, derived from original 18 alphabets), Lonsum Iyek (8 letters), Cheitek Iyek (8 symbols), Khudam Iyek (3 symbols), Cheishing Iyek (10 numeral figures). In addition to these there are 6 vowel letters. The basic character may appear only as the main character of a word and it may be modified using one of the extended symbols (vowel modifiers) to produce the required vocal sound. All the original figures of the Manipuri alphabets are drawn, winded and wreathed from human anatomy and accordingly, the alphabetical names are the names of the different parts of the same where the characters are winded and drawn from [5]. To the best of our knowledge, research in Manipuri script recognition has not yet been widely introduced to the research community while much research on other scripts of different languages has been published and introduced internationally. As the characters of other Indian script are different in their styles of writing, a system adjusted

Manuscript received January 18, 2010; revised June 1, 2010; accepted July 1, 2010.

doi:10.4304/jcp.5.10.1570-1574

for automatic recognition of one script might not perform well for the other one.

This paper is organized as follows: Section II presents the overview of proposed method of the handwritten character recognition system of Manipuri Script. In Section III, the processes of recognition of isolated characters of the input image file are presented. Section IV presents the experimental results of the proposed system. Section V lists some of the remaining challenges in the handwritten Manipuri script recognition.

II. PROPOSED METHOD

There are many steps that need to be taken before handwritten characters on a page can be recognized by a computer. The stages of the proposed technique are as follows (shown in Fig. 1):

1. Scanning, 2. Preprocessing and Binarisation, 3. Segmentation 4. Normalisation and Feature extraction 5. Recognition/ Classification techniques

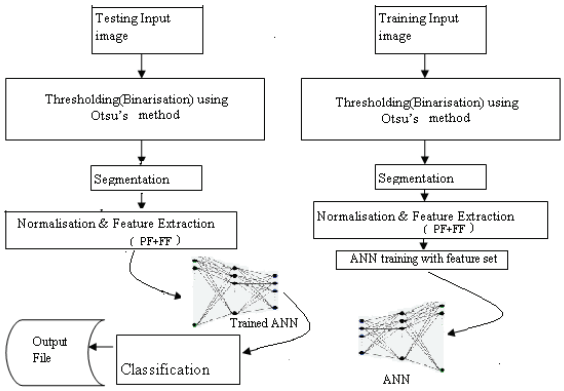

Figure 1. Overview of proposed method

A. Background Noise Removal

Every scanned image is associated with background noise and other noise which creeps is due to errors in scanning.

It is useful to be able to separate out the regions of the image corresponding to objects in which we are interested from the regions of the image that correspond to background. Thresholding often provides an easy and convenient way to perform this segmentation. Otsu's method of thresholding is performed as a preprocessing step to remove the background noise from the picture prior to extraction of characters and recognition of text. This method of thresholding involves in choosing the threshold to minimize the intraclass variance of the thresholded black and white pixels. After thresholding, the image is converted to a 1 bit binary image, thus making it simpler to carry out further operations as the image is reduced to a 2-d matrix of 1s and 0s.Fig.2 shows a sample input handwritten character image. Fig.3 shows the binarised image after thresholding step using Otsu's method.

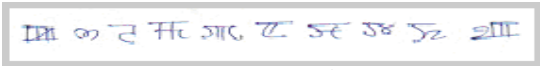

Figure 2. Sample characters of an Input Image

Figure 3 Thresholded image

B. Character Segmentation

From the thresholded image, edges of the objects are found using Sobel method using structure element of size 3. The edges are those points where the gradient of the image is maximum. Fig.4. shows edge detected image. Morphological operation such as image dilation is shown in Fig.5. Image dilation is used for adding pixels to the boundaries of objects in an image using structure element of size 3 so that the operation thickens objects in a binary image. The rule used for the dilation operation is that the value of the output pixel is the maximum value of all the pixels in the input pixel's neighborhood. In a binary image, the morphological dilation function sets the value of the output pixel to 1 because one of the elements in the neighborhood defined by the structuring element is 1.Filling operation on the image is performed as shown in Fig.6. This operation fills holes in the binary image. A hole is a set of background pixels that cannot be reached by filling in the background from the edge of the image. Then connected component analysis are performed on the filled image and basic properties such as area, centroid and bounding boxes of each object are found as shown in Fig. 7 and all the characters are extracted. The extracted characters are cropped to the edges and then re-scaling of the size of characters to 60x80 using nearest-neighbor interpolation are performed.

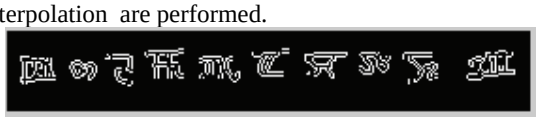

Figure 4. Edge detected image

Figure 5. Dilated image

Figure 6. Filled image

Figure 7. Located characters with bounding boxes

C. Feature Extraction

The features that we extract from the character image are: Probabilistic features and fuzzy features.

Probabilistic features: Two signatures are used to extract the features of the characters. Signature 1 allows the

characters to be rotated and Signature 2 gives a decent amount of scalability to the characters.

For finding the procedure for the signatures of a character, the algorithms are as follows:

To find Signature 1

-Find the means of the X and Y data

-Take these means and complete a matrix of distances

-Normalize

-Bin these distances using histogram

In practice, 21 features are extracted from the rescaled character image for signature 1.

To find Signature 2

-Bin X location data

-Bin Y location data

-Normalize

In practice, 10 features are extracted from the re-scaled image for signature 2. Thus a total of 31 probabilistic features are extracted.

Fuzzy features: The feature extractor breaks down the character image of 60x80 to 6x8 matrices by finding average value in each 10 by 10 blocks. Then compute the fuzzy values of the negative image of the figure where the input range is 0 for black to 1 for white and the value in between thereby giving 48 inputs for the network.

D. ANN Training

Three different feature sets i.e., probabilistic features(31), fuzzy features(48) and combination of both features giving a total of 79 features(31+48) are used to train the two layers feed forward backpropagation neural network separately for the purpose of comparative study. Back-propagation networks provide a very effective method for performing supervised nonlinear classification [4]. As is common in pattern classification problems, the efficiency of the network is strongly affected by the quality of the input features used to train the network. In back-propagation networks, the number of weights in the fully connected network increases as the product of the input neurodes and the hidden neurodes plus the product of the hidden neurodes and the output neurodes.

The number of neuron outputs is 27. Multiple sets of 27 character training pairs were used to train the network. Both the first and second layer has ten logsig neurons each. The traingdx gradient descent momentum and an adaptive learning rate network training function are used. The number of hidden units is 10. The network is trained with 0.01 learning rate, 0.1 performance goal, 0.9 Momentum constant.

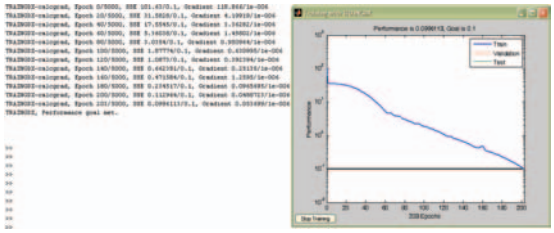

Figure 8. Typical training graphs for characters

III. Recognition of Segmented Characters

As shown in Fig.1, the input file containing the deskewed text lines are thresholded and the lines are segmented. From the extracted text line, the segmentation process follows. Using the same connected component analysis technique; characters are segmented and then resized the same to 60x80 using nearest-neighbor interpolation method.

Each feature set are tested separately for recognition using the previously trained weights of the ANN. The recognized character is then coded for its equivalent ASCII code for the Manipuri script. The process is repeated for all the characters in the extracted line. Then the whole process is repeated for any existing remaining text line in the input file.

IV. Experimental Results

Our data of Manipuri (Meetei/Kanglei) handwritten character samples, having 2 pages of 27 basic characters each, were collected from 11 persons including the students and faculty members of the Department of Computer Science, Manipur University, India. We have used 594 sample images to evaluate the proposed system. The data is divided into two parts.

Training Dataset: The training dataset contains the amples used for learning by the neural network. Out of 594 sample images, 459 character samples are used as training dataset for ANN.

Testing Dataset: From 594 samples, 135 samples are used as testing dataset. For the purpose of assessment/evaluation of the generalization of the system to unseen data, this test dataset is used.

Epoch based backpropagation training was conducted using the training set for different feature sets separately The network was permitted to train for 5000 epochs. The recognition rates are shown in Table 1 and Table 2 for three feature extraction methods: Probabilistic features (PF), fuzzy features (FF) and combination of both features (PF+FF).

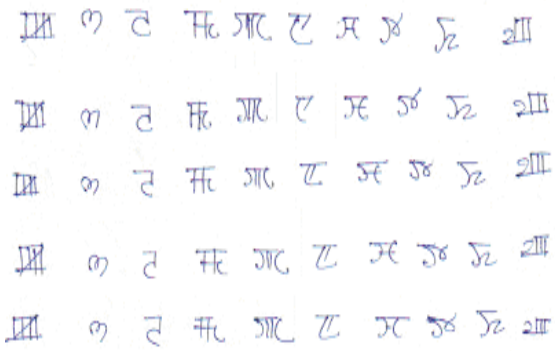

Figure 9. Sample characters from the Dataset.

Figure 10. Output of recognized characters

TABLE I.
PARAMETERS FOR COMPUTING RECOGNITION RATES

Method	No.of Samples	No. of features used	No. of Training samples
PF	594	31	459
FF	594	48	459
PF+FF	594	79	459

TABLE II.
RECOGNITION RATES

Method	No.of test samples	Recognition Rate
PF	135	85.92%
FF	135	88.14%
PF+FF	135	90.3%

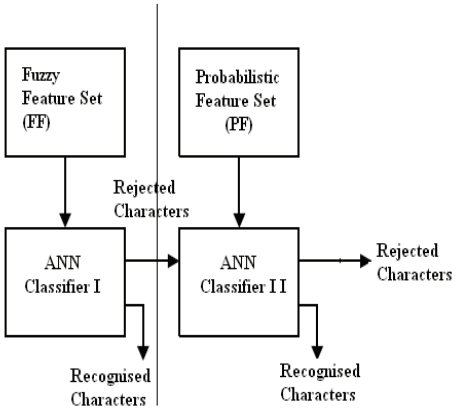

Figure 11 Cascade Recognition system

The accuracy of the proposed 27 class classifier neural network is 90.3% recognition with the combination of probabilistic and fuzzy features while the other two feature extraction methods show low accuracy. The experimental results demonstrate the network's ability to generalize for truly writer-independent recognition. Fig.9 shows some sample characters from the dataset. Fig.10 shows one sample output in ASCII fonts of characters of Manipuri script of the input image.

V. CONCLUSION AND FUTURE WORK

This study documents the strategies, methods and results used and gained from training a back-propagation network with probabilistic features, fuzzy features and combination of both features for character recognition. The experimental results show that the choice of the features affects the performance of the classifier. The generalization of the recognition process has been improved with the size and slant invariant signatures features of the probabilistic feature method. Experimental results indicate that the proposed recognition system performs well and is robust to the writing variations that exist between persons and for a single person at different instances, thus being promising for user independent character recognition and tolerant to random noise degradations of the characters.

Major thrust of future work will be to design a system which can significally improve the recognition performance. In general, there are three ways to simultaneously reduce the error rate, the rejection rate, and at the same time, to increase the system's correct recognition rate:1) extracting more discriminative features 2) using ensemble classifiers 3) employing a cascade classifier system. The experiment using validation set may be performed for performance analysis. To further increase recognition rate of the system we may employ a model to verify the recognized characters from the confusing character pairs. More training samples are to be used and more hierarchical levels of the classification are to be employed, so as to increase recognition performance. In order to achieve the lowest error rate while pursuing the highest recognition rate for the recognition of characters, we need to investigate further a cascade classifier structure as shown in Fig.11.The different feature sets are to be used at different layers of classification; it will make the classifiers complementary from the recognition point of view. The recognition system can use a rejection strategy to reject those characters with relatively low confidence values rather than taking a risk to misrecognize them. The rejected characters are sent to the higher level of classifiers for further recognition. At the first level of the cascade system, most of the characters should be correctly recognized; more difficult characters will be rejected and sent to the higher level classifiers. In other words, we may design classifiers at the higher levels and train them to recognize more difficult characters which are rejected by lower level classifiers. In the training procedure, for an ANN classifier at any level of the cascade classifier system, it will be trained by the rejected characters in the previous level of the classifier.

REFERENCES

[1] Rejean Plamondon and Sargur N.Srihari,, "On-line and off-line handwriting recognition: a comprehensive survey", *IEEE transactions on pattern analysis and machine intelligence*,22, pp. 62-84 , 2000.

[2] Micheal D.Garris,James L.Blue, Gerald T.Candel,Darrin L.Dimmick,Jon Geist,Patrick J.Grother, Stanley A.Janet, & Charles L.Wilson,, "Off-line Handwriting Recognition from frorms", *IEEE Int. Conf. On Systems, Man, & Cybernetics*,.3, pp.2783-2788. 1995.

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

[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] T.C.Hodson, *The Meitheis*, Low price publications, Delhi,1908.

[7] M.D.Garris, R.A.Wilkinson, and C.L.Wilson, "Methods for Enhancing Neural Network Handwritten Character Recognition", *International Joint Conference on Neural Networks*, 1, pp.695-700, 1991.

[8] Rafael C.Gonzalez, Richard E.Woods and Steven L.Eddins, *Digital Image Processing using Matlab*, Pearson Education, First Impression.2006.

[9] Philip D Wasserman, *Neural Computing Theory and Practice*, New York, Van No Strand Reinhold,1989.

[10] Th.Tangkeshwar,P.K.Bansal,Renu Vig and H.S.Kasana," A novel approach to off-line Handwritten character recognition of Manipuri script", *National Conference on Bioinformatics Computing (NCBC'05)*, TIET,Patiala,pp365-371, 2005.

[11] Th.Tangkeshwar, P.K.Bansal, Renu Vig and H.S.Kasana, "Off-line Handwritten Manipuri character recognition using K-L Divergence by probabilistic model", National workshop on trends in advanced computing(NWTAC), Tejpur University, Napaam, Assam., pp163-169,2006.