Geometrical, Profile and HOG Feature Based Recognition of Meetei Mayek Characters

Chandan Jyoti Kumar
Dept. of Computer Science & IT
Cotton College State University, INDIA
Email ID: chandan14944@gmail.com

Sanjib Kr. Kalita
Dept. of Computer Science
Gauhati University, INDIA
Email ID: sanjib959@rediffmail.com

Abstract - Recognizaton of Handwritten characters is a challenging task in the computer vision field. However, if can be done with good accuracy will be very much useful for digitizing various documents. Researchers have made various investigations for developing optical character recognition system. The accuracy of the system highly depends upon the features and classifier used. A lot of work is carried out at International level, in India also lot of work is reported in major regional scripts; Devanagri, Bangla, Gurumukhi are few of them. However, it has been observed that only few researchers have considered working on North East Indian regional scripts. As the scripts vary from one another a lot, the feature working well for one script may not be suitable for another one. Our main focus is to develop an OCR system which can work well for Meetei Mayek script. For recognition of handwritten Meetei Mayek isolated characters, we have considered Geometrical and Histogram oriented gradient feature and profile features, which are used as input to the classifier. Recognition accuracy is calculated for SVM and a comparative study is carried out.

Keywords— Meetei Mayek Script; Geometrical feature; Histogram Oriented Gradient feature; support vector machine; North East Indian regional scripts.

I. INTRODUCTION

North East India also called seven sisters are consisting of seven states of Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland and Tripura. Except the state Assam and Tripura, where the major languages are Assamese and Bengali respectively, this part of the country has mostly tribal residents whose mother tongue is Sino-Tibetan and Austro-Asiatic languages. Meetei is the third most spoken language in this region and belongs to Tibeto-Burman branch of the family of Sino-Tibetan languages. It is an official language in the state of Manipur. The Meetei Mayek script is used for writing this language. This script contains 27 alphabets, out of which 18 are original called Eeyek Eepee and 9 are additional letters which are called Lom Iyek, 8 symbols called Cheitap Iyek , 10 numerals called Cheising Iyek and 4 symbols Khudam Iyek [1]. Fig.1 shows a sample image containing all of them.

In section II we have discussed the state of art work on Meetei Mayek followed by Dataset Details in Section III. Section IV discusses about Geometrical feature, Histogram oriented feature, profile feature, Support vector machine and experimental results and then we conclude in section V.

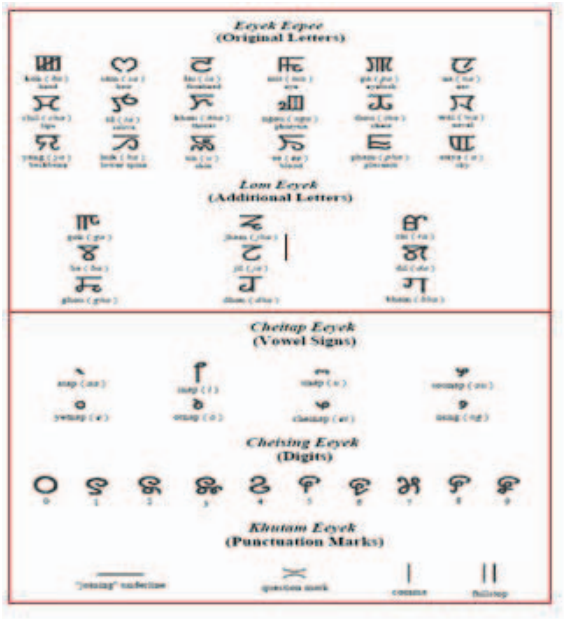


Fig. 1. Meetei Mayek Alphabet

II. LITERATURE REVIEW

K. A. Maring et al. used gabor filter based technique for recognition of Meetei Mayek numerals. They have used Support vector machine for recognition. The accuracy achieved for handwritten numerals is 89.58%, while for printed numerals; the method gives better accuracy, showing a maximum of 98.45%.

T.Thokchom et al. proposed a method for recognizing 27 characters of Meetei Mayek script where they have used fuzzy and probabilistic feature and a combination of these two, finally Artificial neural network performs the recognition. The maximum accuracy achieved is 90.3%, by using the hybrid feature.

Artificial neural network based handwritten numerals recognition is carried by R Laishram et al. Back propagation based technique is used for training the Artificial neural network. The recognition rate varies from digit to digit in the work published [2,3].

978-9-3805-4421-2/16/$31.00

Not much effort is published in the literature for Optical character recognition of Manipuri characters, so for analyzing the effectiveness of a particular feature we require to perform the experiment using that feature. In this work, we have used Histogram based feature recognition of Meetei Mayek characters. Experiment is performed for both SVM and PNN and a comparative study is carried out for recognition accuracy.

III. Dataset Details

The experiment is performed over 500 data sample of each character, for the experimental purpose we are considering all the 27 characters.

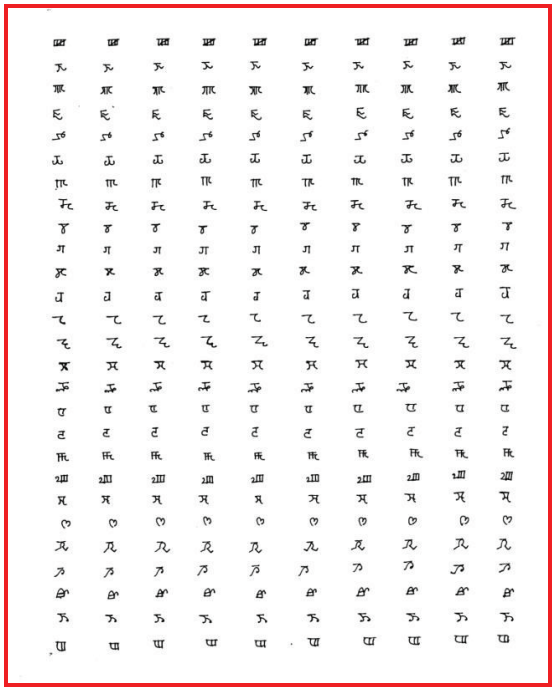


Fig. 2. Sample dataset of Meetei Mayek Script

People from various working background are considered for preparing the handwritten dataset. Although most of the contribution was from student and academic community ,while preparing the data set, we have tried to bring variation in age, working background and place as well, so that the experimental result we get become more general. People were given a page where they had to write all the 27 characters of meetei mayek script, each one ten times and Dataset was collected. Fig.2 above shows a data sample.

IV. Feature Extraction and Classification

Histogram oriented feature is extracted from the characters and the feature vector generated was used for recognition using SVM and PNN classifier.

A. Geometrical Feature

Here we are concerned on finding feature related with the geometry of the character. The feature takes care of various line segments forming the character, as per their orientation, number of holes etc. The feature vector attributes we have considered are:

1) Number of Horizontal lines 2) Number of Vertical lines 3) Number of Right diagonal lines 4) Number of Left diagonal lines. 5) Normalized Length of all horizontal lines 6) Normalized Length of all vertical lines 7) Normalized Length of all right diagonal lines 8) Normalized Length of all left diagonal lines. 9) Normalized Area of the Skeleton. 10)Number of holes in the image. 11)Ratio of Foreground and Background pixel count 12) Normalized Extreme distance with respect to centre of gravity 13) Number of junction point

B. Histogram oriented gradient Feature

Histogram oriented gradient measures how many time gradient orientations occur in a particular zone of an image. The greatest change in intensity of each pixel, both in magnitude and direction is taken care of by the gradient measure. In this method of feature extraction gradient always refer to both magnitude and direction [4]. Gradient feature vector can be computed by means of Sobel operator or Robertz operator or Prewitt operator. In this work,we use Sobel operator to determine histogram oriented gradient vector. The gradient vector used in this research work is initially resolved into two components. One component is called horizontal component and other is called vertical component. Different templates of Sobel operator is used to compute horizontal component and vertical component of histogram oriented gradient vector.

The Sobel templates used to compute the horizontal and vertical component are shown in Fig.3.

+1	+2	+1
0	0	0
-1	-2	-1

Horizontal Component

-1	0	+1
-2	0	+2
-1	0	+1

vertical component

Fig. 3. Sobel masks used

Given an input image I of size M x N at each pixel (i, j) if image I where i=1 to M and j=1 to N, its neighborhood is convolved with these templates to determine horizontal component (GRAD_X) and vertical component (GRAD_Y). The 8 neighborhood of pixel (i, j) is shown as in Fig.4.

(i-1,j-1)	(i,j-1)	(i+1,j-1)
(i,j-1)	(i,j)	(i+1,j)
(i-1,j+1)	(i,j+1)	(i+1,j+1)

Fig. 4. Neighborhood pixel of (i, j)

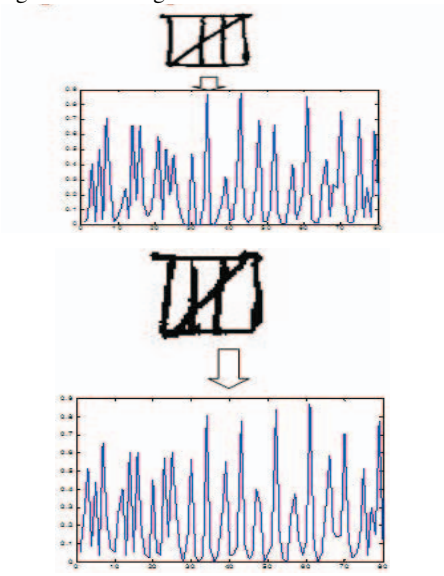
There are two equation which represents mathematical computation of horizontal component (GRAD_X) and vertical component (GRAD_Y).

GRAD_X =I (i+1, j-1) + 2*I (i+1, j) + I (i+1, j+1) - I (i-1, j-1) - 2* I (i-1, j) – I (i-1, j+1)
GRAD_Y =I (i-1, j+1) + 2*I (i, j+1) + I (i+1, j+1) - I (i-1, j-1) - 2* I (i, j-1) – I (i+1, j-1)
The magnitude of histogram oriented gradient can be calculated as:
Magnitude of Gradient (i, j)

$$= \sqrt{(Grad\_X(i,j))^2 + (Grad\_Y(i,j))^2}$$

Direction of gradient can be calculated as:
Angle (i, j) = tan-1 (Grad_Y (i, j) , Grad_X (i, j))

Histogram channels need to be considered over rectangular cells by the working out the unsigned gradient [6]. If a cell overlap half of their area, then it will contribute more than once to the final feature vector. For calculating changes in illumination and contrast, the gradient values were normalized locally, i.e. normalized over each cell. If image is not normalized properly then first it is set to properly normalized state.
We use Histogram windows per bounding boxes. Then the no. of histogram bins gets decided. For each bounding box the angles and magnitude of histogram oriented gradient get calculated as per given equations above. The computation was done over the validation data set, by considering 9 rectangular cells and 9 bin histogram per cell. The nine histograms with nine bins form an 81-dimensional feature vector. Here we get our final Histogram oriented gradient feature vector.



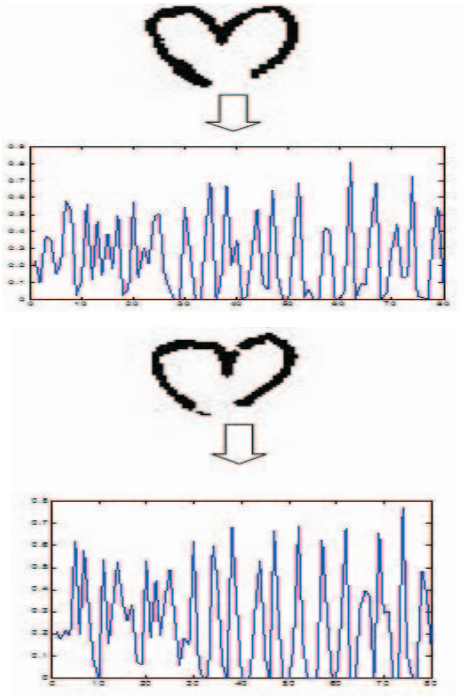


Fig. 5. Hog feature for different Meetei Mayek characters with varying data sample

Once Histogram oriented gradients are calculated they are stored in a Feature vector FV. The size of FV is 9000 X 80, implying that we have 9000 data samples and each of the data sample is having a histogram oriented feature vector of size 80x1. Out of these, 7200 samples are used for training the classifiers and 1800 are used for testing the result.

C. Distance Profiles Feature

In the computation of profile features of character, counting of number of background pixels (distance) from bounding box of the character image to the outer edge of the character is done.
We have computed profiles of four sides top, bottom, left and right. Top profiles are computed by traversing along width of the character in downward direction and bottom profiles are computed by traversing along width of the character in upward direction [7,8].
Similarly, left profiles are computed by traversing along height of the character in forward direction i.e. from left to right and right profiles are computed by traversing along height of the character in backward direction i.e. from right to left. Left profiles show the number of background pixels between left bounding box of the character and left edge of the character.
Thus all these profiles show the number of background pixels between bounding box of the character image and outer edge of the character image along different contours [9]. For dataset normalized to 30×30, distance profiles consist of 120 features as each of four profiles consists of 30 features.

D. Support Vector Machine Classifier

Support Vector Machine classifier is a two-class linear classifier that classifies testing sample based on kernel generated feature space and is designed as a weighted combination of kernel functions on training data samples [5]. The sample space is taken to a higher dimensional space separable by a hyper-plane. Kernel function plays a vital role on accuracy of classification.
For binary classification in an n-dimensional feature space, the following decision function is used:

$$f(x) = w \cdot x + b \qquad (1)$$

Where w is a vector holding the weights of various parameters, x is a vector containing input feature values and b is a bias (−b is also called threshold).Classification is given by sgn[f(x)] .
Kernels of SVM Classifiers
Following are the four different basic kernels which are used in SVM classifications:
Linear: $K(x_i, x_j) = x_i^T . x_j$.
Polynomial: $K(x_i, x_j) = (\gamma x_i^T . x_j + r)^d , \gamma > 0$.
Radial Basis Function (RBF):
$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) , \gamma > 0$.
Sigmoid: $K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + r)$.

5-fold Cross Validation:

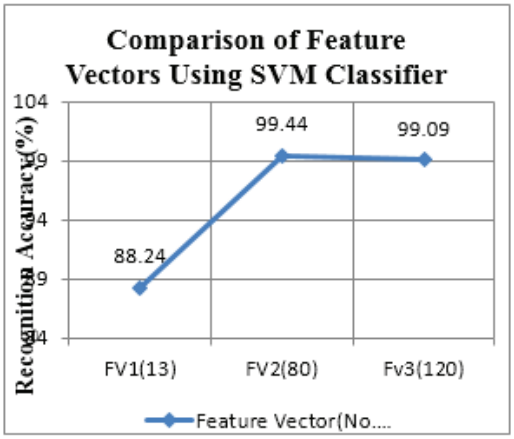
5-fold cross validation with the three classifiers is used in experiments for obtaining recognition results. First of all, 5-fold cross-validation index of the length of size of dataset, which are generated randomly, is created. The entire index contain equal size data sample, indexing is done from integers 1 through 5. Each of these integers point to a particular partition of the dataset, which is divided into 5 disjoint subsets. One division is used for testing and four other divisions are used for training [10]. This is repeated for 5 times, every occurrence we changed the testing dataset to different division and considered the remaining divisions for the purpose of training. As a result we have five different sets of feature vectors containing testing and training dataset, all of them in the size ratio of 1:4.
By the term cross validation accuracy, we are referring to the average accuracy of recognition experimented over these data samples, randomly categorized as training and testing data [11]. In following sections recognition results obtained by character and numeral recognition are discussed. Table.1. shows the recognition accuracy of Meitei Mayek Characters with various Feature Vectors using SVM classifier [12].

V. CONCLUSION

In the current work we have investigated the performance of Geometrical, Histogram oriented Gradient and Profile feature for recognition of Meetei Mayek Script and the experiment is carried out over 13500 data samples. We have used SVM classifier and observed recognition accuracy of different features. Histogram oriented feature is found better than most of the feature used for recognition of characters. However as the size of the feature vector is big it consumes more computational time, whereas the geometrical feature is very small in size consuming limited space and time. The performance of geometrical feature can be improved if we can consider some more relevant feature associated with Meetei Mayek Script. The result published here is based on well segmented, pre-processed, good quality isolated handwritten characters. The scenario may vary if we generalize the dataset, a lot of research need to be carried out in this domain. We hope this work may encourage researchers to work on major north-east Indian regional scripts.

TABLE I. RECOGNITION ACCURACY OF MEITEI MAYEK CHARACTERS WITH VARIOUS FEATURE VECTOR USING SVM

Feature Vector	Feature Vector Size	Accuracy (%)
Fv1	13	88.24
Fv2	80	99.44
Fv3	120	99.09



VI. REFERENCES:

[1]. Tangkeshwar Thokchom , P.K.Bansal , Renu Vig and Seema Bawa, "Recognition Of Handwritten Character Of Manipuri Script", Journal Of Computers, Vol. 5, No. 10, October 2010.
[2]. U. Pal and B.B.Chaudhuri, " Indian Script character recognition: a survey", Pattern Recognition vol 37,pp 1887-1899,2004
[3]. M. Blumenstein, B. K. Verma and H. Basli, A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters, 7th International Conference on Document Analysis and Recognition (ICDAR '03) Eddinburgh, Scotland: pp.137-141, 2003.
[4]. Anil K. Jain, Robert P.W. Duin, and Jianchang Mao," Statistical pattern recognition: a review," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.22, no.1, pp.4-37, Jan 2000.

[5]. ZHAO Bin, LIU Yong and XIA Shao-wei," Support vector machine and its application in handwritten numeral recognition," Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol.2, no., pp.720-723 vol.2, 2000.
[6]. Hailong Liu and Xiaoqing Ding," Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes," Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, vol., no., pp. 19- 23 Vol. 1, 29 Aug.-1 Sept. 2005.
[7]. M. Hanmandlu, J. Grover, V. K. Madasu, and S. Vasikarla," Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals," Information Technology, 2007. ITNG '07. Fourth International Conference on , vol., no., pp.208-213, 2-4 April 2007
[8]. U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol.2, no., pp.749-753, 23-26 Sept. 2007.
[9]. U. Pal, T. Wakabayashi, F. Kimura," Comparative Study of Devnagari Handwritten Character Recognition Using Different Feature and Classifiers," Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on, vol., no., pp.1111-1115, 26-29 July 2009.
[10]. S.Mori ,C Y.Suen and K. Yamamoto,"Historical Review of OCR research and development ", Proceedings of IEEE ,Vol. 80,pp1029-1058,1992.
[11]. S.W.Lee,"Off line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network", IEEE Transactions on PAMI,vol 18,pp-648-652,1996.
[12]. F.El-Khaly and M.A.Sid-Ahmed, " Machine recognition of optically captured machine printed Arabic Text", Pattern Recognition vol 23,pp-1207-1214,1990.