Recognition of Handwritten Numerals of Manipuri Script

Chandan Jyoti Kumar
Department of Computer Science and Engineering
Assam Don Bosco University

Sanjib Kumar Kalita
Department of Computer Science
Gauhati University

ABSTRACT
In this paper a support vector machine based handwritten numerals recognition system of Manipuri Script (Meetei Mayek) is investigated. We have used various feature extraction technique such as background directional distribution (BDD), zone based diagonal, projection histograms and Histogram Oriented Gradient features. In Background Directional Distribution (BDD) features background distribution of neighboring background pixels to foreground pixels in 8-different directions is considered forming a total of 128 features. For the computation of diagonal features, the whole image is divided into 64 zones of equal dimension each of size 4×4 pixels then features are extracted from the pixels of each zone by moving along the diagonal, thus consisting of 64 features in total. Projection Histograms count the number of foreground pixels in different directions such as vertical, horizontal, horizontal, left diagonal and right diagonal creating a total of 190 features. The HOG based feature is computed over the validation data set, was achieved by means of 9 rectangular cells and 9 bin histogram per cell Different combinations of these features are used for forming various feature vectors. These feature vectors are classified by using SVM classifier as 5-fold cross validation with RBF (radial basis function) kernel.

Experimental results show that the proposed 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 recognition of Meetei Mayek numeral.

General Terms
Digital Image Processing, Pattern Recognition, Optical Character Recognition, Indian Script

Keywords
BDD Feature; Projection Histogram; Zone based diagonal Feature; HOG; Support Vector Machine; Meetei Mayek Script, Handwritten Numeral Recognition.

1. INTRODUCTION
Optical character recognition (OCR) is an active area of research, especially for handwritten text. Success of the commercially available OCR system is yet to be extended to handwritten text. This is mainly because of the fact that numerous variations in writing styles of individuals make recognition of handwritten characters difficult. Past works on OCR of handwritten alphabet and numerals have been mostly found to concentrate on Roman script [1], related to English and some European languages, and scripts related to some Asian languages like Arabic, Chinese etc. Among Indian Scripts Devanagari, Tamil , Oriya, Gurumukhi and Bangla [2], have started to receive attention for OCR related research in the recent years. Now a day, lot of effort is going on for the design of OCR of most of the Indian Scripts. But, research in Manipuri Script (Meetei Mayek) recognition has not yet been widely introduced to the research community [6]. Meetei Mayek (Manipuri Script) experts say the script dates back as early as 3900 years ago [3]. Manipuri is a tonal language of Tibeto-Burman language family [4]. This script contains Iyek Ipee/Mapung Iyek, which have 27 alphabets (18 original and 9 letters called Lom Iyek, which are derived from primary 18 alphabets),Lonsum Iyek (8 letters), Cheitek Iyek (8 symbols),Khudam Iyek (3 symbols), Cheishing Iyek (10 numeral figures). In addition it contains 6 vowel letters also.The basic character appear only as the main character in a word and it may be modified using one of the vowel modifiers to produce the required vocal sound. All the original figures of the Meetei Mayek alphabets are drawn, winded and wreathed from human anatomy .The alphabetical names are the names of the different parts of the same where the characters are winded and drawn from [5]. As the characters and numerals of Meetei Mayek Script are different in their styles of writing from other Indian Scripts, so OCR system designed for automatic recognition of other Indian Script might not perform well for the Meetei Mayek. Figure 1, shows a sample of numerals of Meetei Mayek Script.

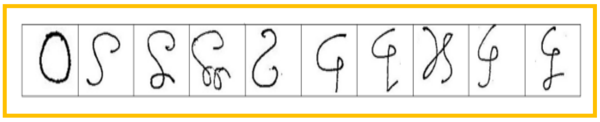


Fig 1: Meetei Mayek Numerals

This rest of the paper is organized as follows: Section 2 presents the literature survey of the handwritten character recognition system of Manipuri and other Indian Script. In Section 3, the processes of recognition of isolated characters of the input image file are presented. Section 4 explains details about SVM classifier. Section 5, presents the experimental results of the system proposed. In Section 6, some of the challenges, conclusion and future works are discussed.

2. LITERATURE SURVEY
Still now, Research in Manipuri script recognition has not yet been widely introduced to the research Community. Only very few works are done for recognition of this script. In 2010, T.Thokchom et.al [6] investigated a back propagation neural network based handwritten characters recognition system of Manipuri Script. From the character matrix they extracted probabilistic and fuzzy features. Using these features the network was trained and recognition tests were performed. They found the Performance with the combined features was robust to the writing variations. Recognition accuracy of the system was 90.3% with the combination of probabilistic and fuzzy feature. R.Vig et.al, using neocognitron simulator, the ANN was trained with the extracted character patterns and then, input pattern file containing a set of test patterns were tested for recognition on the basis of similarity in shape between patterns. Accuracy of this system was almost 90%.Recognition accuracy varies based on feature vector and the classifier used [7]. There are lot of features that are being employed for the recognition of major Indian scripts like

Bangla, Devanagari. In 2005, S.Basu et.al [8] presented a MLP based technique where they used Shadow Features, Centriod Features, and Longest-run Features. The system was designed for recognition of Bangla handwritten characters and it performed well. Experimentally observed recognition accuracy was 86.46%. In 2007, U. Pal et.al [9], presented a MQDF based technique where they used directional information obtained from the arc tangent of the gradient. Using 5-fold cross validation technique they obtained 85.90% accuracy from a dataset of Bangla compound characters containing 20,543 samples.

In 2010, S.Arora et.al [10] made an experiment to compare the performance of two major classifier ANN and SVM. In this paper they discussed the characteristics of the some classification methods that have been successfully applied to handwritten Devanagari character recognition and results of SVM and ANNs classification method, applied on Handwritten Devanagari characters. The result obtained for recognition of Devanagari characters show that reliable classification is possible using SVMs as compared to ANN. But, to the best of our knowledge not much experiment is reported for recognition of Meetei Mayek script using SVM. In this work, we are using SVM as the classifier for recognition of numerals.

3. FEATURE EXTRACTION

We have used four basic types of features for our experiment. Two of the features namely "projection histograms" and "zone based diagonal features" can be categorized as statistical features while third one namely "background directional distribution features" can be categorized as directional features [11]. On the basis of these four types of features we have formed ten feature vectors using different combinations of these four basic features [12].

3.1 Projection Histograms

Projection Histograms are computed by counting the number of pixels having value "1" in different directions. Projection histograms count the number of foreground pixels in specified direction. Four directions of horizontal, vertical and both diagonal (left diagonal and right diagonal) traversing are computed in proposed work. Thus four types of projection histograms: horizontal, vertical, diagonal-left (left traverse) and diagonal-right (right traverse) are created in our approach [13,14]. These projection histograms for a 4*4 pattern are depicted in figure 2. In horizontal histogram these pixels are counted by row wise i.e. for each pixel row. In vertical histogram the pixels are counted by column wise. In diagonal-left histogram the pixels are counted by left diagonal wise. In diagonal-left histogram the pixels are counted by left diagonal wise. In diagonal-right histogram the pixels are counted by right diagonal wise. Lengths of these features are 32, 32, 63 and 63 respectively; according to lines of traversing forming total 190 features.

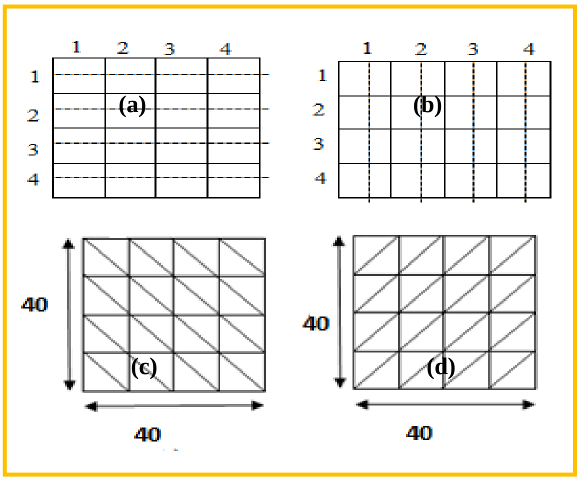
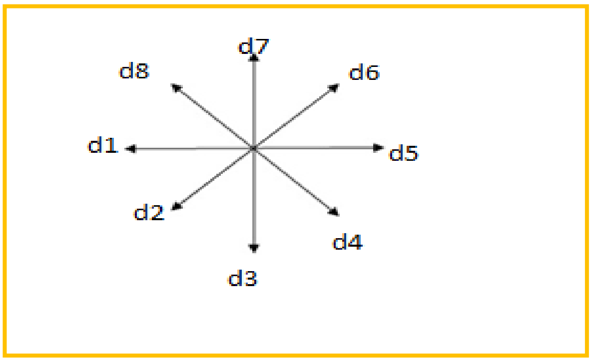


Fig 2: (a) Horizontal Histogram (b) Vertical Histogram (c) Diagonal-left Histogram (d) Diagonal-right Histogram

3.2 Background Directional Distribution

Background directional distribution features compute the directional distribution of background pixels for each foreground pixel. Each image is divided into 16 equal zones each of size 8*8 pixels. For each zone 8 directional distribution features are computed. For calculating directional distribution values of background pixels, for each foreground pixel masks and for each directional values are used. Mask for direction "d3" is shown in figure 3. The pixel at center "X" is foreground pixel under consideration to calculate directional distribution values of background. To compute directional distribution value for foreground pixel "X" in direction d3, for example, the corresponding mask values of neighboring background pixels will be added. In the similar way, we obtained all directional distribution values for each foreground pixel in 8 directions using corresponding mask. Then, all similar directional distribution values for all pixels in each zone are added, Finally 8 directional distribution feature values for each zone are computed. Figure 3,demonstrates the computation of BDD feature. In our approach we have divided image into 16 zones. So, each numeral image is represented using 128 features.



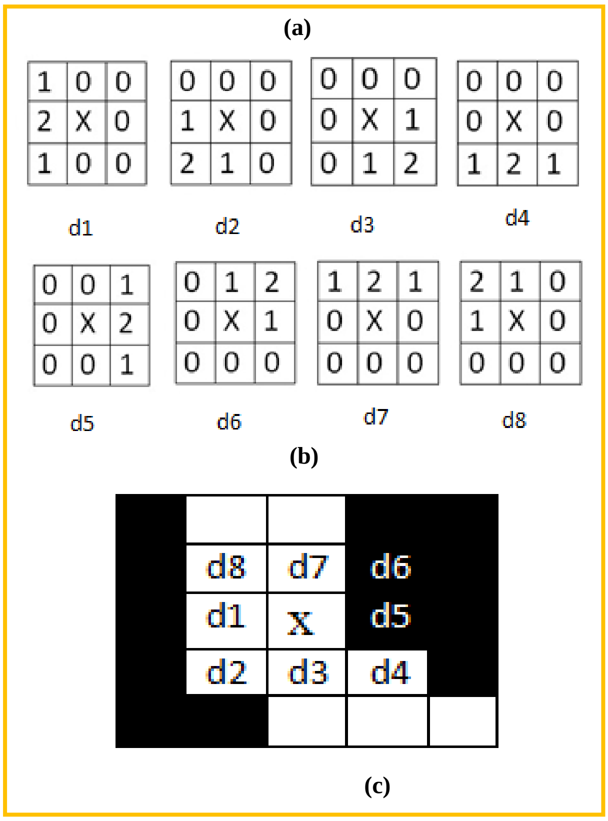


Fig 3: BDD features computation- directions, masks and example: (a) 8 directions used to compute directional distribution, (b) Masks used for computing directional distribution in different directions. (c) An example of the sample.

3.3 Zone Based Diagonal Features

In the computation of diagonal features every character image of size 32×32 pixels is divided into 100 equal zones, each of size 4x4 pixels. Then, the features are extracted from the pixels of each zone by moving along its diagonals. Each zone has 9 diagonal lines and the foreground pixels present along each diagonal line is summed to get a single sub feature and thus 9 sub-features are obtained from the each zone. These 9 sub-features values are averaged to form a single feature value and placed in the corresponding zone. This procedure is repeated sequentially for the all the zones. There may be some zones whose diagonals are empty of foreground pixels. The values of the feature corresponding to these zones are zero. Finally, 64 features are extracted for each character.

3.4 Histogram Oriented Gradient

In Histogram oriented gradient, the gradient measures the magnitude and direction of the greatest change in intensity in small neighborhood of each pixel. In this method of feature extraction gradient always refer to both magnitude and direction. 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.

We computed over the validation data set, which was achieved by means of 9 rectangular cells and 9 bin histogram per cell. The 9 histograms with 9 bins were then concatenated to make an 81-dimensional feature vector. Then, we get our final Histogram oriented gradient feature vector.

3.5 Feature Vectors

To form feature vectors for the purpose of classification, we are taking into consideration all of these four features and at the same time, some features derived from different combinations of above basic features. Table 1. Shows, the different feature used and combination of different features to derive new feature vectors.

Table 1. Sets of Feature Vectors formed with different Combinations of basic Features

Feature Vector	Features
FV1	Profiles
FV2	Background directional distribution
FV3	Zone Based Diagonal
FV4	Combinations of FV1 and FV3
FV5	Combinations of FV2 and FV3
FV6	Combination of Zone based Horizontal and Vertical features
FV7	Combination of FV1 ,FV2 and FV3
FV8	Histogram Oriented Gradient
FV9	Combination of FV1 and Fv8
FV10	Combination of FV7 and Fv9

4. CLASSIFICATION

We have used SVM classifier with Radial Basis Function (RBF) kernel for the recognition of numerals of Meetei Mayek Script. Support vector machines (SVM) are a group of supervised learning methods that with very good Generalization ability that can be applied to classification and pattern recognition. SVMs represent an extension to non-linear models of the generalized portrait algorithm. It is primarily a two class classifier. The Width of the margin between the classes is the major optimization criterion, i.e. the empty area around the decision boundary [16], defined by the distance to the nearest training pattern. These patterns called support vectors, which finally define the function for classification.

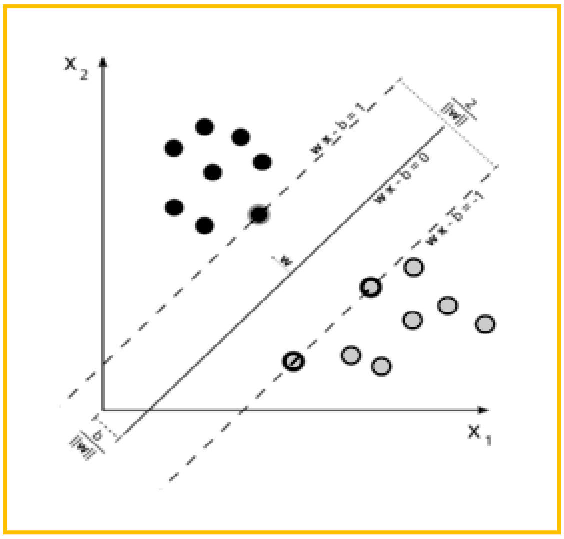


Fig 4: Maximum-margin hyper plane and margins for an SVM trained with samples from two classes. (X1 and X2)

Experiments are performed on LIBSVM 3.0.1, which is multiclass SVM and selecting RBF kernel. Following are the four basic kernels used in SVM classifications:

a) **Linear**: $K(x_i, x_j) = x_i^T . x_j$.

b) **Polynomial**: $K(x_i, x_j) = (\gamma x_i^T . x_j + r)^d$, $\gamma > 0$.

c) **RBF**: $K(x_i, x_j) = e^{(-\Upsilon |Xi-Xj|^2)}$

d) **Sigmoid**: $K(x_i, x_j) = \tanh(\gamma {x_i}^T x_j + r)$.

LIBSVM with Radial Basis Function (RBF) kernel [12], a popular, powerful kernel, is used for this experiment.

5. EXPERIMENTS AND RESULTS

We have tested our technique over a set of 2000 samples of data. These dataset are written by the 20 different people from different age group and of different profession. Figure 5, shows a sample image from the Handwritten dataset of Meetei Mayek numerals used for the Experiment.

In this work, we have used four basic features and a multiple set of combinations of all the four for handwritten Meetei Mayek numerals recognition. Experimental results show that HOG feature is more appropriate as compared to others for this Script. Highest recognition accuracy of 95.16% is obtained for feature vectors FV10 which comprises of Histogram, HOG, BDD and Diagonal features. Second highest results of 94.24% are obtained with FV8 (HOG).Figure 6 shows, a comparison among the recognition accuracy of different feature vectors. Table 2 lists all the accuracy of the experiment for the value of parameter $\Upsilon = 0.4$ and Cost C=500.

Fig 5: A sample image from the Dataset of Meetei Mayek handwritten numerals

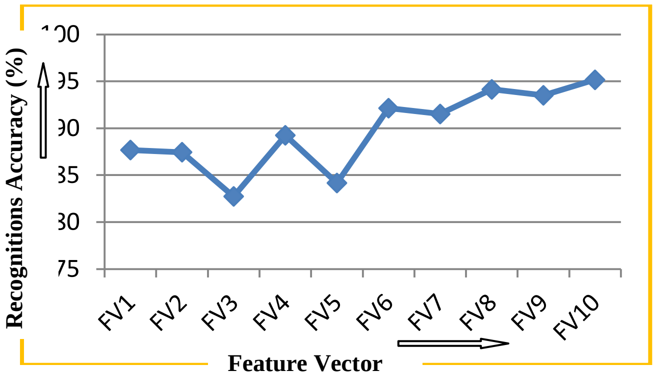


Fig 6: Comparison of results of different features

Table 2. Recognition accuracy of various feature vectors

Feature Vector	Recognition Accuracy (%)
FV1	87.68
FV2	87.45
FV3	83.74
FV4	89.25
FV5	84.17
FV6	92.14
FV7	91.52
FV8	94.14
FV9	93.46
FV10	95.16

6. CONCLUSION AND FUTURE WORKS

In this , a comparative analysis of the performance of different features for OCR of handwritten Meetei Mayek numerals is presented. The result obtained for recognition show that reliable classification is possible using SVMs [15]. Our technique shows maximum accuracy of 95.16%, which is certainly a promising improvement as compared to previous works on Meetei Mayek script.

The SVM-based method described here for Handwritten Meetei Mayek numerals can be extended to degraded text recognition also [17]. Although SVM shows higher accuracy than ANN, new classifier like MQDF can be used for recognition purpose. This work can further extended to enhance the performance by adding some more relevant features.

7. REFERENCES:

[1] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros and George Carayannis, "Handwritten document image segmentation into text lines and words." Pattern Recognition 43,2010.

[2] Plamondon.R., and S.Srihari "Online and offline Handwriting Recognition : A Comprehensive Survey." IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 22 No.1, 2000.

[3] D.N.S Bhat and M.S. Ningomba , "Manipuri Grammar", Central Institute of Indian Languages,1995.

[4] Chungkham Yashawanta Singh, Manipuri Grammar, Rajesh Publications,New Delhi, 2000.

[5] R M K Sinha, " A journey from Indian scripts processing to Indian language processing" , IEEE Ann. Hist. Computer, vol 31, no 1, pp 831, 2009

[6] Tangkeshwar Thokchom, P.K.Bansal, Renu Vig and Seema Bawa, "Recognition of Handwritten Character of Manipuri Script", Journal Of Computers, Vol. 5, No. 10, Oct 2010.

[7] N. Sharma, U. Pal, F. Kimura, and S. Pal, "Recognition of offline hand-written Devnagari characters using quadratic classifier," inProc. Indian Conference. Of Computer. Vision. Graph. Image Processing, 2006, pp. 805–816.

[8] Subhadip Basu , Nibaran Das, Ram Sarkar, Mahantapas Kundu,Mita Nasipuri and Dipak Kumar Basu, "Handwritten Bangla alphabet recognition using an MLP based classifier", 2nd National Conf. On Computer Processing of Bangla, Pp. 285-291, Feb-2005, Dhaka.

[9] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten Bangla Compound Character Recognition using Gradient Feature" , 10th International Conference on Information Technology,2007.

[10] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, L. Malik , M. Kundu and D. K. Basu, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, May 2010.

[11] Anita Rani, Rajneesh Rani and Renu Dhir , "Combination of Different Feature Sets and SVM Classifier for Handwritten Gurumukhi Numeral Recognition", International Journal of Computer Applications (0975 – 888) Volume 47– No.18, June 2012.

[12] Anita Rani, Rajneesh Rani and Renu Dhir,"Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5, May 2012.

[13] Mahesh Jangid, "Devanagari Isolated Character Recognition by using Statistical features", International Journal on Computer Science and Engineering (IJCSE) Vol. 3 No. 6 June 2011.

[14] Ratnashil N Khobragade, Dr. Nitin A. Koli and Mahendra S Makesar , "A Survey on Recognition of Devnagari Script", International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013.

[15] Manal A. Abdullah, Lulwah M. Al-Harigy, and Hanadi H. Al-Fraidi, "Off-Line Arabic Handwriting Character Recognition Using Word Segmentation" Journal Of Computing, Vol. 4, Issue 3, Mar 2012.

[16] Hande Adiguzel, Emre Sahin and Pınar Duygulu, "A Hybrid Approach for Line Segmentation in Handwritten Documents" , International Conference on Frontiers in Handwriting Recognition ,2012.

[17] Thoudam Doren Singh , "Taste of Two Different Flavours: Which Manipuri Script Works Better" Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 11–18, Atlanta, Georgia, 13 June 2013.

8. AUTHOR'S PROFILE

Dr. Sanjib Kr. Kalita is an Assistant. Professor, Dept. of Computer Science, Gauhati University. Dr. Kalita, Previously worked with various government colleges and Universities in the position of HOD. He was the head in the dept. of Computer Science, Handique Girls' College, Guwahati (2002-09).He headed the dept. of Computer Science, Gauhati University, Kokrajhar campus (presently Bodoland University) (2009 -12). He has done his Ph. D. from Gauhati University. He has authored over thirteen peer-reviewed articles, and one book. His area of research is Speech Processing, Signal Processing, Image Processing and Data Mining.

Mr. Chandan Jyoti kumar received his B.Tech from NIT Silchar and M.Tech from NIT Jalandhar. Currently he is working as a Assistant Professor in Assam Don Bosco University. He has authored over six peer-reviewed articles. His current areas of research are OCR, Image Processing, Machine Learning and Pattern Recognition.