Structured based Feature Extraction of Handwritten Marathi Word

C.Namrata Mahender
Dept of CS and IT,
Dr.B.A.M.University,
Aurangabad.

K.V.Kale
Dept of CS and IT,
Dr.B.A.M.University,
Aurangabad.

ABSTRACT
Writing which has been the most natural method of collecting, storing and transmitting information through the centuries, now serves not only for the communication among humans, but also for the communication of humans and machines. The free style handwriting recognition is difficult not only because of the great amount of variations involved in the shape of characters, but also because of the overlapping and the interconnection of the neighboring characters.

In this paper we are presenting a structured based feature extraction for handwritten Marathi word. And rule based recognition was applied which has given 85 to 90 % recognition rate.

General Terms: Pattern Recognition, OCR .

Keywords: Structured Feature extraction, Recognition and Rule based approach.

1. INTRODUCTION
Writing which has been the most natural method of collecting, storing and transmitting information through the centuries, now serves not only for the communication among humans, but also for the communication of humans and machines.

From the recorded history, only humans were able to recognize and interpret the handwriting of other human being. The term "Handwriting" is defined as meaning to a surface consisting of artificial graphic mark conveying some messages through the marks conventional relation to language [1]. Fueled by the curiosity to uncover the secret of human minds, many scientists began to focus their attention on attempting to mimic intelligent behavior. One of such attempt is to imitate the way the human read and recognize printed and handwritten matter.

Optical Character Recognition (OCR) is the most crucial part of Electronic Document Analysis Systems. The solution lies in the intersection of the fields of pattern recognition, image and natural language processing. Although there has been a tremendous research effort, the state of the art in the OCR has only reached the point of partial use in recent years. Nowadays, cleanly printed text in document with simple layouts can be recognized reliably by off-the-shelf OCR software [3 7]. There is only limited success in handwriting recognition, practically for isolated and neatly hand-printed characters and word for limited vocabulary. However, in spite of the intensive effort of more than thirty years, the recognition of free style and cursive handwriting as in the example of figure 1 continues to remain in the research arena.

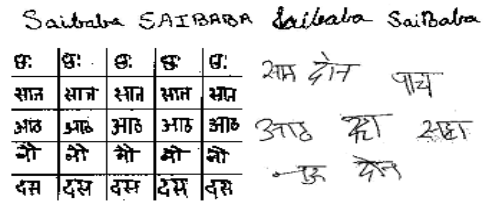

Figure 1: An Example of Handwriting in English and Devanagari script

The difficulty of the free style handwriting recognition is not only because of the great amount of variations involved in the shape of characters, but also because of the overlapping and the interconnection of the neighboring characters. In addition to the peculiarities of an author's idio-script, which means one writer can be identified among thousands; there are the peculiarities of writing in different situations, with different media and for different purposes. Sometimes it may not be possible to extract the characters from the whole word. Furthermore, when observed in isolation, characters are often ambiguous and require context to minimize the recognition errors.

In order to reach the ultimate goal of fluent machine reading many sub problems have been investigated applying some constraints to the recognition system. One constraint may be imposed on data acquisition equipment, by using a digitizer as in the on-line recognition task is performed concurrently. This approach avoids complicated pre-processing operations and captures the temporal information. On the other hand, the input in the off-line recognition systems is a pre-written document image that is converted into a bit pattern data through an optical scanner. The data is contaminated with various sources of noise. Another constraint is the restriction of the system to the writing of a specific user. In this case, the system is "writer dependent". It is possible to make the system learn accurately the writing style of the user and have a satisfactory performance in recognizing a specific writing. Thirdly, the size of the vocabulary can be limited. The system looks for the whole word in lexicon. Which best fits the unknown word. Thus, it avoids the recognition errors due to imperfect segmentation and increases the recognition rates. Lastly, some constraints can be applied on the writing style using pre-printed guidelines or boxes on forms or envelopes.

Therefore, the methods and the success of the recognition rates depend on the level of constraints on handwriting. Considering the roman script, the difficulty is less for handwriting produced as a sequence of separated characters than for the cursive script.

For other writing systems, character recognition is hard to achieve, as in the case of Chinese characters, which is characterized by complex shapes and a huge number of symbols. In the case of Devanagari script the situation gets more complicated due to number of vowels, consonants and the half consonants, i.e. similar characters with lot of modifiers and the extensive use of Header line called as Shirorekha [1,2,3],.

In this we are presenting a novel structured based feature extraction technique for handwritten Marathi words of numerals.

2. FREESTYLE HANDWRITTEN MARATHI WORD

The free style handwriting recognition is difficult not only because of the great amount of variations involved in the shape of characters, but also because of the overlapping and the interconnection of the neighboring characters. Figure 2 shows the sample data used.

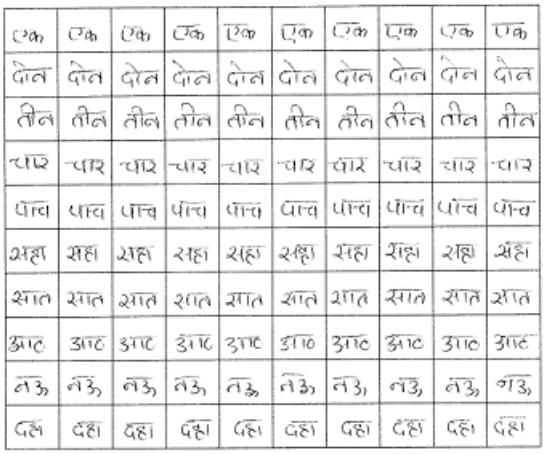

Figure 2: Shows on of the sample data used in our work.

The collected data is preprocessed; resized and thin images are used for further processing.

2.1 Feature Extraction

There are lots of different types of tilt in the characters some are due to the nature of the vowels and consonants but some are due to the writing style as shown in the figure 3 even a straight line in two images are not that straight which is effecting the values calculated for finding proper features [6 7 20].

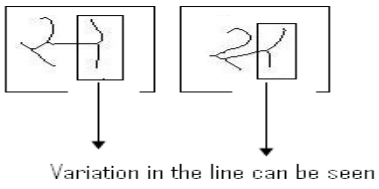

Figure 3: Variation in the line drawn in the same characters

Thus we wanted to extract the character into basic components for this we have used the concept of Psychology of reading.

2.2 Psychology of reading

Visual recognition of words has been widely investigated by psychologist during the past century and has produced two very different interpretations. Holistic theories suggest that words are identified directly from their global shapes. Holistic theories of reading propose that reading is accomplished using stored encodings of shapes of words as shown in figure 4. Hierarchical theories on the other hand, hypothesize that a words recognized from letters from features detected in stimulus.

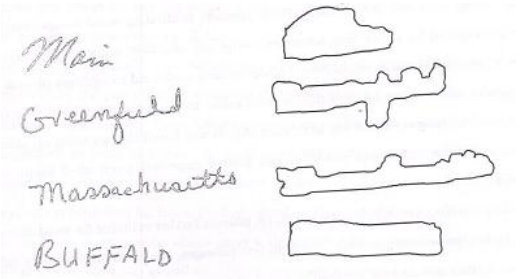

Figure 4: The word shape of English words written different styles

2.3 Polygon or box fitting Approach

Taking into consideration of things discussed above we have utilize the befits of letter shape of letters to reduce the complex nature of Contours and as mentioned above instead of word shape. Word level template of characters designed using the following parameter:

1) Length of words and 2) Probability of occurrence of a single character with respect to other characters represented .

The basic component of the approach in detail:

Polygon or box fitting : the concept of fitting the polygon or box to the character is carried in Top-Down approach for this purpose five variety of box has been used that is closed box , Lower open ,left open, right open and down open and a line characters with no boxes as shown in figure 5.

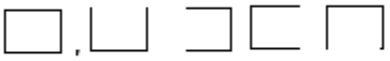

Figure 5:All the shapes for Box Fitting

Steps for fitting the box:

a) The character is scanned and the location where the box can be fixed is located.

b) Then the box is fixed according to the character.

c) But the character like ‘sa’ have joints with it like ligatures due to which proper boxes are not fixed thus left to right scan is done and step a and b steps are repeated.

d)Lastly the boxes are resized to 7 X 7 matrix and matched to one of the template for generating the string for the character.

Example of fitting the box in character 'ka' partial section scanning from left to right are shown in figure 6

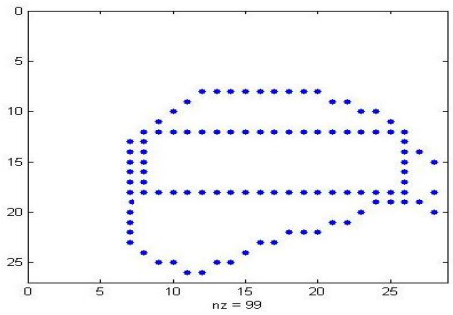

Figure 6: Complete box drawn

Some sample of box fitting of few characters is shown in following figures 7 to 9:

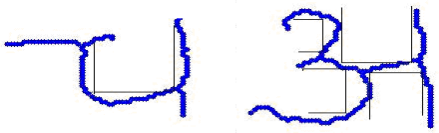

Figure 7: showing Box fitting for Cha and AH

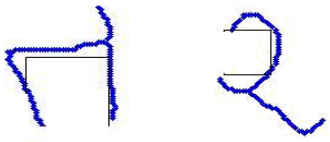

Figure 8: showing Box fitting for Ta and Ra

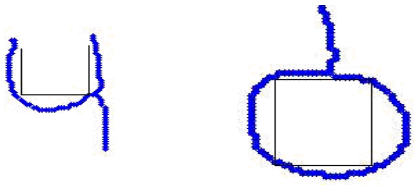

Figure 9: showing Box fitting for pa and laa

Few characters having two templates due to the variations found in the database like Pa and cha having same template

The first problem can be overcome by just standardizing one of the template depending upon the maximum character converges to it, but this will drop the recognition rate. Second problem can be addressed by looking as the nature of both the character occurrences with respect to each other and other characters. It was found that pa is having no relation other than cha , while cha is having relation with ra. So it helped us to recognize them based on the probability of the occurrences of the other characters that takes the consideration of based on the concept of bayes net, in this case both the character has same length size in word view.

3. CLASSIFICATION AND RECOGNITION

Once the feature selection finds a proper representation, a classifier can be designed using the number of possible approach available in the literature. Recognition plays the important role in assigning a label to a character based on the information provided by its descriptors. We used a rule based approach for the structured based features which are discussed in detail in remaining sections.

3.1 Rule based approach

Rule base approach is based on If Condition then Action structure. Our structure based features are represented by rule based approach. For recognition every possible condition is verified for the feature.

Variables:

WordBit; Length of the character in word either one or two in this case.

TestSample1; First character in the word

TestSample2; Second character in the word

Table:1 Label for the respective box

L1;	
B1;	
UB;	
LB;	
RB;	
DB;	

Rule Based Algorithm:

If WordBit = 1

then If TestSample1 = = 'L1' then Bitstream 5

Elseif TestSample1 = = 'LBRB'OR 'LBUB' then Bitstream 3

Elseif TestSample1 = = 'RBRB' then Bitstream 9

If WordBit = 2

then If TestSample1 = = 'L1L1' then Bitstream 1

Elseif TestSample1 = = 'DB' then Bitstream 4

Elseif TestSample1 = = 'UB' then

If TestSample2 = = 'UB' then Bitstream 7

Elseif TestSample2 = = 'LB' then Bitstream 6

Elseif TestSample1 = = 'LBUBDB' then Bitstream 10

Elseif TestSample1 = = LBLBUBDB' then Bitstream 11

and

If TestSample2 = = 'B1B1' then Bitstream 2

Elseif TestSample2 = = 'L1' then Bitstream 5

Elseif TestSample2 = = 'LB' then Bitstream 8

Elseif TestSample2 = = 'UB' then Bitstream 6

Elseif TestSample2 = = 'DB' then Bitstream 4

Elseif TestSample2 = = 'B1' then Bitstream 12

Elseif TestSample2 = = 'LBUBDB' then Bitstream 10

3.2 Codebook Generation for Recognition

Codebook generation is an essential stage of Hidden Markov Model for training and clustering in HMM before training. In discrete density HMM's represented every observation as one of a number of discrete vector prototypes. By observing the frequency of each of these prototypes, the probability of each can be estimated. In their original form, the observation may take the same form as these prototype vectors, take on larger variety of discrete values or may take on values from some continuous distribution. In either of the last two cases, before training a model must be first created i.e. a codebook of prototype. Such that any observation can be represented by one of these prototypes. Thus in our system we are trying to create a codebook prototype for each pseudocode.[21,22]

We created the codebook, so that a predetermined number of clusters must be created that will partition the input space into linearly separable regions, one for each codebook. This partition is most often done by selecting an initial set of prototype vectors and iteratively improving on them. Gray et al. [26] describes methods of selecting an initial codebook. Methods of random selection include selecting the first set of observations from the training data to act as the initial codebook vectors, and selecting training observations at regular interval to act as the initial vector, the latter of which act as a better starting point if data are highly correlated.

Once an initial set of codebook vectors have been selected, classifying each to its closest prototype vectors clusters training observations. Using these new clusters each codebook vectors can be recalculated to more accurately reflect the real cluster of vectors that are falling within it cluster.

Clusters are created using data from the entire set of training observations as a single pool of data or by using class labels, provided by the training observations to create clusters from each individual class pool of data. The former techniques is known as intrinsic or unsupervised while the later is known as extrinsic or supervised clustering [27]. As examples these labels may be characters from an alphabet, words, phonemes or some smaller subgroup of one of these. Humans generally provide labels by a very tedious job called truthing. Due to the large number of man-hours it may be the case that higher-level labels will be available, but labels of higher granularity, such as characters are not available. In this case a supervised clustering can be performed at the word level, while unsupervised clustering must be performed at the sub-word level.

Bitstream used in work is nothing but 13 bit long stream of zeros and ones for Marathi letters. The Bitstream value represented in the rule based structure shows the on position of bit for the character in the bitstream which is unique for each character, which is acting as a codebook for us.figure 10 which shows the codebook in Matrix form for Marathi character.

	ए	क	द	न	त	च	र	ष	स	ह	अ	ट	ॐ
ए	1	0	0	0	0	0	0	0	0	0	0	0	0
क	0	1	0	0	0	0	0	0	0	0	0	0	0
र	0	0	1	0	0	0	0	0	0	0	0	0	0
न	0	0	0	1	0	0	0	0	0	0	0	0	0
त	0	0	0	0	1	0	0	0	0	0	0	0	0
च	0	0	0	0	0	1	0	0	0	0	0	0	0
र	0	0	0	0	0	0	1	0	0	0	0	0	0
ष	0	0	0	0	0	0	0	1	0	0	0	0	0
स	0	0	0	0	0	0	0	0	1	0	0	0	0
ह	0	0	0	0	0	0	0	0	0	1	0	0	0
अ	0	0	0	0	0	0	0	0	0	0	1	0	0
ट	0	0	0	0	0	0	0	0	0	0	0	1	0
ॐ	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 10:Character Bitstream in matrix form for Marathi character

4. RESULTS

For each character 200 samples were taken, and and all 200 character were tested, against their class and with other class template of characters. Following table 2 shows the results for recognition were best result is for character "ahi" and the lowest is for "ahoo" as major things of "ahoo" collide with many different classes and the character is also not so clear.

Table 2 : Results based on Structural Approach for sorted data

Character	Total sorted Data Samples	RS	% Rec
ए	200	180	90.00
क	200	174	87.00
द	200	160	80.00
न	200	178	89.00
त	200	170	85.00
च	200	160	80.00
र	200	160	80.00
प	200	168	84.00
स	200	155	77.50
ह	200	140	70.00
अ	200	150	75.00
ठ	200	170	85.00
ऊ	200	147	73.50
Total	2800	2262	80.78

Where, RS is Recognized Samples Used and %Rec is Recognition Rate of each character

5. CONCLUSION

Free style handwriting recognition is a difficult problem, not only because of the great amount of variations in handwriting, but also because of the overlapping and the interconnection of the neighboring characters. Furthermore, when observed in isolation, characters are often ambiguous and require context to minimize the classification error. In this paper we aimed at developing systems for limited domain applications. A novel approach for feature extraction that is box based method is applied which is actually taken on the basis of the concept of psychology of reading and a rule based recognition system was developed which has shown better results compared to statistical based approaches.

On an average the performance of recognition of the system is 85 to 90 %, but the major constraint of the system is the limited vocabulary.

6. REFERENCES

[1] R.Phamondon and S.N Shrihari, "On-line and Offline Handwriting Recognition: A comprensive survey, IEEE Transaction on Pattern Analysis and Machine Intelligence", 22, 63-84,2000.

[2] C.Y.Suen,R.Legault,C.Nadal,M.Cheriet and L.Lam, "Building a new generation of handwriting recognition systems, Pattern Recognition Letters, 14,305-315,1993.

[3] S.Impedovo, L.Ottaviano and S.Occhinegro," Optical Character Recognition – a survey", International Journal of Pattern Recognition and AI, 5,1-24,1991.

[4] Jean.R.Ward and Thedore Kuklinski, "A Model for Variability effects in Handwriting Character Recognition Systems in IEEE Trans Sys.Man.Cybernatics, Vol 18, No 3, pp 438-451, 1988.

[5] L.Lam, S.W. Lee and C.Y.Suen, "Thining Methodologies. A Comprehensive survey",IEEE Trans. Pattern recognition and Machine Intelligence, Vol.14,pp.869-885,1992.

[6] A.C.Downtown, C.G.Leedham, "Preprocessing and Presorting of Envelope Images for Automatic Sorting Using OCR", Pattern Recognition, Vol:23, No:3-4,pp: 347-362,1990.

[7] S.N. Shrihari and S.W.Law, "Character Recognition, Technical Report", CEDAR-TR-95-1.

[8] C.Weliwilage, A. Harvey, A.Jennings, "Whole of word Recognition Methods for Cursive Script".

[9] W.K.Pratt,1991 Digital Image Processing,wiley interscience.

[10] R.G.Casey, "Moment Normalization of Handprinted Characters",IBM Journal of Research Development, 548-557, 1970.

[11] R.S.Gonzales,P.Wintz,Digital Image Processing Addison-wesley publishing co,1987.

[12] M.Y.Chen,A.Kundu and J.Zhou,"Offline Handwritten Word Recognition using a Hidden Markov Model Type Stochastic Network",IEEE Trans.Pattern recognition and Machine Intelligence,vol.16,pp.481-496,1994.

[13] A.Atici,Fatos,Yarman-Vural, "A heuristic method for Arabic Character recognition" , journal of signal processing,vol 62,pp.87-99,1997.

[14] R.G. Casey and E.Lecolinet, " A survey of methods and stratergies in character segmentation", IEEE Trans. Pattern recognition and Machine Intelligence, Vol.18,pp.690-706,1996.

[15] C.E Dunn and P.S.P.Wang, "Character Segmentation techniques for Handwritten Text –A survey", Proc. of 11th International Con on Pattern Recognition", The Hague, The NetherLands, pp 577-580, 1992.

[16] C.Y.Suen, "Character Recognition by Computer and Applications in Handbook of Pattern Recognition and Image processing", Young T.Y, Fu, King-Su, Academic Press in San Diego, CA, 569-586, 1986.

[17] M.Shridhar and a Badreldin,"Recognition of Isolated and Simply Connected handwritten Numerals', Pattern Recognition", Vol 19, No 1, pp1-12, 1986.

[18] W .Y.Kim. and P.Yuan, "A practical Pattern Recognition System for Translation, Scale and Rotation Invariance", CVPR`94,Seattle, Washington,June 1994.

[19] C.Y.Suen, M.Berthod and S.Moris, "Automatic Recognition of Handprinted Characters -the state of the Art", Proceeding of the IEEE Vol:68, No 4,pp 469-487, 1980.

[20] Mohiuddin and J.Mao, A Comparative study of different classifiers for handprinted character recognition, Pattern Recognition pp. 437-448,1994.

[21] A. J.Elms, "A connected character recognizer using Level Building of HMMs", IEEE 12TH IAPR International Conference on Pattern Recognition IIpp. 439-441,1994.

[22] A.Kundu,Y.He and P.Bhal, " Recognition of Handwritten word: First and second order HMM based Approach", Pattern Recognition, Vol 22, pp. 283-297,1989.

[23] Siganesh Madhavanath and V.Govindaraju," The Role of Holistic Paradigm in Handwritten Word Recognition", IEEE PAMI, Vol 23, No2, Feb-2001.

[24] Watanable, LQin, N. Sugie. "Structure Recognition Methods for Various Types of Documents", Machine vision and Applications vol, no 6, pp163-176 1993

[25] J. Rocha , T. Pavilidis. "A Shape Analysis Model" ,IEEE Tran on Pattern and machine Intelligence , Vol.16,No.4,pp394-404,1994.

[26] R.M.Gray," Vector Quantization", "Readings in Speech Recognition Morgan Kaufmann Publishers",Inc., pp.75-100,1990.

[27] A.K.Jain and R.C.Dubes," Algorithm for Clustering Data",Prentice Hall, Englewood Cliffs, 1998.