MARATHI CHARACTER RECOGNITION USING ANT MINER ALGORITHM

[1]URMILA SHINDE, [2]VANITA MANE, [3]RAJASHREE SHEDGE

[1,2,3]Computer Department, RAIT, Nerul, Navi Mumbai, Maharashtra, India.
E-mail: urmigd@gmail.com, vanitamane1@gmail.com, rajashree.shedge@gmail.com

Abstract- Optical Character Recognition (OCR) is an interesting and challenging field of research in pattern recognition, artificial intelligence and machine vision and is used in many real life applications. The work done for the recognition of Devanagari handwritten script is negligible in literature despite it is being used by millions people in India and abroad and it has numerous applications. Research on Optical Character Recognition OCR of Devnagari script is very challenging due to the complex structural properties of the script that are not observed in most other scripts. Devnagari is the script for Marathi. The Marathi language contains 49 distinct characters, 12 vowels and 37 consonants. Recognition of Devnagari characters poses great challenge due to the large variety of symbols and their proximity in appearance. Feature extraction and classification are the two very important steps in Optical character recognition. In this paper we have used three sets of feature extraction techniques namely Hu's moment invariant, zoning with Hu moment and Zoning with Hu's moment and radon transform. Here we have proposed Ant Miner Algorithm (AMA) for classification. The AMA is a rule-based approach. The rules are incrementally tuned during the training. The result of this experiment is a 96.94% recognition rate of the training set and 82.21% recognition rate of unseen data test

Keywords- OCR, Devnagari Character Recognition, Off-line Handwriting Recognition, Segmentation, Feature Extraction.

I. INTRODUCTION

Optical Character Recognition (OCR) is the technique which enables a machine to automatically recognize the characters or scripts written in the users' language. Marathi script is derived from Devanagari. It is an official language of Maharashtra. Marathi script consists of 12 vowels and 37 consonants making 49 alphabets. Marathi is written from left to right. It has no upper and lower case characters. Every character has a horizontal line at the top called as the header line. The header line joins the characters in a word.

While a large amount of literature is available for recognition of English script, relatively less work has been reported for the recognition of Indian languages. Main reason for this slow development could be attributed to the complexity in the shapes of Indian scripts, and also the large set of different patterns that exists in these languages, as opposed to English. For handwritten Devnagari character recognition system and accuracy reported is not high and dataset used are not large. The study shows that, OCR implementation in Marathi language is more challenging due to its writing style complexity. To handle the above difficulties, we have proposed Ant miner algorithm for classification to increase the recognition rate. The Ant-Miner algorithm (AMA) is a modified version of the Ant Colony Optimization (ACO), which is used for classification based on a set of rules. Initially the list of rules is empty and the training set consists of all the training cases. In each iteration of training, it discovers one classification rule. This rule is added to the list of discovered rules. After the training it is ensured that all the training cases are correctly covered by this set of rules. Rest of the paper is organized as follows. Section 2 describes the Literature survey for this study. Section 3 describes the proposed system in detail.Section 4 represents result and discussion. Finally, section 5 concludes the paper and directs the future extensions of this work.

II. LITERATURE SURVEY

A.Character Recognition Model

For any character recognition system there are four major stages namely, Pre-processing, Segmentation, Feature extraction, Classification & Recognition. These stages are shown in the following Fig. 2.1.

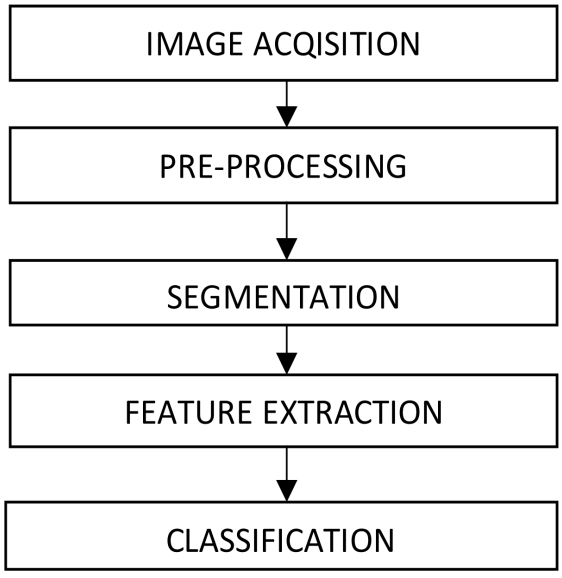

Figure 2.1: Block Diagram of Character Recognition System

B. Related Work

Devanagari script started in early 1970s. An extensive research on printed Devanagari text was carried out by Veena Bansal and R. M. K. Sinha. First system for hand-written numeral recognition of Devanagari characters was proposed by R. Bajaj et al. P. M. Patil and T.R. Sontakke also presented an algorithm for ndwritten Devanagari numeral recognition which was rotation, scale and translation invariant. U. Pal et al presented a system for off-line handwritten character recognition of Devanagari using directional information for extracting features. A technique for accuracy improvement of Devanagari character recognition system was proposed by U. Pal et al using two features based upon directional and curvature information in the characters and applied to the combination of two classifiers namely, support vector machines and modified quadratic discriminate function. A comparative study of various features and classifiers used for handwritten Devanagari character recognition was done by U. Pal et al. N Sharma et al. proposed a method where the features are obtained from the directional chain codes information of the contour points.

A Multi-feature multi-classifier scheme for handwritten Devanagari characters is proposed by S. Shelke and S. Apte, which combines neural network and template matching recognition Approaches. Recognition of Non-Compound Handwritten Devnagari characters using a Combination of MLP and Minimum Edit Distance was proposed by S. Arora et al. for compound characters involving two stages. A methodology for off-line isolated handwritten Devanagari character recognition is proposed by Mahesh Jangid. Brijmohan Singh and Ankush Mittal proposed the two different methods for extracting features from handwritten Devnagari characters, the Curvelet Transform and the Character Geometry.

III. PROPOSED SYSTEM

In our proposed system we recommend the use of Ant Miner algorithm for classification. We first will acquire character image using scanner to get character in digital format. After data acquisition, it will be preprocessed to be suitable for feature extraction and classification. In preprocessing, image will be processed for Background Elimination, Noise Reduction, Width Normalization and thinning. Then the character image will be fed for feature extraction where first the zoning will be applied on preprocessed images to extract features and then the Hu's moment will be applied. Thus we will use two feature extraction techniques here to retrieve more refined features. Extracted features will then be fed to the Ant miner classifier for training and testing. Flow of our system is shown in following Fig. 3.1.

A. Preprocessing

The purpose of the pre-processing phase is to make characters standard and ready for feature extraction. The preprocessing step is applied both in training and testing phases. The pre-processing stage primarily involves some of the following steps:

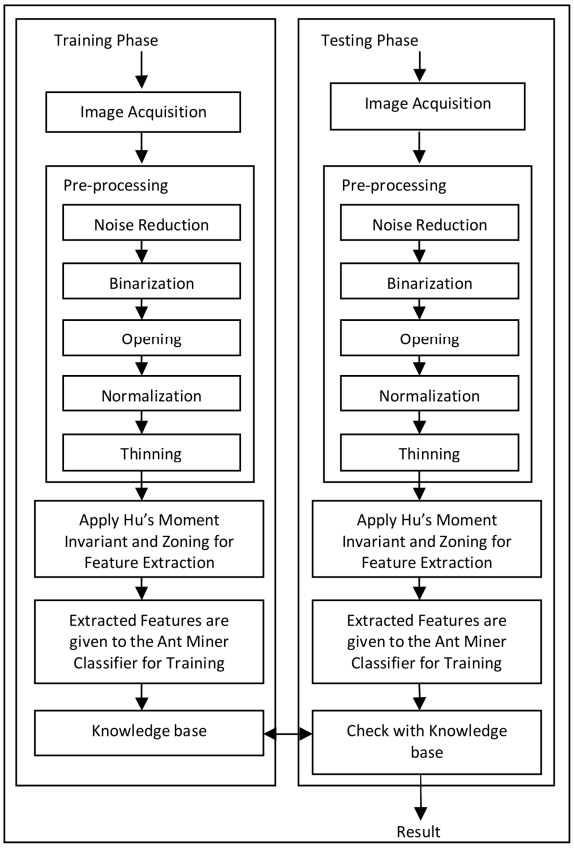

Figure 3.1: Proposed system

1) Image Averaging (Noise Reduction)

During the scanning process, some distortion in images may be introduced due to poor quality of pen, light hand handwriting & poor paper quality on which the characters are written etc. It causes some noise in these images. Recognition using computer requires noise free & distortion free images of the characters. The following fig. 3.2 and fig. 3.3 represent original image and averaged image.

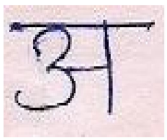

Figure 3.2: Original Image

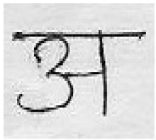

Figure 3.3: Averaged Image

2) Image Binarization

It allows us to reduce the amount of image information (removing colour and background), so the output image is black-white. The black-white type of the image is much more easily used for further processing. Fig. 3.4 represents binarized image.

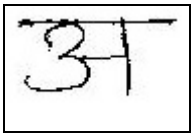

Figure 3.4: Binarized Image

3) Opening

The character area is alienated from the background by using the well known segmentation methods of vertical and horizontal projection. Thus, the white space surrounding the character is discarded. Morphological operation viz Erosion and Dilation are applied to perform this step. Fig. 3.5 represent images processed for Erosion and Dilation respectively.

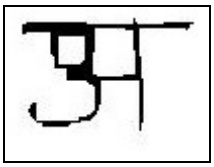

Figure 3.5: Opening

4) Normalization

It is necessary because the characters written by hand vary greatly in size and shape. To account for variety in shape, the character is normalized.Fig.3.6 represent normalized image.

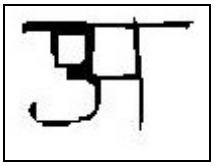

Figure 3.6: Normalization

5) Thinning

Size of the image is abridged. In this procedure, unnecessary character areas are removed. To find the features of the objects, the boundary detection of the image is done. Fig. 3.7 represent thinned image.

Figure 3.7: Thinning

C. Feature Extraction Techniques

In this stage, the features of the characters that are used for classifying them at recognition stage are extracted. This is an important stage as its effective functioning improves the recognition rate and reduces the misclassification.

In feature extraction, we will extract features of character using Hu's moment invariant, zoning with Hu moment and Zoning with Hu's moment and radon transform. .

1) Hu's Moment Invariant

Moment invariants are firstly introduced by Hu. Hu's Uniqueness Theorem states that if is piecewise continuous and has nonzero values only in the finite part of the plane, then statistical moments of all orders exist. Moments and functions of moments have been extensively employed as invariant global features of images in pattern recognition. Moment set can be computed and used to exceptionally explain the information contained in the image segment. Hu derived six absolute orthogonal invariants and one skew orthogonal invariant based upon algebraic invariants, which are independent of position, size and orientation. For our character image we first apply zoning and then Hu's moment on original character and extract features from them and store it in database. Hu's moment gives 7 features which is independent of scale, position and orientation. Therefore we apply this moment on preprocessed characters; so that character becomes invariant to scale, position and orientation.

2) Zoning

Foreground pixel distribution

Suppose that im(x,y) is a handwritten Devanagari character image in which the foreground pixels are denoted by 1's and background pixels are denoted by 0's. Feature extraction algorithm sub-divides the character image recursively. At granularity level 0 the image is divided into four parts and gives a division point (DP) (x_0, y_0). The following algorithm I shows that how x_0 is calculated and likewise y_0.

Algorithm I

Step 1: Input im (x_{max}, y_{max}) where x_{max} and y_{max} be the width and the height of the character image

Step 2: Let v_0 [x_{max}] be the vertical projection of image (fig 3.7)

Step 3: Create v_1 [2*x_{max}] array by inserting a '0' before each element of v_0

Step 4: Find x_q in v_1 that minimizes the difference between the sum of the left partition [1,x_q] and the right partition [x_q, 2*x_{max}] or left partition should be greater than right if not able to equally divide.

Step 5: x_0=x_q / 2;

Step 6: If x_q mod 2 = 0

Two sub-images [(1, 1), (x_0, y_{max}) and (x_0, 1), (x_{max}, y_{max})]

Else

Two sub-images are [(1, 1), (x_0, y_{max}) and (x_0+1, 1), (x_{max},y_{max})]

The number of sub-images, at the specified granularity level (L) will be $4^{(L+1)}$. Let L=0 then the number of sub-images are 4 and when the L=1 it will be 16. The number of DP (Division Point) equals to 4^L. After dividing image into number of zones, again we will apply Hu moment on each zone for getting more fine features.

We will combine all features extracted from Zoning, Hu's moment invariant. Ant miner classifier uses these features for classification.

D.Classification

In character recognition system, features of stored character images are compared with features of character to be recognized. When maximum matching is found, I/P character is recognized as that particular stored character. Our proposed algorithm for character recognition is Ant Miner Algorithm. So before starting with Ant miner algorithm, it is very necessary to understand the Ant colony optimization.

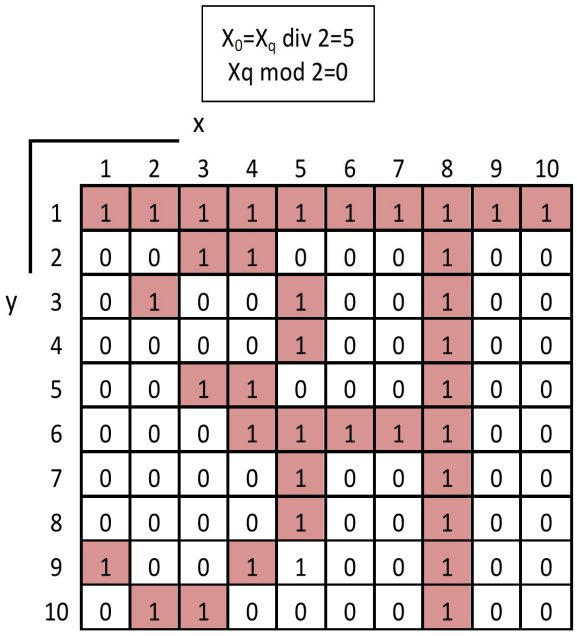

Figure 3.8: Vertical division of an image array (x_{max}=10,y_{max}=10)

1) Ant Colony Optimization

An Ant Colony Optimization algorithm (ACO) is essentially a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation. In a colony of social insects, such as ants, bees, wasps and termites, each insect usually performs its own tasks independently from other members of the colony.

In this paper we are interested in a particular behavior of real ants, namely the fact that they are capable of finding the shortest path between a food source and the nest (adapting to changes in the environment) without the use of visual information. This intriguing ability of almost-blind ants has been extensively studied by ethologists. They discovered that, in order to exchange information about which path should be followed, ants communicate with one another by means of pheromone (a chemical substance) trails. As ants move, a certain amount of pheromone is dropped on the ground, marking the path with a trail of this substance. The more ants follow a given trail, the more attractive this trail becomes to be followed by other ants.

2) Ant-Miner: A New ACO Algorithm for Classification

The AMA was proposed by Parpinelli and his Colleagues. It applies the ant colony optimization heuristic for the classification which is used to discover an ordered list of classification rules. To discover classification rules of the form: IF (Attribute1= val1 AND Attribute2= val2 AND...Attributem= value) THEN (predicted class). In our case all the different rules can have equal number of terms in their antecedent (the IF part). The consequent of a rule is a predicted class (the THEN part), that is the value that the rule predicts for the class attribute when an example satisfies the conjunction of terms in the rule antecedent. Classification rules have the advantage of representing knowledge at a high level of abstraction. A high-level description of Ant-Miner is shown below.

Algorithm II

Training set = all training cases;
Rule list = empty;
REPEAT
i=0;
Pheromone Initialization;
REPEAT
i=i+1;
Anti constructs a classification rule;
Prune the current constructed rule;
Update the pheromone of the trail followed by Anti;
The best rule is memorized;
UNTIL (i = No_of_Ants) or (Anti constructed the same rule as the previous Ants continually No_Rule_Converge times)
The best rule is added to the rule list;
Remove the cases covered by the selected rule from the training set;
UNTIL (Number of cases in the Training set less than Max_uncovered_cases)

Following the algorithm II; after the pheromone is initialized, many rules are constructed in the inner Repeat-Until loop with the rule pruning and the pheromone updating method. The loop will stop when ants construct the same rule continually more than No_Rule_Converg times or the number of rules is equal to the number of ants. When the inner Repeat-Until loop is completed, the best rule will be added to the rule list.

Then, all training cases which are predicted by this rule are removed from the training case set. Pheromone is initialized again. This cycle is controlled by the Outer Repeat-Until loop. The Repeat-Until loop will finish when the number of

uncovered training cases is less than a threshold, called Max_uncovered_cases.

a) Pheromone Initialization

The pheromone values are initialized for each character as per the followings at time t =0.

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^{a} b_i} \quad (1)$$

Where a is the total number of attributes, i is the index of an attribute, j is the index of a value in the domain of attribute i, and b_i is the number of values in the domain of attribute i.

b) Pheromone Updation

The pheromone levels are updated for all terms in a rule by an Ant and it is based upon the quality Q of that rule. The quality value is measured as "sensitivity × specificity", which is defined as follows:

$$Q = \frac{TN}{FP+TN} \cdot \frac{TP}{TP+FN} \quad (2)$$

Here TP denotes True Positive, FN means False Negative, TN determines True Negative, FP denotes False Positive and TP / (TP + FN) is the sensitivity; TN / (FP + TN) is the specificity. Once a rule has been accepted, the amount of pheromone increment to be done to each of the terms in that rule, which is determined by the following formula.

$$\tau_{ij}(t+1) = \tau_{ij}(t) + \tau_{ij}(t).Q \forall_{i,j} \in R \quad (3)$$

Where, the R_t is the set of rules generated for a single script during training and R_{best} stores the best rule for the script after the completion of the training. Since in here we deal with a static scenario, the pheromone evaporation step is merely skipped to reduce the complexity.

A GUI based software tool has been developed using Matlab 2009a, ver.7.9 for the Optical Character Recognition, especially for handwritten Marathi scripts. The PC specification is as follows:

- MS Windows 8
- Pentium ® Dual-Core CPU with 2 GB RAM
- Processor speed 3.00 GHz.

The handwritten scripts are collected from twenty users, 126 characters from each user. Hence 2520 scripts are available. The system is trained with 2275 characters.

The handwritten character images are collected using a 300 dpi digital scanner. Hence N = 2275 i.e., the maximum number of characters used during training. For testing we are collected 5 samples of each character, so total 49*5=245 characters are available for the testing. For each character the ant system is initialized. Using the feature extraction process as discussed above, the features of character are prepared. An ant uses these features to recognize the character.

IV. RESULTS AND DISCUSSION

Experimental results are shown in this section. The Figure 4.1 shows the feature matrix of the trained characters. Using these feature ant recognize the character. figure 4.2 shows how character is recognized as the best match among all of the good sketches using the proposed Any miner Algorithm recognition method. Rank 1 accuracy gives best match. For each image, we have listed the match rank. As the rank increases accuracy will decreases.

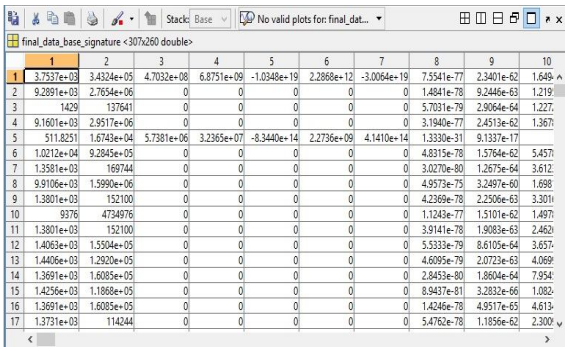

Figure 4.1: Feature matrix of trained character

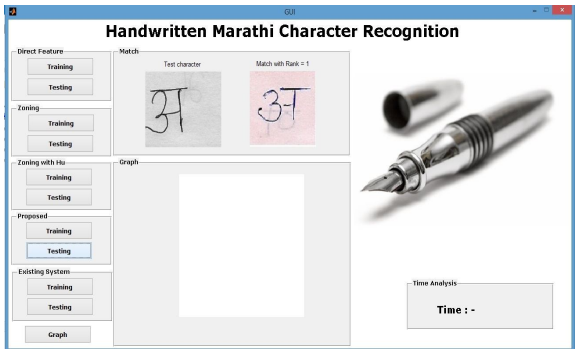

Figure 4.2: Rank 1 accuracy of testing image using proposed system

The result is show in Table I and Table II compares the recognition rate of the proposed system with the affine moment.

TABLE I
RESULT OF HANDWRITTEN DEVNARARI CHARACTER TESTING WITH AN ANT-MINER CLASSIFIER

Group of characters	No. of characters	Rank Accuracy in %	
		Training Data	Testing Data

Vowels	720	97.22	83.33
Consonants	1800	96.66	81.09
Average		96.94	82.21

TABLE II
RECOGNITION RATE OF PROPOSED SYSTEM AND AFFINE MOMENT

Recognition	Recognition Rate
Proposed System	82.21
Existing System	Less than 60%

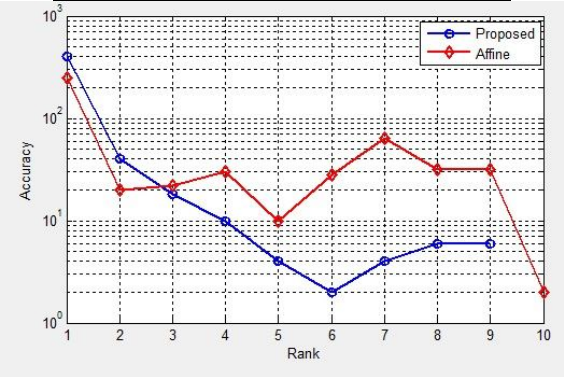

Figure 4,3: Performance comparison of Proposed system and existing system

CONCLUSION

The technique discussed above is a novel technique as it involves the rule constructions which have not been used much for recognition especially for Devnagari Characters. The Devnagari involves 12 Vowels and 37 Consonants. Thus we have lot of problems associated with the recognition of these characters. The technique proposed here can be used to obtain a higher recognition rate as compared to the implemented methodologies. The use of the ant miner algorithm will also prove a novel idea for the classification and yielding best results for the characters. In future, we can plan to experiment on recognition of compound character to get higher recognition accuracy from our system.

REFERENCES

[1] P.E.Ajmire, RV Dharaskar, V M Thakare, "A Comparative Study of Handwritten Marathi Character Recognition", In Proc. National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012)

[2] Shreya N. Patankar Leena R. Ragha, "Zonal moments based Handwritten Marathi Barakhadi recognition" In Proc. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012

[3] M. Hanmandlu and O.V. Ramana Murthy, "Fuzzy Model Based Recognition of Handwritten Hindi Numerals", In Proc. International Conference on Cognition and Recognition, pp. 490-496, 2005.

[4] M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madasu, "Fuzzy Model based recognition of Handwritten Hindi characters", IEEE Computer society, Digital Image Computing Techniques and Applications , 2007

[5] S. Kumar and C. Singh, "A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition", In Proc. International Conference on Cognition and Recognition, pp. 514-520,2005.

[6] V. Bansal, "Integrating Knowledge Sources in Devanagari Text Recognition", Ph.D.Thesis, IIT Kharagpur, 1999.

[7] V. Bansal, R.M.K. Sinha, "Partitioning and Searching Dictionary for Correction of Optically Read Devanagari Character Strings", Proc. 5th Int. Conf. Document Analysis and Recognition, Bangalore, India, Sept. 20-22, 6pp. 53-656,1999.

[8] V. Bansal, R. M. K. Sinha, "On How to Describe Shapes of Devanagari Characters and Use them for Recognition", Proc. 5th Int. Conf. Document Analysis and Recognition, Bangalore, India, Sept. 20-22, pp. 410-413,1999.

[9] R. Bajaj, L. Dey, S. Chaudhary, "Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers", Sadhana, 27, pp. 59-72, 2002.

[10] P.M. Patil, T. R. Sontakke, "Rotation, Scale and Translation Invariant Handwritten Devanagari Numeral Character Recognition using General Fuzzy Neural Network", Pattern Recognition, vol. 40, pp. 2110-2117, 2007.

[11] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Off-line Handwritten Character Recognition of Devanagari Script", Proc. 9th Int. Conf. Document Analysis and Recognition, Parana, Sept.23-26, pp. 496-500,2007.

[12] U. Pal, S. Chanda, T.Wakabayashi, F. Kimura, "Accuracy Improvement of Devanagari Character Recognition Combining SVM and MQDF", Proc. 11th Int. Conf. Frontiers in Handwriting Recognition, Qu´ebec, Canada, Aug. 19-21, pp. 367-372,2008.

[13] U. Pal, T.Wakabayashi, F. Kimura, "Comparative Study of Devanagari Handwritten Character Recognition using Different Feature and Classifiers", Proc. 10th Int. Conf. Document Analysis and Recognition, Barcelona, Spain, Jul. 26-29, pp. 1111-111,2009.

[14] N. Sharma, U. Pal, F. Kimura, and S. Pal, "Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier", Springer, 2006

[15] S. Shelke, S. Apte, "A Novel Multi-feature Multi-Classifier Scheme for Unconstrained Handwritten Devanagari Character Recognition", Proc. 12th Int. Conf. Frontiers in Handwriting Recognition, Kolkata, India, Nov. 16-18, pp. 215-219,2010.

[16] Sandhya Arora, D. Bhattacharjee, Mita Nasipuri, "Recognition of Non-Compound Handwritten Devnagari Characters using a Combination of MLP and Minimum Edit Distance", IJCSS Vol.4 Issue 1.

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

[18] Brijmohan Singh, Ankush Mittal, Debashis Ghosh,"An Evaluation of Different Feature Extractors and Classifiers for Offline Handwritten Devnagari Character Recognition", Journal of Pattern Recognition Research 2 (2011) 269-277, Received December 12, 2009. Accepted September 14, 2011.

[19] H. Ming-Kuei, "Visual pattern recognition by moment invariants" Information Theory, IRE Transactions, vol. 8, pp.179-187, 1962.

[20] Zhihu Huang, Jinsong Leng "Analysis of Hu's Moment Invariants on Image Scaling and Rotation" 2010 2nd International Conference on Computer Engineering and Technology, Volume 7 , IEEE, 2010.

[21] Sandeep Patil, Shailendra Dewangan "Neural Network-based Offline Handwritten Signature Verification System using Hu's Moment Invariant Analysis" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-1, October 2011

[22] Nada M. A. Al Salami "Ant Colony Optimization Algorithm", UbiCC Journal, Volume 4, Number 3, pp. 823-826, 2009.

[23] Smaldon J., Freitas A. A. "A New Version of the Ant-Miner Algorithm Discovering Unordered Rule Sets", GECCO'06, Seattle, Washington, USA, July 8–12, 2006.

[24] Parpinelli R.S., Lopes H.S., Freitas A.A. "An ant colony algorithm for classification rule discovery". In Abbas H., Sarkar R., and Newton C. (Editors.). Data Mining: a Heuristic Approach, London: Idea Group Publishing, pp. 191-208.

[25] Nilima P. Patil K. P. Adhiya Surendra P. Ramteke, "A Structured Analytical Approach to Handwritten Marathi vowels Recognition"in Proc. International Journal of Computer Applications (0975 – 8887), Volume 31– No.3, October 2011.

★★★