Fine Classification & Recognition of Hand Written Devnagari Characters with Regular Expressions & Minimum Edit Distance Method

First Dr. P. S. Deshpande
V.N.I.T., Nagpur, India
Email: psdeshvnit@yahoo.co.in

Second Latesh Malik
G.H.Raisoni College of Engineering, Nagpur,Maharashtra ,India
Email: lateshmalik@rediffmail.com

Third Sandhya Arora
Meghnad Saha Institute of Technology, Kolkata ,India
Email: sandhyabhagat@yahoo.com

***Abstract*— Regular expressions are extremely useful, because they allow us to work with text in terms of patterns. They are considered the most sophisticated means of performing operations such as string searching, manipulation, validation, and formatting in all applications that deal with text data. Character recognition problem scenarios in sequence analysis that are ideally suited for the application of regular expression algorithms. This paper describes a use of regular expressions in this problem domain, and demonstrates how the effective use of regular expressions that can serve to facilitate more efficient and more effective character recognition.**

***Index Terms*— Regular expression, edit distance method, feature extraction, preprocessing**

I. INTRODUCTION

An OCR has a variety of commercial and physical applications in reading forms, manuscripts and their archival etc.Such a system facilitates a keyboard less user computer interaction. Also the text which is either printed or handwritten can be directly transferred to the machine. It becomes a reading machine for the visually handicapped when interfaced with a voice synthesizer. An elaborate list of OCR applications has been presented by Govindan[35].

Development of OCR system for the Indian language script are gaining importance because of their large market potential .With the explosion of information technology there has been a dramatic increase of research in this field since the beginning of 1980.

A number of strategies for character recognition are available in literature. The recognition of digits, a subfield of character recognition, is subject of much attention since the first years of research in the field of handwriting recognition. However, research in this field has basically considered: a) feature extraction methods, b) classification method, and c) system architectures based on different strategies, such as combination of multiple classifiers, the use of multiple templates, and the use of verification modules.

The investigation of feature extraction methods has gained considerable attention since a discriminative feature set is considered the most important factor in achieving high recognition performance. Trier et al [25] present an interesting survey of feature extraction method for off-line recognition of segmented characters. The authors describe important aspects that must be considered before selecting a specific feature extraction method.

In general, the feature extraction methods for digit recognition reported in the literature are based on two types of features: statistical and structural. The statistical features are derived from statistical distributions of points, such a zoning, moments, projection histograms, or direction histograms [17,24,35]. Structural features are based on topological and geometrical properties of the character, like strokes and their directions, end-points, or intersections of segments, and loops[9,33]. Many researchers have explored the integration of structural and statistical information to highlight different character properties, since these types of features are considered to be complementary [7,9].

Alongside these investigation of feature extraction methods, many other studies have addressed classification methods. Different classifiers have been used for handwritten digit recognition, such as statistical [28]structural [33] and neural networks [37]. Recently, significant contributions towards the improvement of recognition rates have been made by means of different

combination strategies [4],and the use of support vector machines [19].

However, when we talk about the recognition of Devnagari character, the problem becomes more complicated. Character recognition is complicated by presence of multiple loops, conjuncts, upper and lower modifiers and the number of disconnected and multi-stroke characters[18],in a word where all characters are connected through header line. In fact, we can say that character recognition is still an open problem.

In this paper, we propose a scheme for unconstrained off-line handwritten Devnagari character recognition using regular expressions, minimum edit distance method, based on features obtained from chain code. Here the character is converted into chained code encoded string and regular expressions are matched with it. Rejected samples are sent to minimum edit distance classifier. Rest of paper is organized as follows. In section 2,3 ,we discuss about Devnagari language, its character set and work done so far in this direction. Section 5 discuss about feature extraction, method of regular expression creation, matching and minimum edit distance method. Experimental results are discussed in section 6.

II. FEATURES OF DEVNAGARI SCRIPT

Devnagari is the script for Hindi which is official language of India. It is also the script for Sanskrit, Marathi and Nepali languages. Devnagari script has 50 characters which can be written as individual symbols in a word(as shown in figure 1).

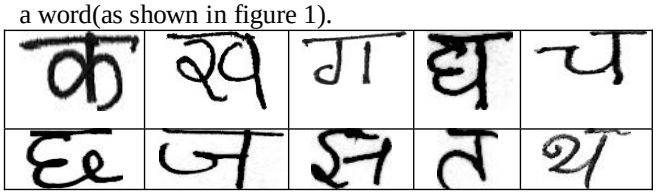

Figure 1: Samples of handwritten Devnagari basic characters (Consonants)

The characters may also have a *half* form. A half character in most of the cases is required to touch the following character, resulting in a composite character. Some characters of Devnagari script take the next character in their shadow because of their shape(as shown in figure 2).

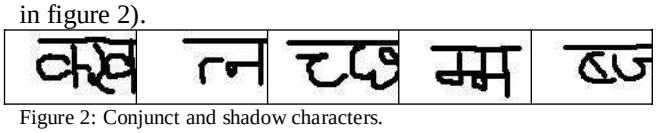

Figure 2: Conjunct and shadow characters.

The script has a set of modifier symbols which are used only to modify a character. These symbols are placed either on top, at the bottom, on the left, to the right or a combination of these.

Top modifiers are placed above the ***shirorekha***, which is a horizontal line drawn on the top of the word. The lower modifiers are placed below the character which may or may not touch the characters. More than one lower modifier may also be placed below one character. A character may be in shadow of another character, either due to a lower modifier or due the shapes of two adjacent characters [35](as shown in figure 3).

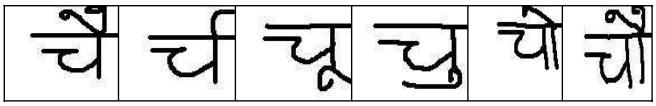

Figure 3: Upper and lower modifiers with basic character CH

Modifiers make Optical Character Recognition (OCR) with Devnagari script very challenging. OCR is further complicated by compound characters that make character separation and identification very difficult.

III. DEVNAGARI OCR

First research report on handwritten Devnagari characters was published in 1977 [15] but not much research work is done after that. At present researchers have started to work on handwritten Devnagari characters and few research reports are published recently.

Sinha et al [27,28] have reported various aspects of Devnagari script recognition. Sethi et. al. [10] have described Devnagari numeral recognition based on structural approach.The primitive used are horizontal line segment, vertical line segment , right slant and left slant. A decision tree is employed to perform analysis based on presence/absence of these primitives and their interconnection. A similar strategy was applied to constrained hand printed Devnagari character [9]. Neural network approach is used for isolated characters have also been reported [11].

Bansal et. Al. [1], have used translation & scaling invariant moments and structural description of a character and reported accuracy of 93% at character level of printed Devnagari characters. Bhattacharya et al[34] proposed a Multi-layer perceptron (MLP) neural network based classification approach for the recognition of Devnagari handwritten numerals and obtained 91.28% results. They considered a multi- resolution features based on wavelet transform in their proposed system.

IV. PREPROCESSING

Histogram based global binarising algorithm is used to convert them to two tone(0 and 1). In preprocessing header lines are completed as straight lines(as shown in figure 4).Detection of header line is based on

No-of-ones-row>=.50*col and pos_header >=.50*row

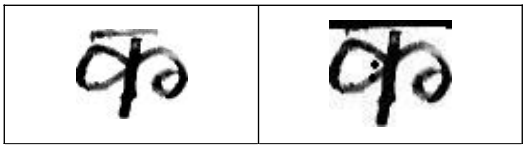

Figure 4:Completed header line

Some of vertical lines are not completed and due to unequal thickness of vertical lines are not completed and due to unequal thickness of vertical lines features may not be properly extracted. Columns of unequal thickness of vertical line can be detected as (No_of_ones_col >=.80*row) and completed(as shown in figure 5).

Figure 5: Completed vertical line

Noise at isolated locations can be detected as it is not connected with character and length of noise is also negligible in comparison with any connected curves or lines in a character(as shown in figure 6).

Figure 6: Removal of isolated dot near header line

V. PROPOSED TECHNIQUE

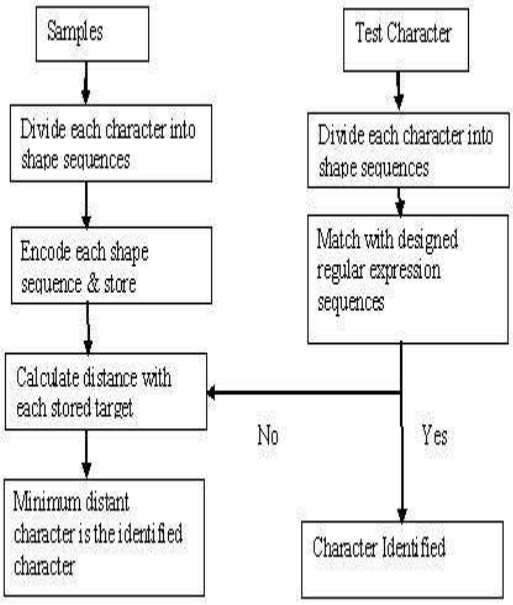

Figure 7: Components of proposed technique

A. Direction Features

Direction features have been used for many scripting recognition system and tested using neural network and HMM based classifiers[20].

Directional features are used in proposed system with the use of segments and type of connectivity is the representation of direction. Proposed method consider a shape and is converted into 1's and 0's. The segments are defined of continuous 1's in one column .Shape is scanned from top to bottom and left to right and is encoded according to the connectivity between segment of one column with the previous column. Information stored for particular segment is : staring position of segment, end point of segment, length of segment, column number, connectivity code, sequence number(as shown in figure 8)

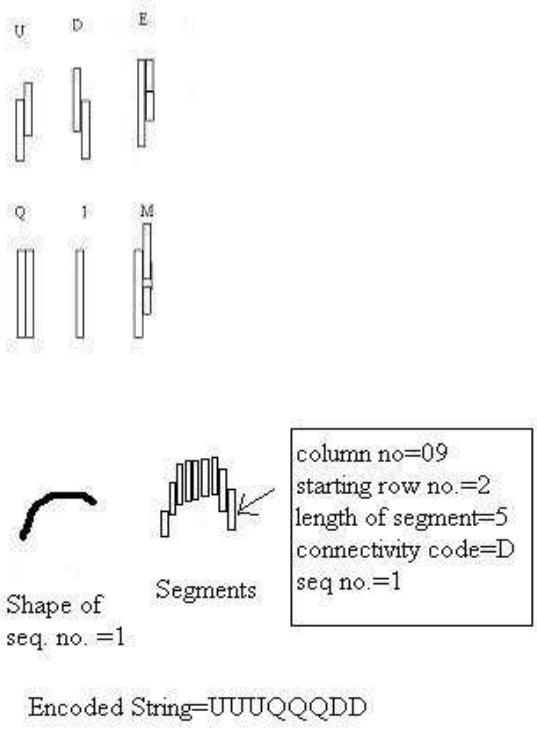

Figure 8: Operators used , extracted features

B. Shape Analysis and Extraction Of Various Shapes In a Character For Efficient Regular Expression Matching

In previous paper[5] , we introduced the use of regular expressions in character recognition. Character was encoded into string and regular expressions were created on full string representing full image of a character , main and common patterns are found in encoded string[5]. Problem faced by previous method that features were merging of upper and lower half, recognition of character as more than one and rate of character overlapping was more.

To overcome the above problems strokes and shapes are extracted (as shown in figure 9) based on connectivity analysis using the following algorithm.

Initialize
Sequenceno=1;
Scan segments sequentially from top to bottom and left to right .
For each segment repeat step 4
If Starting position of a segment in one column- Starting position of a segment in previous column)<=threshold value(may be taken as 5 estimated from 5000 samples) , assign same sequenceno to a segment ;
Else
sequenceno= sequenceno+1
Assign new sequenceno to the segment.

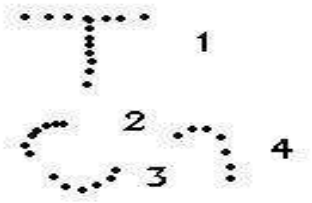

Figure 9: Extraction of shapes in a character

Sample string sequences of character KA with 4 sequences is shown below.

ka1.jpg,4,
1.DUDUQQQQQQUQQQQQQQQDQQQQUQQEEDQ DQQQQQQQQQQQQQQQQ,
2.UUUEUUUUDDUQDQQQUUDDU,
3.DDDQQQQQUUUQUUUU,
4.UQDQQQQUUDDDDDDI

C. Regular Expressions

Regular expressions are extremely useful, because they allow us to work with text in terms of patterns. They are considered the most sophisticated means of performing operations such as string searching, manipulation, validation, and formatting in all applications that deal with text data.

Regular expressions are represented as a character strings. Character strings and regular expressions are so closely related, the regular expression type should "look like" a character string. Use of regular expressions is already found in literature for reconfigurable packet inspection[23], characterizing human gene splice sites[15], creating protein sequence patterns[22], network intrusion detection [8] etc.

Regular expressions are never used in character recognition . Approximate string matching by regular expressions can be used to cater the variations in handwritten Characters, because we live in an error-prone world.

D. Feature Sequence Representation

Character is converted into sequences(as shown in figure 9) and then sequence of geometrical properties of the character ,like strokes and their directions, end points, or intersection of segments, and loops can be denoted with regular expressions. If a character have sequence like loop, intersection, loop and then end point can be represented as

[^loop]*loop{1}[^intersection]*(intersection)+
[^loop]*loop{1}[^end point]*(end point) +

E. Character Recognition

In any character recognition problem, structural features are based on topological and geometrical properties of the character ,like strokes and their directions, end points, or intersection of segments, and loops. Regular expression represents the above structural features as

strokes and their directions are encoded character for particular direction and + sign(one or more occurrence).

end points are represented as encoded character of the particular end point and {1} (only one occurrence)

intersection of of segments with header line, vertical bar, loops etc are represented as particular encoded character and as many time as intersections points occur in sequence.

loops are represented by division of one segment into two (E), encoded into single character and {1} (only one occurrence)

Joint Representation

Joints with header line: This type of one joint has regular expression representation of starting and ending of joint by at least one occurrence of connectivity code representing expanding connectivity(U) and shrinking connectivity(D).

Joints with vertical bar **:**This type of joint can be starting of stroke or ending of stroke. Ending stroke has at least one occurrence of expansion connectivity code(U).

Joints with loops**:** Treated in the same way as with vertical bar.

In this way regular expressions can be designed manually for each character depending on shapes and joints present in each character .Operators for connectivity is shown in figure 8 and used in table 1.Regular expressions for some of the characters are shown in table 1.

Table 1: Regular expressions for some Devnagari characters

Characters	No. of sequences	Regular expression
	3	1.[^U]*U+[^E|^M]*(E|M)+[^W]* 2.U+(E|D|M)+[^W]* 3.[^DD]*(DD)+[^QQ]*(QQ)+[^UQ*U]*(UQ*U)+[^DI]*(DI)+[^W]*
	2	1.[^U]*U+Q*(E|D)+Q*[^W]* 2.U+Q*[E|D|U]+Q*D*Q*U+[^W]*
	2	1.[^U]*U+Q*([E|D]+Q*)+U+Q+[E|D]+Q+[^W]* 2.[*QQ]*(QQ)+Q*U+[^W]*
	4	1.[^U]*(UQ*)+[^E|^D|^M]*(E|D|M)+Q+[^W]* 2.U+E+[^W]* 3.[^D+Q+U+]*(D+Q+U+)+[^W]* 4. [^DDDI]*[^DDDI][^W]*

Table 2: Representation by regular expression for sequence of shapes(as shown in figure 9) in character KA

Part No.	Representation
1	At least one joint with header line(U+D+)
2	Directional curve upward and starting of loop(U+E+)
3	Directional curve downwards(D+)
4	At least one end of curve(I)

F. Minimum Edit Distance Method

Regular expression matching results in accuracy of 70-75%.To improve the accuracy of the system remaining 25-30% of samples which cannot be identified with regular expressions are tested with minimum edit distance method.

Distance is a measure of similarity between two strings which is referred as source string (s) and target string(t). The distance is the number of deletions, insertions or substitutions required to transform s into t. Distance is calculated by

```
cost := 0    if s[i] = t[j]
cost := 1    if s[i] <> t[j]
d[i, j] := minimum(
 d[i-1, j] + 1,    // deletion
 d[i, j-1] + 1,    // insertion
 d[i-1, j-1] + cost   // substitution
)
```

10 samples of each handwritten character are taken as the target string and source string is matched with all target strings and distance is calculated. Minimum distant target is the identified character.

In the proposed technique character is divided into shapes and sequence of shapes is also maintained. Distance has to be calculated for each shape sequence in one character and summation of each will define the total distance.

$$\text{A target character} = \sum_{i=0}^{i=n} \text{shapet}(i)$$

$$\text{A source character} = \sum_{i=0}^{i=m} \text{shapes}(i)$$

$$\text{distance} = \sum_{i=0}^{i=\max(m,n)} d[i,j](\text{shapes}(i)-\text{shapet}(i))$$

VI. EXPERIMENTAL RESULTS

Data used for the present work was collected from different individuals. From experiments we noted that the overall recognition accuracy of the proposed scheme for characters is 82% for 50 writers. Experiment has been conducted of 5,000 characters with 50 writers. Results are provided in table for 5 writers. Characters are first identified with regular expression matching, if they do not match with any pattern they are passed to minimum edit distance filter. Proposed method is tested for 50 different characters ,50 writes and results are shown in table 4 .

95% accurarcy is achieved for characters like न, ज, ढ where the number of loops, connectivity with header line, end of loops are less in comparison with other characters. We also noticed more misclassification of characters like ध, रव, ए, इ due to more sequence of shapes in each character.

Our results are very promising indicating that recognition rate can be increased if the number of samples for minimum edit distance are increased. In future experiments we hope to use larger database for testing and samples.

Table 3: Result showing recognition rate for 5 writers and 500 samples

Writers-> No. of samples	1 100	2 200	3 300	4 400	5 500
Recognition rate with regular expression matching	85	164	240	300	375
Recognition rate with minimum edit distance method (Remaining samples)	13	24	36	68	75
Total recognition rate	98%	94%	92%	92%	90%

We obtained 82% accuracy using 5,000 samples of data ,but Kumar & Singh[32] reported 80% accuracy and tested 200 data samples, N. Sharma et. al. tested 11270 samples and reported accuracy of 80.36% but the number of samples required for training are much higher. In our system 70-75 % samples recognized with regular expressions where no training is required and for minimum edit distance method only 10 samples of each character is stored for distance calculation. The details about the comparison results are given in table 4.

Table 4: Comparison results for Devnagari characters

Sr.No	Method proposed by	Data size	Accuracy obtained
1	Kumar & Singh[32]	200	80%
2	N. Sharma et al[23]	11270	80.6%
3	Proposed method	5000	82%

VII. CONCLUSION

India is a multilingual and multi script country comprising of twelve different scripts. But not much work has been done towards offline handwriting recognition. Proposed technique is more powerful as such there is no need of scaling and thinning. This technique is never tried in the field of character recognition and document analysis. From the experiment we obtained encouraging results.

References

[1] Bansal V, Sinha R. M. K., Integrating Knowledge Resources in Devnagri. Text recognition system, IEEE Transaction on System, Man & Cybernatics Part A: Systems & Humans. V30 n 4 july 2000.p 500-505

[2] Britto Jr. R. Sabourin, F. Bortolozzi, and C. Y. Suen (2001). A two-stage hmm based system for recognizing handwritten numerals strings. In proc. 6th international conference on document analysis and recognition, Seattle-USA September, PP.396-400.

[3] C –L. Liu, K. Nakashima, H. Sako, and H. Fujisawa,Handwritten digit recognition using state-of –the-art techniques. In Proc of 8th International Workshop on Frontiers of Handwritings Recognition. pp. 320-325, 2002.

[4] C.Y. Suen, K Diu, N. W. Strathy ,Sorting & recognizing cheque and financial documents, Document Analysis System Theory and Practice, Springer, Belin pp 173-187.

[5] Deshpande P. S.,Malik Latesh, Arora Sandhya, Characterizing Hand Written Devnagari Characters using Evolved Regular Expressions TENCON 2006, Page(s):1 – 4

[6] EI Yacobi, M. Gilloux, R. Sabourin, C. Y. Suen (1999). An hmm-based approach for off-line unconstrained handwritten word modeling and recognition. IEEE Transactions on pattern Analysis and machine intelligence, 21 (8): 752-760.

[7] Heute L. T. Paquet, J. V. Moreall, A Structural /Statislical feature base vector for handwritten character recognition, Pattern Recognition Letters, 19, pp. 629-641, 1998.

[8] Hutchings, B.L., Franklin, R., Carver, D., Assisting network intrusion detection with reconfigurable hardware,10thAnnual IEEE Symposium on 22-24 April 2002 Page(s):111 – 120

[9] J. Cai, Z.Liu ,Integration of structural and statistical information for Unconstrained Handwritten Numeral Recognition. Proceedings of the Fourteen International conference on pattern Recognition, Vol. 1. Pp.378-380, 1998.

[10] J. X. Dong. A. Krzyzak, C.Y. Suen ,A multi-net learning framework for pattern recognition. Proc. Of the sixth International conference on Document analysis and Recognition, Seattle, pp. 328-332, 2001.

[11] J.C. Sant and S. K. Mullick, Hand written Devnagri script recognition using CTNNSE algorithm,International Conference on Application of Information Technology in south Asia Language, Feb 1994.

[12] J.X. Dong. A. Krzyzak, C. Y. Suen .Local learning framework for recognition of lowercase handwritten characters. In Proc. Int. Workshop on Machine Learning and Data Mining in pattern Recognition ,Leipzig, Germany, pp. 226-238, 2001

[13] Jing-Jing Li; De-Shuang Huang; MacCallum, R.M.; Xiao-Run Wu ,Characterizing human gene splice sites using evolved regular expressions ,IEEE International Joint Conference on Volume 1, Page(s):493 - 498 vol. 1, 31 July-4 Aug. 2005

[14] K. Cheung, D. Yeung, A Bayesian Framework for deformable Pattern Recognition with Application to Handwritten Character Recognition. IEEE Transaction on pattern Analysis & machine intelligence, Vol 20, No.12 , December; pp1382-1388, 1998.

[15] K. Sethi , Machine recognition of Constrained hand printed Devnagri, Pattern recognition, Vol 9 pp69-75, 1977.

[16] K. Sethi and B. Chattarjee, Machine recognition of Handprinted Devnagri Numerals. Journal of Instituions of Electronics & Telecommunication Engineers, India Vol 22, pp 532-535, 1976.

[17] Kimure, M. Shredhar, Segmentation recognition algorithm for hand written numeral strings, Machine Vision Application, No.5, pp 199-210, 1992.

[18] L. Koerich ,Large Vocabulary off-line handwritten word recognition. PhD thesis, Ecole de Technologic Superieure, Montreal-Canada, August 2002

[19] L. S. Oliveira, R. Sabourin ,Support Vector Machines for Handwritten Numerical String Recognition. 9th international workshop on Frontiers in handwriting Recognition, October 26-29, Kokubunji, Tokyo, Japan pp 39-44, 2004.

[20] M. Blumenstein, Y. Liu, B. Verma An Investigation of the modified direction feature for cursive character , Pattern Recognition ,Volume 40, Issue 2, page 376-388,2007.

[21] M. Morita, L. S. Oliveria, R Sabourin ,Unsupervised Feature Selection for Ensemble of Classifiers. IWFHR-9, Pages 81-86, 2004.

[22] Mulder,M.;Nezlek,G.S., Creating protein sequence patterns using efficient regular expressions in bioinformatics research , 28th International Conference 2006 ,Page(s):207 – 212

[23] N.Sharma, U. Pal et al ,"Recognition of off line handwritten devnagari characters using quadratic classifier", ICVGIP 2006,LNCS 4338,pp. 805-816,2006.

[24] Nafiz Arica and Fatos T. Yarman Vural, An overview of Character Recognition Focused on Offline Handwriting,IEEE Transactions on System, Man, and Cybernetics-Part C Vol. 31 No 2 May 2001.

[25] O. D. Trier. A. K. Jain, T. Taxt ,Feature Extraction methods for Character Recognition a survey. Pattern Recognitions, Vol 29, no. 4 pp.641-662, 1996.

[26] P. D. Goder, M. A. Khabou, Automatic Feature Generate for handwritten Digit Recognition. IEEE Transaction on Pattern Analysis & Machine Intelligence, Vol 18, No. 12, pp1256-1261.

[27] R. M.K. Sinha, H. N. Mahabale ,Machine Recognition of Devnagri Script , IEEE Transaction on System ,Man & Cybernetics, Vol SMC- 9, 435-441, 1979.

[28] R. M.K. Sinha, Rule based Contrival Post processing for Devnagri text recognition, Pattern Recognition 20(5), pp. 475-485,1987.

[29] S. Britto Jr. r. Sabourin, F. Bortolozzi, C. Y. Suen ,Foreground and Background information in an HMM-Based Method for Recognition of isolated Characters and Numerals String, 9th International workshop on Frontiers in handwriting Recognition, October 26-29, 2004, Kokubunji, Tokyo, Japan, pp371-376

[30] S. Oh, C. Y. Suen ,"Distance Features for neural network-based recognition of handwritten Characters". International Journal on Document Analysis and recognition. 1(2): 73-88, 1998.

[31] S. S. Marwah, S. K. Mullick and R. M. K. Sinha , Recognition of Devnagri Character using hierachical binary decision trace classifier ", IEEE international conference on system, Man, Cybernete oct. 1994.

[32] Satish Kumar and Chandan Singh,"A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition", Int Conf. and Recognition ,pp-514-520,2005.

[33] T. Hirano, Y. Okada, F. Yoda, Structural Character Recognition Using Simulated Annealing. Proceedings of the fourth International conference on document Analysis and Recognition, Vol. 2. pp.507-510, 1997.

[34] U. Bhattacharya, B.B. Chaudhari, R.Ghosh, and M.Ghosh, "On recognition of Handwritten Devnagari Numerals", In proc. Of the workshop on Learning algorithms for pattern recognition (in conjunction with the 18th Australian Jopint Conference on Artificial Intelligence), Sydney, pp. 1-7, 2005

[35] V. K. Govindan, A. P. Shivprasad, Character Recognition A Review, Pattern Recognition, Vol 23, No. 7 , pp 671-683, 1990.

[36] Veena Bansal, Integrating Knowledge Sources in Devanagari Text Recognition, PhD thesis, Indian Institute of Technology, Kanpur, India, March 1999.

[37] Zhang, M.Fu, H. Yan ,A Modulator Classification Scheme with Elastic Net models for handwritten digit Recognition, Vol. 31, No. 12 pp.1849-1864, 1998.

First Author (M'76–SM'81–F'87) This author has completed Ph.D. from Nagpur University, Nagpur , India and M.Tech. from IIT,Powai, India.

He is currently working as Sr. Lecturer in Department of Computer Science & Engineering, VNIT, Nagpur , India.He is author of books "C & Data Structure", "Data Warehousing using Oracle", "Sql/plSql for Oracle 9i" published by Dreamtech India Pvt Ltd.

Dr. P.S.Deshpande is life member of ISTE, and SAE-India member.

Second Author (M'06) This author became a Member (M) of IEEE in 2006. She has completed M.Tech. (Computer Science & Engineering) from Banasthali Vidyapith, Rajasthan, India and B.E. (Computer Engineering) from University of Rajasthan,India . She is gold medallist in B.E. and M.Tech.

She is currently working as Assistant Professor in Department of Computer Science & Engineering at G.H. Raisoni College of Engineering, Nagpur University, Nagpur, MS, India. She has teaching experience of 8 years.

Mrs. Latesh Malik is life member of ISTE and presented 5 papers in national and 4 papers in international conferences.

Third Author (M'06) She has completed M.Tech. (Computer Science & Engineering) from Banasthali Vidyapith, Rajasthan, India and B.E. (Computer Engineering) from University of Rajasthan,India .

She is currently working as Assistant Professor in Department of Computer Science & Engineering at Meghnad Saha Institute of Technology,Kolkata,WB, India. She has teaching experience of 8 years.

Mrs. Sandhya Arora presented 5 papers in national and 2 papers in international conferences.