Recognition of Printed Devnagari Characters With Regular Expression in Finite State Models

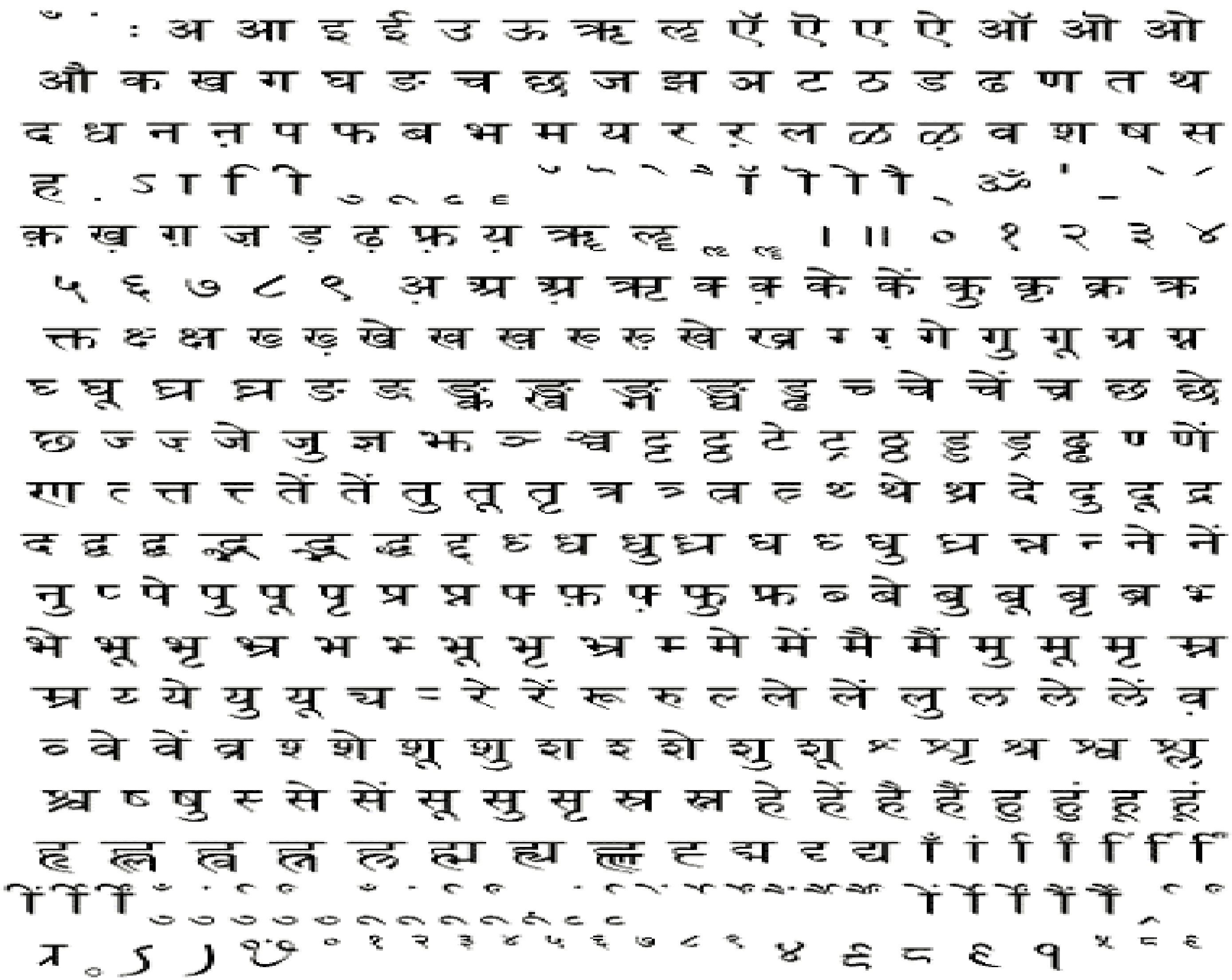
Latesh Malik
Asst. Prof.
G.H.Raisoni College of Engineering, Nagpur
Dr. P.S. Deshpande
Asst. Prof.
V.N.I.T. , Nagpur

Previous Work

Objective:

1. Devnagari is used by number of Indian languages including Sanskrit, Hindi and Marathi. Hindi is the world third most commonly used language after Chinese and English
2. An OCR has a variety of commercial and practical applications in reading forms, manuscripts and their archival etc. Such a system facilitates a key board less user computer interaction. Also the text which is printed can be directly transferred to the machine .

Example Characters

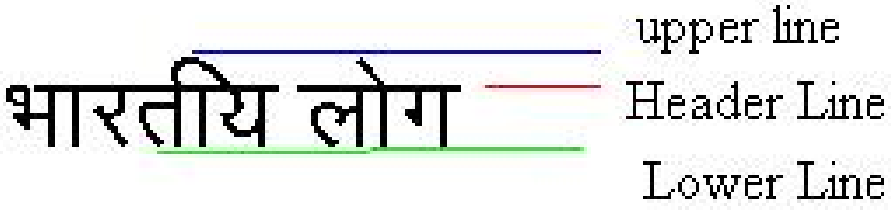


Properties of Devnagari Script

- It is written and read from left to right
- Characters are distinguished by presence of matras
- Matras are dependent vowels used for representing a vowel sound that is not inherent to the consonants.

Properties Of Devnagari Script

- Header lines for words
- Upper modifiers
- Lower Modifiers



Work done for printed Devnagari script

- Chaudhari and Pal (ISI, Kolkata) have developed a devnagari OCR system which is being marketed as custom solution is not yet available as an off the shelf product.
- A feature based tree classifier is used to recognize the basic characters. Error detection and correction for the OCR based on dictionary search has led to the recognition accuracy of 91.25% at word level and 97.18% at character level.

Work done for printed Devnagari script

- Bansal et al proposed Hindi text recognition system by integrating knowledge sources. They achieved accuracy of 87% at character level

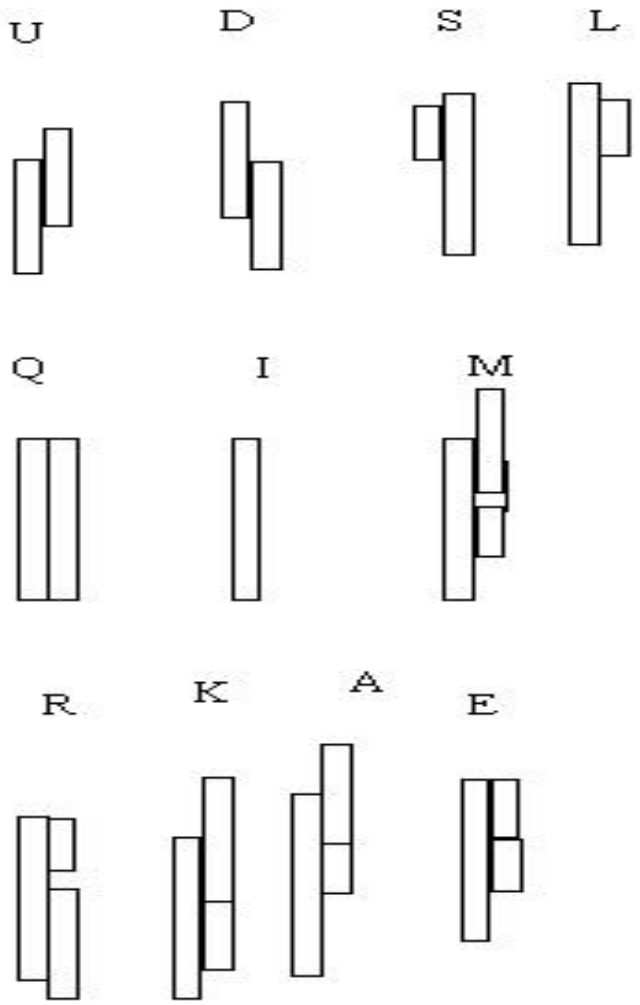
Proposed Approach

- Image is converted into binary form.
- Apply operators to convert into string.
- Shape , joints are extracted in the form of string and design regular expressions.
- Matching of regular expression will recognize a character.

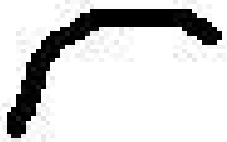
Feature Extraction

- **Joints**
- Joint with vertical line(loop, curve, line joint)
- Joint with header line(loop, curve, line joint)
- **Shapes**
- Ascending curve
- Descending curve
- Straight line
- Division of one curve into two
- Merging of two curves into one
- Starting of one curve from vertical (line, curve)
- Starting of two curves from vertical (line, curve)

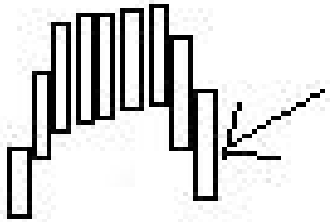
String Operators



Encoding in String



Shape of
seq. no. =1



Segments

column no=09
starting row no.=2
length of segment=5
connectivity code=D
seq no.=1

Encoded String=UUUQQQDD

Regular Expressions

- 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) +

Designing of Regular Expression

QQQQQSQSQSSSSQLKQQQQQQQQQQQQQQQQQQSSQQQLQQQQQQ+(SQ)+

- [^SSS]*(SSS)+ [^K]*K+ Q+ (SS)+Q+LQ+

- QQQQQQSQSQSQYQUDQQDQLDQQLQQQQQLQQQQQQQQQQQQQQSSSQQQSQSLUQLQLQLQIQ
- Q+(SQ)+Y+[^QUD]*(QUD)+[^QQ(D|L)]*(QQ(D|L))+[^SSS]*SSS([^QL]*(QL))+(Q|I)+

Size Invariance

Regular expression created for one size of a character matches with the string generated from any size of the same character even if generated string is not same. Normalization (thinning etc..) of character is not required

Conclusion

- Achieved accuracy of 100% for printed characters
- Use of regular expression in the field of character recognition is found to be fruitful and new concept. In this paper modifiers are not taken into account , and they need to be processed differently.
- It can be applied to other languages as well as handwritten characters

Classifier 2(Edit distance method)

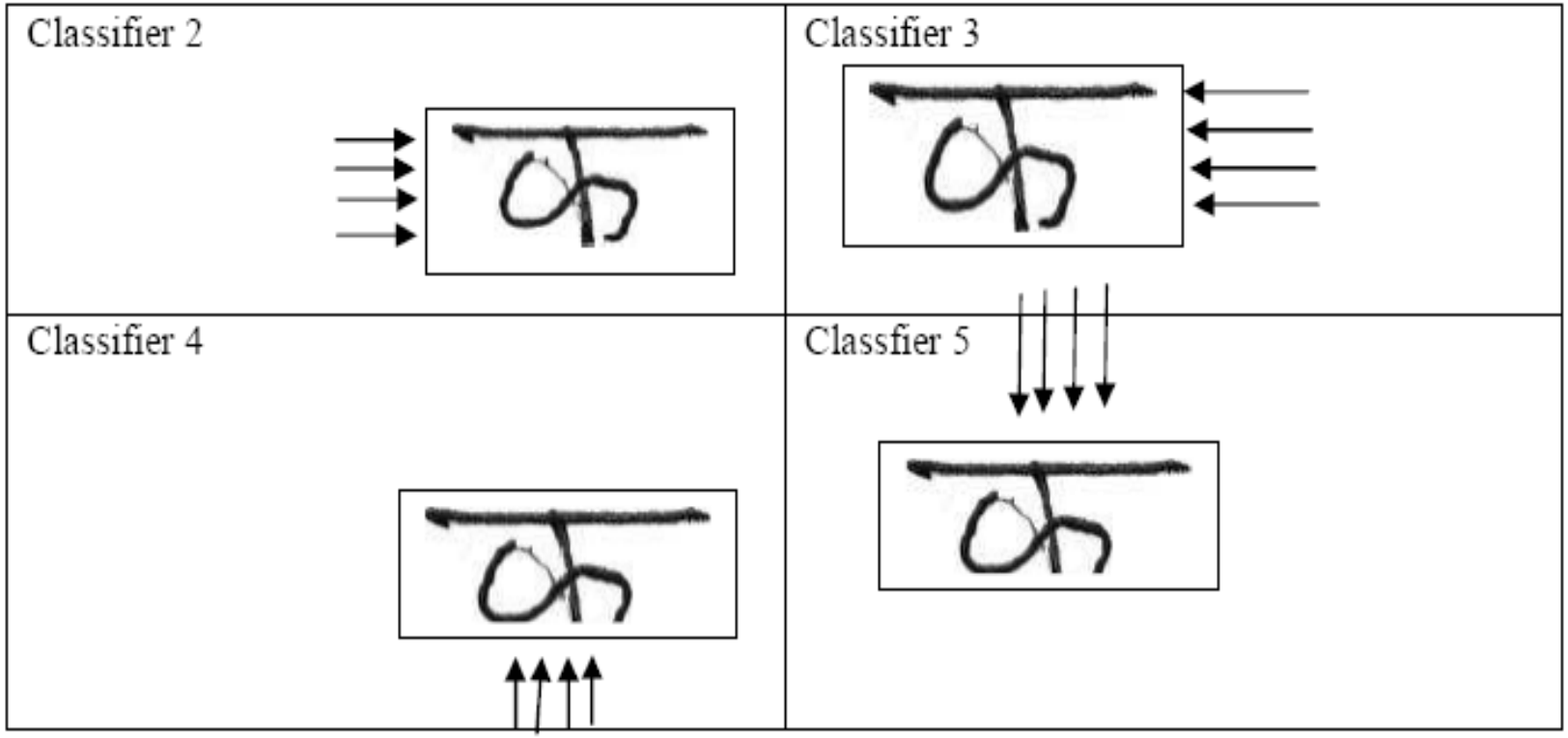
Binary character image is scanned from left to right

Structural pattern is found.

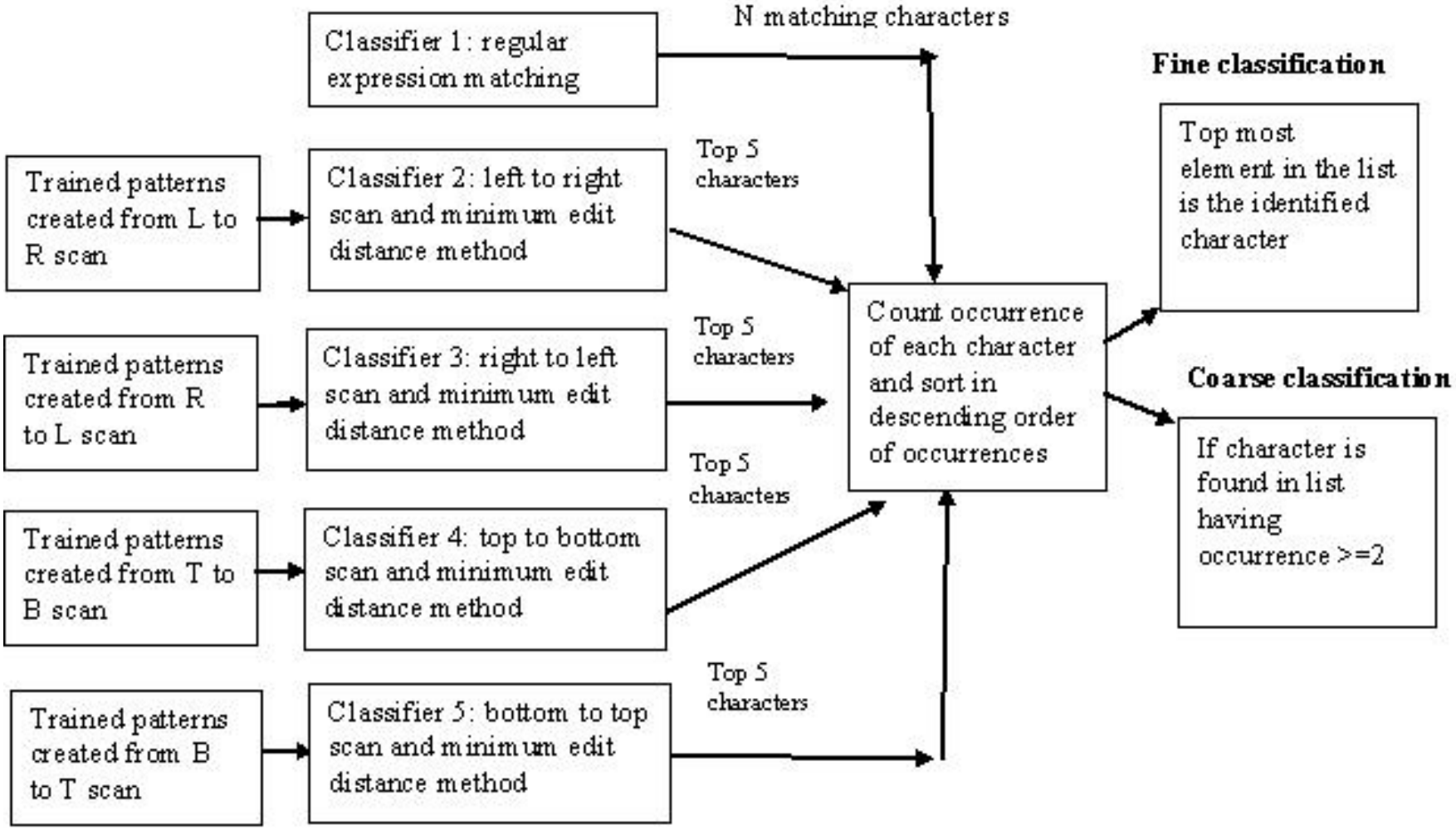
Structural pattern is compared against all patterns stored in training file

Distance is calculated using edit distance method

Classifier 2(L to R scan)



Classifier Combination



Majority voting scheme

No of matches from regular expressions 1

Character is K.jpg count 9

Character is anta-jya.jpg count 2

Character is pha.jpg count 6

Character is danta-s.jpg count 2

Results of two different classifiers if applied in isolation

Method	Result
Minimum edit distance method (L to R scan)	70%
Regular expression matching	67 %

Top choices results of combined classifier

S. No.	Proposed method result(Combined 5 classifiers)	Accuracy obtained
1	Top 1 choice	85%
2	Top 2 choices	88%
3	Top 3 choices	91%
4	Top 4 choices	92%
5	Top 5 choices	95%

Recognition at various level

S.No.	Level	Accuracy
1	Horizontal line separation	98%
2	Word isolation	94%
3	Character isolation	85%
4	Character Recognition (Coarse classification)	95%
5	Character Recognition (Fine classification)	85%

References

- Jing-Jing Li, De-Shuand Huang, Robert M. Maccallum And Xiao-Run Wu, " Characterizing Human Gene Splice Sites Using Evolved Regular Expressions, Proceedings Of International Joint Conference On Neural Networks,Montreal Canada" 2005,pp. 493-498
- K. Jaynathi, A.Suzuki, H. Kanai,Y. Kawazoe, M. Kimura, K. Kido, " Devanagari Character Recognition Using Structure Analysis",IEEE Trans, 1989,pp 363-366 .
- M. Garofalakis,R. Rastogi, K Shim," Mining Sequential Patterns With Regular Expression Constraints",IEEE Transaction On Knowledge And Data Engineering,Vol 14, No. 3, May-June 2003, pp 531-552
- K.V. Prema, N.V. Subba Reddy, "Two Tier Architecture For Unconstrained Handwritten Character Recognition",Spandan, Vol 27, Part 5, October 2002 , pp 585-594.
- Sinha R.M.K. And Mahabala H.N., " Machine Recognition Of Devanagari Script", IEEE Trans On System,Man, And Cybernetics, Vol. SMC-9, No. 8,1979, pp 435-441.
- V. Bansal And R.M.K. Sinha,"On How To Describe Shapes of Devanagari Characters and Use Them for Recognition", Proc. 5 th Conf. Document Analysis and Recognition, Banglore, India, Sept. 1999,pp. 410-413.
- P. Deshpande, L.Malik, S. Arora," Character Recognition with Histogram Band Analysis of Encoded String and Neural Network", Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, December 16-18, 2005 , pp354-359
- *Charles C. Tappert, Ching Y. Suen and Toru Wakahara,* "The State of the Art in On Line Handwriting Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No. 8, August 1990.
- B.B. Chaudhari , U. Pal, "An OCR system to read two Indian Language script Bangla and devnagai(hindi). In Proc. 4th International conference on Document analysis and recognition" ,pages 1011-1016 ,Germany 1997.
- Haunfeg Ma, David Doermann , "Adaptive Hindi OCR using generalized housdorff image comparison", UMIACS-TR-2003-87

Paper Presentation On Work

1. Deshpande P.S.,Malik L.,"Character recognition using relation between connected segments and neural network", "WSEAS Transactions on Computers", 2006, Pp 229-234 (JOURNAL).
2. Dr. P.S. Deshpande, L. Malik,"Characterizing Hand written Devnagari Characters using Regular Expressions" at IEEE TENCON 2006, Hong Kong from 14-17 November 2006.
3. Dr. P.S. Deshpande, L. Malik,"Recognition of Hand written Devnagari Characters using Percentage Component Regular Expresstion and Tree Classifier" at International Conference on Image Processing organized by University Visvesvaraya College of Engineering Bangalore from 10-13 August 2007
4. "Recognition of Hand Written Character Recognition using Connected Segments and Minimum Edit Distance" at IEEE TECON 2007 , Taipai, Oct 30-Nov 2, 2007
5. "Fine Classification & Recognition of Hand Written Devanagari Characters with Regular Expressions & Minimum Edit Distance Method" accepted in Journal of Computers, ISSN : 1796-203X Volume : 3 Issue : 5 Date : May 2008 Academy publisher(JOURNAL).
6. "Recognition of Handwritten Devnagari Script" submitted to journal

Other related paper presentations

1. S.Arora,L.Malik,” **Classification Of Gradient Change Features Using MLP For Handwritten Character Recognition** “, **International conference** in emerging applications on IT , 10-11 FEB 2006 , Organised by CSI Kolkata.
2. L.Malik, K. Hande ,” **Character Recognition using fourier descriptors** “, Emerging trends in computational science and information processing, 1-3 April 2006, J.N.E.C., Aurangabad
3. L.Malik, R. Welekar ,” **Cursive character recognition using four connected segments and minimum edit distance** “, First **International conference** on Information Technology, 19-21 March 2007, Haldia Institute of Technology, WB
4. Sandhya Arora, Dhebotosh Bhatcharjee,Mita Nasipuri, Latesh Malik ,” **A two stage classification approach for Handwritten Devnagari Characters** “, **ICCMIA 2007,** 13-15 Dec 2007, Siwakasi, TamilNadu India
5. Latesh Malik, Ekta Satija ,” **Analytical Text Segmentation and Recognition System for Devnagari Script** “, **ICEAETS 2008,** Rajkot, Gujrat, 13-14 Jan 2008
6. Snehal Dalal, L.Malik ,” **A Survey of Methods and Strategies for Feature extraction in handwritten Script Identification** “, ICETET 08, GHRCE Nagpur 16-18 July 2008

RECEIVED GRANT

- Received grant form AICTE under RPS of Rs. 7.5 Lakhs for project “ Recogntion of hand written manuscript in Devnagari Script”
- Duration of project is 2 years.

Thanking You