International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Recognition of a Medieval Indic-'Modi' Script using Empirically Determined Heuristics in Hybrid Feature Space

R. K. Maurya[1*], S. R. Maurya[2]

[1*] University Department of Computer Science (UDCS), University of Mumbai-98, Maharashtra, India
[2] S. K. Somaiya College of Arts, Science & Commerce, Vidyavihar, Mumbai-77,Maharashtra, India
e-mail: [*]rajeshmaurya@udcs.mu.ac.in

Available online at: www.ijcseonline.org

Received: 21/Jan/2018, Revised: 29/Jan/2018, Accepted: 12/Feb/2018, Published: 28/Feb/2018

***Abstract*—** The 'Modi' script originated as a cursive variant of the script during the 17th century CE and used to write the Marathi language spoken in the Indian state of Maharashtra. Modi script evolved over time and found in many styles of writing. There is no standardization for writing characters and numerals of 'Modi' script but largely written without lifting the pen. The cursive nature of the script and lack of standardization in writing style pose challenges in digital recognition of documents written using Modi script including historical ones. Being largely medieval era script, modern document recognition systems lack support for recognizing handwritten texts using 'Modi' script. In this paper, we have described the framework of digital recognition of characters of handwritten 'Modi' script using empirically determined heuristics for determining the contribution of features from hybrid feature space for recognition of the 'Modi' character. The hybrid feature space uses normalized chain code together with feature vector encompassing a number of holes, endpoints, and zones associated with the character. The proposed framework for digital recognition of 'Modi' character using empirically determined heuristics provides a naïve model for recognizing a class of Indic scripts especially based on the cursive style of writing. The average and best recognition performance for the proposed method was measured to be 91.20% and 99.10% respectively.

***Keywords*—** Chain Code, OCR, Medieval Script, 'Modi' Script Recognition, Empirical Heuristic-based OCR

I. INTRODUCTION

The ancient and medieval era of the Indian history had well-evolved scripts and created good literary documents. These literary documents mainly handwritten should be preserved for future. 'Modi' is one such medieval script that was used to write the Marathi language, which is the primary language spoken in the state of Maharashtra in western India.

'Modi' Script was greatly used by political and administrative authorities as well as businessmen to keep their accounts and other important documents. 'Modi' originated as "cursive" style of writing Marathi near about last 600 years and was mostly used since 17th century till the 1950s when Devanagari replaced it as the written medium of the Marathi language. The cursive style of the 'Modi' script often makes it difficult to read and thus could also be used as shorthand script for faster writing.

There are two major classes of scripts in India-scripts with and without Sirorekha (a head bar). 'Modi' script belongs to the former class. Importantly 'Modi' is generally written without lifting the pen which produces cursive property in the character of the script. It neither accounts for vowel length nor does include conjunct consonants like Devanagari and some other Indic scripts. It was also used to encrypt the message since not all people were well versed in reading this script.

Presently Marathi uses Devanagari script while most of the documents found in archives are written in 'Modi'. Since even as of today there are only a handful number of expert people who can transcribe documents written in 'Modi', restoring these documents have become difficult. This gives the motivation to develop a system that would recognize documents written using 'Modi' and preserve them for future. The documents written using 'Modi' script shows a high degree of variations in writing style including discrete and legitimate to cursive, and fused character styles. There are little-known efforts done for digital recognition of hand documents written using ''Modi' script. The digital recognition of the script can help to preserve the knowledge repositories found in the historical documents.

This paper claims results of digital recognition of 'Modi' script characters for non-fused handwritten discrete

characters of 'Modi' script. The input documents are treated as two-dimensional intensity map. After fundamental preprocessing of these image documents, the character recognition follows matching features of the character in feature space that contain spatial and shape characteristics of the character. In this paper, we have developed and experimented technique to recognize characters of 'Modi' script in hybrid feature space and present their recognition performance. The paper is based on empirically determined heuristics for matching in feature space that comprises of a spatial feature of the character including number of holes, number of endpoints, zone description and normalized chain code analysis on archived image documents of 'Modi' script.

II. RELATED WORK

Many types of research in past decades have tried to address the issues of recognizing Indian scripts. Indian scripts have complex structural compositions which makes designing OCR relatively difficult. Initial works by R. M. K. Sinha and others in [1], [2] and [3] have addressed the issues with recognizing modern Devanagari script successfully mainly using knowledge sources and rule-based contextual post-processing. Their work is limited to the typeset image documents and more modern way of writing. Character recognition for Devanagari script has been studied by V. Bansal in [4] that used a two-pass algorithm for the segmentation and decomposition specifically for composite characters/symbols into their constituent symbols.

There are good attempts for recognition of other Indic scripts, but largely work on medieval Indic script such as 'Modi' Script recognition is limited. Remarkably, there is a large collection of valuable medieval documents written using 'Modi' script which itself has evolved over the time and thus shows great variations. These documents have great contribution in literature in central and western parts of India. 'Modi' has been less explored for the purpose of automation of OCR. D. N. Besekar et al. have done some preliminary work in recognition of 'Modi' script in their papers [5], [6]. The work in the paper [5] describes the structural similarities of standard characters and handwritten characters in 'Modi' script while [6] outlines the theoretical analysis of 'Modi' script and other issues in recognizing the script. More specific attempt form them include mathematical morphological approach [7] and chain code based recognition [8] on standard characters of 'Modi' script. The accuracy may vary when simply using chain code based recognition approach as 'Modi' follows free-form writing style and has many similar looking characters.

This paper attempts to create a framework for generating OCR for simple medieval documents containing 'Modi' scripts using feature space analysis together with chain code tests and highlight performance on individual 'Modi' script characters.

III. CHALLENGES OF OCR FOR 'MODI' SCRIPT

'Modi' script has cursive characteristics due to its writing mechanism. Generally 'Modi' is written without lifting the pen, thus forming complex pattern in many cases. The hand written documents showed large variations in character construction and often lead to more complex problem of touching and fused characters.

- Few of the critical challenges in recognizing 'Modi' scripts include:
- Availability of only handwritten documents and absence of typeset template for target match.
- Presence of high degree of self-occluding structures in the 'Modi' characters.
- Cursive writing style for all characters.
- Variation in character pattern due to change in speed of writing
- Occurrence of fused characters in the sentences of the input documents.

Developing an OCR for 'Modi' script invites solving mentioned problems. The characters of 'Modi' script have evolved over a period of time and there are many variations to write the same character. Since there are no standard typed documents and standard character set/style available, recognition is a challenging task. Due to its continuous writing style, 'Modi' often creates occluding structures and fused or overlapping words and sentences which makes recognition more difficult. The speed of writing also brings large variation in the document created in 'Modi'. Overall all these problems make the development of standard OCR for 'Modi' more difficult.

In this paper, we have proposed the methodology to recognize character set of 'Modi' script independent of writing style by focusing on most prominent features of the character including shape. With exception of few highly similar characters, the recognizer performs quite well. Relevant details should be given including experimental design and the technique (s) used along with appropriate statistical methods used clearly along with the year of experimentation (field and laboratory).

IV. METHODOLOGY

The inputs to the 'Modi' OCR described in this paper are handwritten documents that contain 'Modi' script characters. The input documents are subject to a number of preliminary processing steps, in order to make it suitable for stages of

character analysis and recognition. Pre-processing aims to generate improved forms of inputs that are easy for the OCR systems to process and produce better results. The preprocessing methods include binarization, noise removal, skeletonization, smoothing, and segmentation of characters. Once preprocessing is complete, the segmented characters are ready for post-processing and recognition.

The work presented in this paper focuses on a more generic approach by using empirically determined heuristic for determining the contribution of individual features of the character rather than the template-based approach [8] and their variants. A hybrid feature space is constructed using normalized chain code [9],[10],[11] shape descriptions together with the traditional feature vector. The benefit of using empirically determined heuristics on this hybrid feature space can be easily realized as it allows more flexible input including characters of varies sizes.

The framework described in this paper accepts images of handwritten 'Modi' script documents saved in most widely used formats including jpeg, tiff, BMP or PCX. The input image is binarized to detect foreground character pixels from the background. The framework is independent of input image formats as it works on binarised form rather than grey scale images. The binarized image is processed to have foreground as character pixels in white while the background in black. The binarized image is treated for removal of noise and all connected components that have fewer than p pixels. For practical reasons, p=10 is used except conjoint characters. Conjoint characters with a dot (.) are dealt separately. The image is now sent to segmentation module which segments line, word, and character using histogram analysis [13]. The segmented characters are then normalized and resized to 40x30 pixels for standard database storage and matching.

The next step in preparing segmented character is to skeletonize each segmented image. Then each image segment consisting of characters are sent for identification and creation of feature vectors. The feature vector space in the proposed model is defined as

$$\vec{F}_S \equiv \vec{V}_{F(e,h,z)} \cup \left|\vec{V}_{N\_Chain}\right| \quad \dots(1)$$

with

$$\vec{V}_{F(e,h,z)} = F_e i + F_h j + F_z k \quad \dots(2)$$

where $\vec{F}_S$ is the hybrid feature space vector with vector $\vec{V}_F$ describing number of end points (F_e), holes (F_h) and zone (F_z) descriptions for the character under evaluation. F_z described the presence of the character divided in nine regions. $\left|\vec{V}_{N\_Chain}\right|$ is the normalized chain code vector for the skeletonized character. $\left|\vec{V}_{N\_Chain}\right|$ overcomes small variations in the shape due to writing style. Next, the read character module computes the coherence of the input characters defined as feature space vector $\vec{F}_S$ on a set of training hybrid feature space vector database for characters constructed using empirically determined heuristics.

The heuristics for each feature in the feature vector is determined experimentally by applying the technique on more than 120 test cases from each character set and testing their correct recognition and classification on the known character set. The database was populated with nearly 3200 samples. The database is self-organized into different character sets based on their proximities of similarity in feature descriptions and chain code shapes. The cumulative empirical probabilistic weights for each match for the segmented input character I can be found using the following term

$$W_{EP}(I) = \sum_{j=1}^{m_i} \sum_{k=1}^{4} w_{i,k} * \vec{F}_{S_{i,k}} \quad \dots(3)$$

$w_{i,k}$ denotes the empirically determined heuristic weights corresponding to each feature element in the hybrid feature vector space . The index i denotes the unique character class and j denotes the index of samples within each class i with

$$1 \le i \le n,$$

$$1 \le j \le m_i$$

$$\text{and} \quad 1 \le k \le 4$$

m_i is the maximum number of records in a particular character class from the test database. The value of k ranges from 0 to 4 corresponding to feature vector elements number of holes, endpoints, zones and normalized chain code. $\left|\vec{V}_{N_{Chain}}\right|$. This new input sample is tested againt the database and maximum match to a character is determined as defined below

$$C = max\left(corr(W(C_i), W_{EP}(I))\right) \quad \dots(4)$$

$W(C_i)$ is cumulative character weight stored in the knowledge database for the ith known character. the C is the matched output character. W_{EP} has potential to highlight potential features in the classification and recognition process againsts known characters in the database. It is possible that due to non-standardization in writing style there may be the large variation in the character. W_{EP} is used to correlate in its hybrid feature space with the database character. For some special cases where W_{EP} is critically less than a defined threshold θ (i.e. $W_{EP} << \theta$), the input suggests more than one probable characters.

The experiment validates the improvement of recognition when normalized chain code analysis is used together with

the feature vector describing the shape for the matching purpose. The highest contribution is taken for chain codes of the characters in the recognition process. The zonal description has next highest significance followed by the number of endpoints and the number of holes. The number of holes and end points are crucial, experiments have shown that together with zonal and shape description using chain code, the result is improved for classification and recognition of the characters.

V. ALGORITHM FOR EMPIRICALLY DETERMINED HEURISTICS WEIGHT METHOD

The proposed model of character recognition for 'Modi' script follows a simple heuristic approach. Each element of the feature vector is associated with an experimentally determined heuristics that determine the contribution of a specific feature in the overall character recognition process. These heuristics are used as a decision parameter to determine the percentage contributions of respective features from the feature vector. The feature vector for the script characters is stored in the format described in Figure 1.

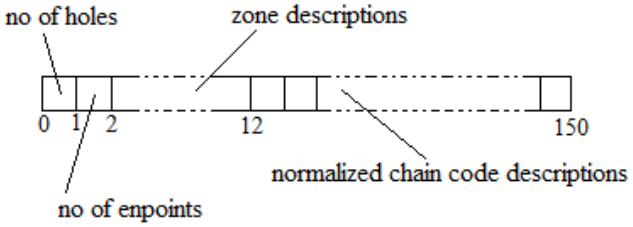


Figure 1. Hybrid Feature Vector Format for 'Modi' Characters

The naïve pseudo code for recognition of the segmented characters based on heuristics feature vectors matching described below.

Algorithm for creating hybrid feature vector space and matching 'Modi' script characters

```
/* feature_vector is a vector containing description of
script characters containing number of holes, number of
  end points, region vector and normalized chain code
                      vector.*/
1.  Input image document with handwritten 'Modi' script
    characters
2.  Load database directory
3.  Preprocess and segment script characters
4.  Repeat step 4 to 6 for each segmented n character
5.  [Create hybrid feature vector]
  5.1 Compute feature vector
            a.Compute number of holes
            b.Compute number of endpoints
            c.Compute zonal descriptions
            d.Compute feature vector
```

$$\vec{V}_{F(e,h,z)} = F_e i + F_h j + F_z k$$

```
  5.2 Create normalized chain code
  5.3 Create hybrid Feature vector
```

$$\vec{F}_S \equiv \vec{V}_{F(e,h,z)} \cup |\vec{V}_{N\_Chain}|$$

```
6.  Repeat for each record m in the hybrid feature
    database
  6.1 if  Fs(n,1) = Db(m,1)
  /*no_of_holes(feature_vactor)==no_of_holes(db_char)*/
      feature1_per=heuristics1;
  6.2 if  Fs(n,2) = Db(m,2)
      /*no_of_endpoints(feature_vector)==
      no_of_endpoints(db_char) */
      feature2_per=heuristics2;
  6.3 [Zonal matching in nine regions]
      initialize zone_count=0;
      for i=3:11
            if feature_vector(n,i)==db_char(m,i)
          zone_count=zone_count+1;
       /* counting character presence in nine zones*/


      feature3_per=round(zone_count*heuristics3/9);
7.  [Normalized chain code Matching]
  Initialize ccode_count=0;
  6.4 for i =12:n
           if feature_vector(n,i)== db_char(m,i)
       ccode_count=ccode_count+1;
           /* match shapes using chain codes in the
                Vector*/
  feature4_per=round(ccode_count*heuristics4/150);
8.  [Compute match percentage]
    match_per=feature1_per+feature2_per+feature3_per
                +feature4_per;
  /* record the value of matching factor with database
  Character*/
9.  /*update & match percentage and corresponding
    character using maximum value computed and note
    corresponding index*/


10. /*Find index and value of highest matched character
    and return character*/
```

The general framework for character recognition for Modi script using empirically determined heuristics model is shown in Figure 2.

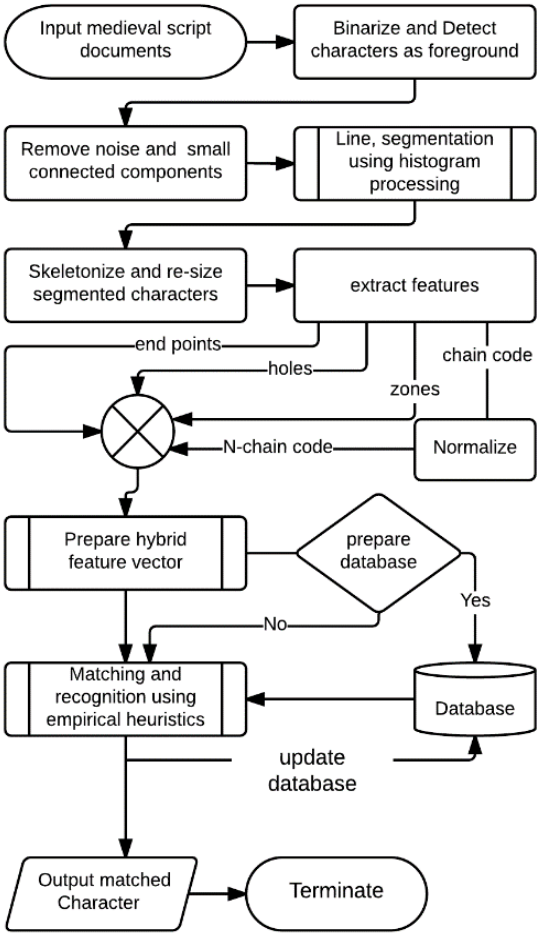


Figure 2. Framework for 'Modi' OCR Using Empirical Probabilistic Heuristics on Hybrid Feature Space

The proposed model encompasses a high degree of flexibility in updating the knowledge database by allowing new character samples to be included in the test sets.

VI. EXPERIMENTAL DATA

The experimental setup for implementing the model required scanned documents containing 'Modi' script characters. The experiment is performed on nearly 3200 test sample characters from the scanned documents and results is noted. For each character, the sample test space created from nearly 120 test characters for each character. The is then further normalized chain codes of the character helps to narrow down the search space which further combined with empirical weights of feature vectors to determine the cumulative probabilistic category and mapped to the most probable matching symbolic character of the 'Modi' script. Few samples of input to the empirical spatial correlation model are shown in Figure 3.

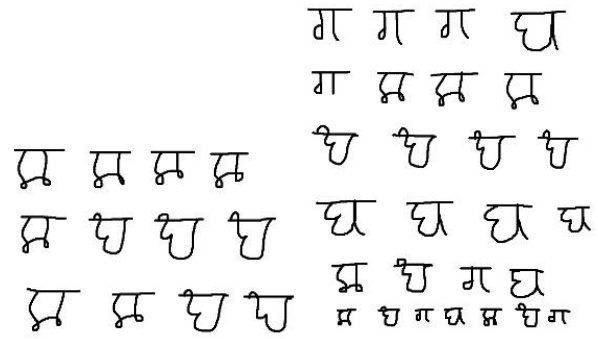
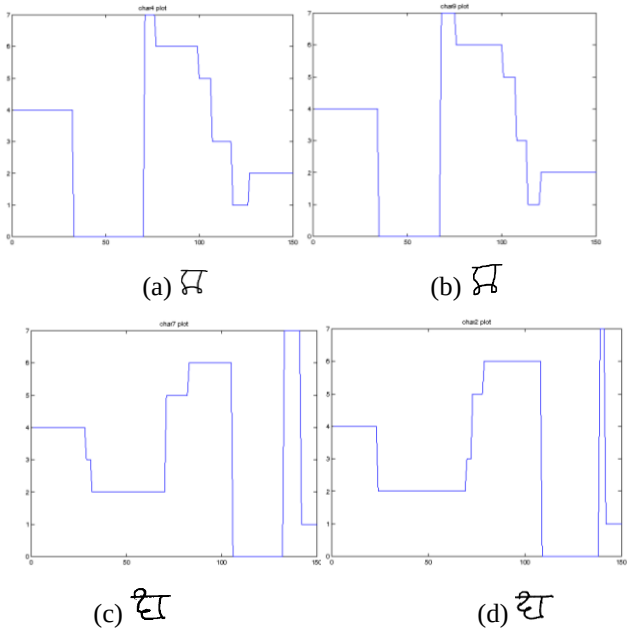


Figure 3. Input samples to 'Modi' OCR

The segmented input characters from the image document is passed to recognition module based on the empirical heuristic matching of feature vector containing chain code of the character.

VII. RESULTS AND DISCUSSIONS

The framework of 'Modi' OCR was applied on thirty-three characters of 'Modi' script on 60 handwritten documents with nearly 3200 characters. The recognition performance is highly influenced by chain code description of the character, thus giving higher heuristic value to chain code contribution. The contribution of shape analysis using empirically determined heuristic based chain codes can be verified from the normalized chain code graphs. Some sample normalized chain code graphs are given in Figure 4.



(a) (b)

(c) (d)

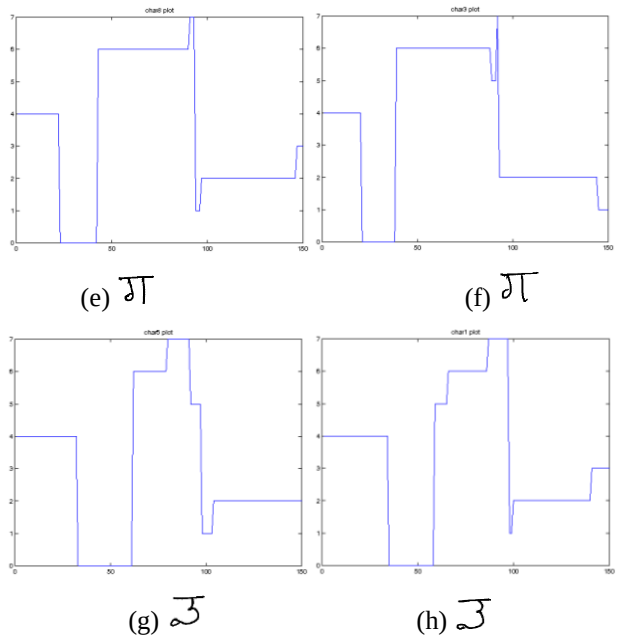


Figure 4. Normalized chain code graphs from the feature vector

Table1: Performance of OCR on numbers and characters of 'Modi' script

Input	Recognition Rate for 'Modi' Numerals	
	Average Performance	Best Performance
Numbers	98.16 %	99.14%

The results obtained show promising performance on discretely written 'Modi' script documents. The recognition rates individual numeric handwritten sample is presented in Table 2. The Table 2 presents the average recognition sampled over 33 documents containing nearly 130 variants of individual symbols of the 'Modi' scripts.

Table 2. Recognition rate on 'Modi' characters

Sample characters	Performance Statistics		
	Correct match (in %)	**Sample characters**	**Correct match (in %)**
	91.03		87.24
	88.33		95.00
	99.01		99.10
	87.80		88.96
	87.07		91.44
	84.01		84.89
	92.66		98.26
	88.46		89.17
	79.47		96.64
	83.33		96.47
	97.47		81.99
	75.56		95.55
	83.22		80.09

VIII. CONCLUSION AND FUTURE SCOPE

We have described a simple and effective approach for recognizing medieval characters of 'Modi' script using experimentally determined the heuristic approach for feature matching and highlighted significant performance improvement over recognition rate using simple chain code analysis. The handwritten characters were extracted from the input image, normalized and feature vectors were compared with the database. Average recognition rate was with proposed method was 91.20% even for highly cursive style of writing. Experimental results suggest that the proposed empirical model is suitable for legibly written discrete 'Modi' script character. The study also generates scope for a model that is independent of the size of handwritten characters including those with loops and occluding structures. The results also suggest that the framework is independent of the style of writing and it can be easily applied to cursive and non-cursive variants of 'Modi' script. In the future work, we will try to improve the performance and reduce the search space of the OCR further to speed up the process and obtain more accuracy by employing analysis of feature space vector of the characters and applying other soft computing techniques. In addition, the empirical heuristics can be improved by applying context-based knowledge.

REFERENCES

[1] R.M.K.Sinha, "Rule based contexual post-processing for Devanagari text recognition", Pattern Recognition, 20(5), pp. 475-485, 1987.

[2] Veena Bansal, R. M. K. Sinha, "Segmentation of touching and fused Devnagari characters", Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'98),371 - 376, December 21-23, New Delhi, 1998

[3] R. M. K. Sinha, H. N. Mahabala, "Machine Recognition of Devanagari Script", IEEE International Conference on Systems, Man and Cybernetics, pp. 435-441, 1979.
[4] Bansal, V. ; Sinha, R.M.K. A Complete OCR for Printed Hindi Text in Devanagari Script, Sixth International Conference on Document Analysis and Recognition, 2001. Proceedings.
[5] A S Ramteke, G S Katkar, Recognition of Off-line Modi Script : A Structure Similarity Approach, "International Journal of ICT and Management" ISSN No. 2026-6839, February 2013
[6] D. N. Besekar, A S Ramteke, "Theoretical analysis of MODI script according to recognition point of view, some issues involved with character recognition of MODI script", International Journal of Computer Applications, February 2013
[7] D. N. Besekar, "Recognition Of Numerals Of Modi Script Using Morphological Approach",Shodh Samiksha Aur Mulyankan Vol.III, Issue-27, april 2011
[8] Anil K. Jain, Template-based online character recognition, Pattern Recognition, Volume 34, Issue 1, January 2001, Pages 1–14
[9] D. N. Besekar, A S Ramteke, "Chain Code Approach For Recognizing Modi Neumerals , Indian Journal Of Applied Research", December 2011
[10] Guojun Lu. "Chain Code-Based Shape Representation and Similarity measures", Visual Information Systems, 1997,Springer Berlin Heidelberg, 1997
[11] Nor Amizam Jusoh and Jasni Mohamad Zain, Application of Freeman Chain Codes: An Alternative Recognition Technique for Malaysian Car Plates", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.11, November 2009
[12] Sadanand A. Kulkarni et al,"Offline Handwritten MODI Character Recognition Using HU, Zernike Moments and Zoning", arXiv:1406.6140 , 2014
[13] Dos Santos, Text Line Segmentation Based on Morphology and Histogram Projection, 10th International Conference on Document Analysis and Recognition, 2009. ICDAR '09.
[14] V.Di Lecce, G.Dimauro, A.Guerriero, S.Impedovo, G.Pirlo, A.Salzo, Zoning Design For Hand-Written Numeral Recognition, Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, 2000
[15] Sang Hak Lee, Hyung Il Koo, Nam Ik Cho, "Image segmentation algorithms based on the machine learning of features", Published in Pattern Recognition Letters, Volume 31 Issue 14, pp. 2325-2336,October, 2010.
[16] Hung-Ming Sun, "Multi-Linguistic Optical Font Recognition Using Stroke Templates", Proceedings of the 18th International Conference on Pattern Recognition - Volume 02 , Volume 02, August 2006.
[17] John Cowell, Fiaz Hussain, "Two template matching approaches to Arabic, Amharic and Latin isolated characters recognition", Machine Graphics & Vision International Journal , Volume 14 Issue 2, pp. 213-232, Jan 2005
[18] Zabih, R. Kolmogorov V, "Spatially coherent clustering using graph cuts", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 2, pp. 437-444, June 2004.
[19] Mohammed Ali Qatran G, "Template matching method for recognition musnad characters based on correlation analysis", ACIT' Proceedings, 2011
[20] Vassilis Papavassiliou, ThemosStafylakis, VassilisKatsouros, GeorgeCarayannis, "Handwritten document image segmentation into text lines and words", Pattern Recognition, Volume 43 Issue 1, pp. 369-377 January 2010.
[21] Tanmoy Som, Sumit Saha, "A new approach for slant angle correction and character segmentation of handwritten document", social science research network,2008.
[22] Saha, Sumit, "Handwritten Character Recognition Using Fuzzy Membership Function", 2008 International Journal of Emerging. Technologies in Sciences and Engineering,Vol.5, No.2, pp. 11-15, Dec 2011.
[23] Swapnil Khedekar , Vemulapati Ramanaprasad , Srirangaraj Setlur , Venugopal Govindaraju, "Text–Image Separation in Devanagari Documents", Proceedings of the Seventh International Conference on Document Analysis and Recognition, p.1265, August 03-06, 2003
[24] Rafael Palacios, Amar Gupta, "Training neural network for reading handwritten amount of checks", Working Paper 4365-02, May 2002.
[25] S. Basavaraj Patil, N V Subbareddy, "Neural network based system for script identification in Indian documents", Sadhana Vol. 27, Part 1, February 2002, pp. 83–97.
[26] V. Bansal, R.M.K. Sinha, "Partitioning and Searching Dictionary for Correction of Optically-Read Devanagari Character Strings", Proceedings of the international conference on Document Analysis and Recognition (ICDAR-99), Bangalore, India, pp. 653- 656, 1999
[27] M. Hanmandlu, Harish Kumar, K. R. Murli Mohan, "Neural Based Handwritten Character Recognition", Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR), pp 241 , 1999.
[28] Thein M. Ha, Horts Bunke, Off-Line, "Handwritten Numeral Recognition by Perturbation Method", IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 19 , Issue 5 (May 1997) pp 535 - 539
[29] B.B. Chaudhuri, U. Pal, "An OCR system to read two Indian language scripts: Bangla and Devanagari", Proceedings of 4th International Conference on Document Analysis and Recognition, pp. 1011-1015, 1997.

Authors Profile

Mr. R. K. Maurya is an academician and researcher holding Master Degrees including M.Sc (Comp Sc), M.Phil (IT), MCA and MBA (IT). He is currently associated with University Department of Computer Science, University of Mumbai, India. He is a member of IEEE, ACM, a life member of the CSI, ISCA, IETE and several other scientific societies related to computing. He has published more than 12 books in computing-related areas with publishers of international repute and also has research papers in reputed l journals. His research work focuses Image Processing, Spatial Computing and Analysis, Data Analytics and Computational Intelligence. He has more than 15 years of experience in teaching, Industry and research.

Mrs. S. R. Maurya is currently heading the Department of Computer Science at S. K. Somaiya College, Mumbai. She has more than a decade of experience in teaching and research. She has a keen interest in data analytics, spatial data management and visualization and machine intelligence. She is a member of reputed institutions including CSI, IEEE, ISCA, ACM IETE etc. She has several books and papers in international journals and conferences to her credit as published contributions.

