A Survey on Recognition of Devnagari Script

Ratnashil N Khobragade1
Assistant Professor
PG Dept of Computer Science
Sant Gadge Baba Amravati

Dr. Nitin A. Koli
Head, Computer Center
Sant Gadge Baba Amravati
University, Amravati

Mahendra S Makesar
Assistant Professor
Prof Ram Meghe College of
Engg and Mgt, Amravati

Abstract:

This paper describes a set of preprocessing, segmentation, feature extraction, classification and matching techniques, which play very important role in the recognition of characters. Feature extraction provides us methods with the help of which we can identify characters uniquely and with high degree of accuracy. So many approaches have been proposed for pre-processing, feature extraction, learning/classification, and post-processing. The objective of this paper is to review these techniques, so that the set of these techniques can be appreciated.

Keywords: OCR, preprocessing, segmentation, feature extraction and classification.

1. Introduction:

1.1 Indian Language Characteristics

India is a multi lingual multi script country with twenty two scheduled languages, namely, Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri (Meithei), Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu**.** These languages are written using only twelve scripts.

Devnagiri script used to write Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Mathili. Sindhi is written using Devnagiri script in India and Urdu script in Pakistan. Assamese, Manipuri and Bangla languages are written using Bengali script. Gurmukhi script is used to write Punjabi language. All other languages have their own script. In Indian language scripts, the concept of upper case and lower-case characters is not present. Most of the Indian languages are derived from Ancient Brahmi and are phonetic in nature and hence writing maps sounds of alphabets to specific shapes. All these languages, except Urdu, are written from left to right.

1.2 Characteristics of Devanagari Script

Devanagari is used in many Indian languages like Hindi, Nepali, Marathi, Konkani, Sindhi etc. More than 300 million people around the world use Devanagari script. This script forms the foundation of Indian languages. So Devanagari script plays a very major role in the development of literature and manuscripts. Devanagari script has about 11 vowels and 33 consonants. Devanagari script is written from left to right and it does not have any upper or lower case letters. It is usually recognized by a horizontal line that connects the top of the characters in a word. However, in some words, all the characters are not connected. The alphabets consisting of consonants, vowels, conjuncts in the Devanagari script are now enumerated [1]. In handwritten recognition difficulty is mainly caused by the large variations of individual writing style. So many approaches have been proposed for pre-processing, feature extraction, learning/classification, and post-processing. The objective of this paper is to review these techniques, so that the set of these techniques can be appreciated and use for recognition of Marathi Manuscript.

2. Review of Literature:

The literature survey carried out related to technology impact in the study of different text recognition techniques use on different languages of printed and handwritten scripts. Research in Indian offline character recognition started with the recognition of printed characters, irrespective of the script and then extended to the recognition of handwritten numbers and characters in many Indian scripts including Devanagari. Rotation invariant texture features and their use in automatic script identification results for the Chinese documents revealed that a significant part of the errors was due to documents misclassified as Persian. The reason for this is not entirely clear [2]. Some research is also devoted towards segmentation of touching characters, recognition of handwritten compound characters and words in various Indian scripts. OCR work on printed Devanagari script started in early 1970s. An extensive research on printed Devanagari text was carried out by Veena Bansal [3]. First system for hand-written numeral recognition of Devanagari characters was proposed by [4]. U. Pal et. al. [5] presented a system for off-line handwritten character recognition of Devanagari using directional information for extracting features. Debashis Ghosh et. al. [6] reported in survey an overview of the different script identification methodologies under each of these categories. B.V.Dhandra et. al. [7] presented an automatic technique for script identification at word level based on morphological reconstruction is proposed for two printed bilingual documents of Kannada and Devnagari containing English numerals. The technique developed includes a feature extractor and a classifier. Water reservoir analogy is used, to extract individual text lines from printed Indian documents containing multioriented and/or curve text lines [8]. Unexpected noise in a sequence might “break” the normal transmission of states for this sequence, making it unrecognizable to trained models. The strategy use will compensate for some of the negative effects of this noise. System achieves a 98.88 percent accuracy rate on handwritten digits. Different kinds of degradation [9] identified from printed Gurmukhi script. Such as touching characters, broken characters, heavy printed characters, faxed documents and typewritten documents. Problems associated with each kind of degradation [10]. A simple and an efficient off-line handwritten character recognition system using a new type of feature extraction, namely, radon feature extraction is proposed by M. K. Mohahmed Althaf, M. Baritha Begum with recognition accuracy of 90% for 270 features[11].

Handwritten character recognition aims at converting handwritten characters in images into text that can be stored, edited or can be converted into text speech. This field of research finds applications in various areas that aim in automation so as to reduce the human efforts like postal automation, bank automation [12], form filling etc. Handwritten character recognition for Indian scripts [13] is

quite a challenging task due to several reasons. These scripts have a character set with a large number of characters in it. The shape of the characters is complex and they may have modifiers, present above, below or in line with the character. The modifiers are the vowels that change their shape when connected to the consonants. Moreover, some character pairs are almost similar to each other that make them quite difficult to classify. Another reason is presence of compound characters in some scripts like Bangla, Devanagari etc. where two or more consonants are joined together to form a special character. Further, handwritten characters tend to show a large variation in the basic shape of the characters since the pen ink, pen width, accuracy of acquisition device, the stroke size and location in the character, different writing styles of different person, physical and mental situation of the writer affects the writing style and in turn the recognition accuracy to a considerable extent.

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 [14]. A structured analytical approach to handwritten Marathi vowels recognition [15], Multistage handwritten Marathi compound character recognition using neural networks presented by S. Shelke and Shaila Apte [16], Vikas Dongre and Vijay H Mankar present the Devnagari document segmentation using histogram approach [17]. From the survey by R. Jayadevan et. al. [18], it is noted that the errors in recognizing printed Devanagari characters are mainly due to incorrect character segmentation of touching or broken characters.

3. Proposed Model:

We intend to take the following steps in the proposed model for Marathi manuscript recognition and its text interpretation as shown in figure 1.

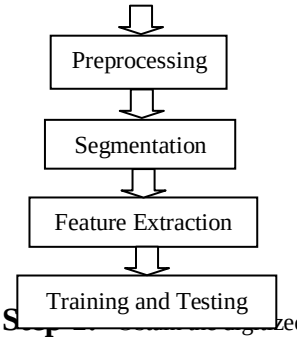

Figure 1. Sequential Steps in Proposed System

S zed image of Marathi character.

Step 2:- Preprocessing:

Pre-processing phase is applied to remove unwanted parts from the image by applying one or more technique such as Binarization, Complement, Size normalization, Morphological Operation, Noise removal using filters, thinning, cleaning techniques and filtering mechanisms, thresholding, skeletonization tecnniques can be used [19,20].

Step 3:- Segmentation:

It is one the most important process that decides the success of character recognition technique. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. Marathi words can further be splitted in to individual character for classification and recognition by removing Shirorekha [19].

Step 4:- This step can be subdivided into following four sub divisions-

1. Feature Extraction
2. Training
3. Classification
4. Matching Techniques

3.1 Feature Extraction:

Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class.

According to C. Y. Suen [18], Features of a character can be classified into two classes: Global or statistical features and Geometrical or topological features.

3.1.1 Global or Statistical Features

Global features are obtained from the arrangement of points constituting the character matrix. These features can be easily detected as compared to topological features. Global features are not affected too much by noise or distortions as compared to topological features. A number of techniques are used for feature extraction; some of these are: moments, zoning, projection histograms, n-tuples, crossings and distances.

i) Moments : In this case the moments of different points present in a character are utilized as a feature. Heutte et. al [21] said that these are most commonly used methods in character recognition . M K Hu [22] brought the concept of classical moment invariants. Hu's other moments are statistical measure of the pixel distribution about the centre of gravity of the character. S. S. Reddi [23] proposed Radial and angular moments where as Zernike moments were proposed by Teh and Chin [24].

ii) Zoning: According to this technique the character matrix is divided into small portions or zones. The densities of pixels in each zone are calculated and used as features. This concept was suggested by Hussain et al. [25].

iii) Projection histograms : Projection histograms give us the number of black pixels in the vertical and horizontal directions of the specified area of the character. This concept was introduced by M. H. Glauberman [26] in a hardware OCR system. Projection histograms may be vertical, horizontal, left diagonal or right diagonal.

iv) n-tuples: According to this method the position of black or white pixels in a character image is considered as a feature. The n-tuple method has been developed by Tarling and Rohwer [27]. This method provides a number of important properties of pixels.

v) Crossings and distances: Some researchers have obtained features by analyzing the counts the character image is crossed by vectors in certain directions or at certain angles.

3.1.2 Geometrical and Topological Features:

Various global and local properties of characters can be represented by geometrical and topological features

with high tolerance to distortions and style variations. This type of representation may also, encode some knowledge about the structure of the object or may provide some knowledge as to what sort of components make up that object. Various topological and geometrical representations can be grouped in four categories:

i) Extracting and Counting Topological Structures: In this category, lines, curves, splines, extreme points, maxima and minima, cups above and below a threshold, openings, to the right, left, up and down, cross (X) points, branch (T) points, line ends (J), loops (O), direction of a stroke from a special point, inflection between two points, isolated dots, a bend between two points, horizontal curves at top or bottom, straight strokes between two points, ascending, descending and middle strokes and relations among the stroke that make up a character are considered as features [28,29].

ii) Measuring and Approximating the Geometrical Properties: In this category, the characters are represented by the measurement of the geometrical quantities such as, the ratio between width and height of the bounding box of a character, the relative distance between the last point and the last y-min, the relative horizontal and vertical distances between first and last points, distance between two points, comparative lengths between two strokes, width of a stroke, upper and lower masses of words, word length curvature or change in the curvature[30-33].

iii) Coding: One of the most popular coding schemes is Freeman's chain code. This coding is essentially obtained by mapping the strokes of a character into a 2-dimensional parameter space, which is made up of codes. There are many versions of chain coding. The character frame is divided to left-right sliding window and each region is coded by the chain code [34-36].

iv) Graphs and Trees: Words or characters are first partitioned into a set of topological primitives, such as strokes, holes, cross points etc. Then, these primitives are represented using attributed or relational graphs. Image is represented either by graphs coordinates of the character shape or by an abstract representation with nodes corresponding to the strokes and edges corresponding to the relationships between the strokes. Trees can also be used to represent the words or characters with a set of features, which has a hierarchical relation [36].

3.2 Training: An Artificial Neural Network as the backend can use for performing classification, training and recognition task. Support Vector Machine (SVM) had been developed by Vapnik in the framework of Statistical Learning Theory. We can use SVM classifier, Feed Forward, MLPs, Hopfield Network for training the system.

3.3 Classification: The feature vector obtained from previous phase is assigned a class label and recognized using supervised and unsupervised method. The data set is divided into training set and test set for each character. Character classifier can be one or more of the following, Bayes classifier, Nearest neighbour classifier, Radial basis function, Support vector machine, MLP, Quadratic classifier, Linear, Modified discriminant functions, Gaussian distribution function, KNN, and Neural networks with or without back propagation. A number of classification methods were purposed by different researchers some of these are template matching, statistical methods, syntactic methods, artificial neural networks, kernel methods.

3.3.1 Template matching: This is one of the simplest approaches to patter recognition. In this approach a prototype of the pattern that is to be recognized is available. Now the given pattern that is to be recognized is compared with the stored patterns. The size and style of the patterns is ignored while matching.

3.3.2 Statistical methods: The purpose of the statistical methods is to determine to which category the given pattern belongs. By making observations and measurement processes, a set of numbers is prepared, which is used to prepare a measurement vector. Statistical classifiers are automatically trainable. The *k*-NN rule is a non parametric recognition method. This method compares an unknown pattern to a set of patterns that have been already labeled with class identities in the training stage. A pattern is identified to be of the class of pattern, to which it has the closest distance.

3.3.3 Syntactic or structural methods: Syntactic methods are good for classifying hand written texts. This type of classifier, classifies the input patterns on the basis of components of the characters and the relationship among these components. Firstly the primitives of the character are identified and then strings of the primitives are checked on the basis of pre-decided rules. Generally a character is represented as a production rules structure, whose left-hand side represents character labels and whose right-hand side represents string of primitives. The right-hand side of rules is compared to the string of primitives extracted from a word. So classifying a character means finding a path to a leaf.

3.3.4 Artificial neural networks: A neural networks composed of inter connected elements called neurons. A neural network can trained itself automatically on the basis of examples and efficient tools for learning large databases. This approach is non-algorithmic and is trainable. The most commonly used family of neural networks for pattern classification task is the feed-forward network, which includes multilayer perception and Radial-Basis Function (RBF) networks. The other neural networks used for classification purpose are Convolutional Neural Network, Vector Quantization (VQ) networks, auto-association networks, Learning Vector Quantization (LVQ). But the limitation of the systems based on neural networks is their poor capability for generality.

3.3.5 Kernel methods: Some of the most important Kernel methods are Support Vector Machines , Kernel Principal Component Analysis (KPCA), Kernel Fisher Discriminant Analysis (KFDA) etc. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification. In a classification task usually data is divided into training and testing sets. The aim of SVM is to produce a model, which predicts the target values of the test data. Different types of kernel functions of SVM are: Linear kernel, Polynomial kernel, Gaussian Radial Basis Function (RBF) and Sigmoid.

3.4 *Matching techniques* : After the classification the matching will be performed on the trained data set with the help of algorithm.

Step 5: Character recognition and repeat the steps for the entire characters.

Conclusion:

In this paper, we have presented a survey of preprocessing, segmentation, feature extraction, classification and matching techniques for optical character recognition of general scripts. We can make use of these techniques for recognition of Marathi manuscript as well as for other languages also. A lot of research has been done in this field. Still the work is going on to improve the accuracy of the above techniques. However the different methods of preprocessing, segmentation, feature extraction, classification and matching techniques discussed here are very effective and useful for new researchers.

References:

[1] O. V. Ramana Murthy, Sujoy Roy, Vipin Narang, M. Hanmandlu, "Devanagari Character Recognition in the Wild*", International Journal of Computer Applications (0975 – 8887) Volume 38– No.4, January 2012* .

[2] Prathima G and Guruprasad K S Rao, "A Survey of Nandinagari Manuscript Recognition System", *International Journal of Science & Technology ISSN (online): 2250-141X www.ijst.co.in Vol. 1 Issue 1, November 2011.*

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

[4] 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, pp. 496-500, Sept. 23-26, 2007.*

[5] Debashis Ghosh, Tulika Dube, and Adamane P. Shivaprasad, "Script Recognition - A Review", *IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 32, No. 12, December 2010.*

[6]B.V.Dhandra et. al., "Word-wise Script Identification from Bilingual Documents Based on Morphological Reconstruction", *IEEE IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 32, No. 12, December 2006.*

[7]Albert Hung-Ren Ko et. al., "Leave-One-Out-Training and Leave-One-Out-Testing Hidden Markov Models for a Handwritten Numeral Recognizer: The Implications of a Single Classifier and Multiple Classifications", *IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 12, December 2009.*

[8] U. Pal and Partha Pratim Roy, "Multioriented and Curved Text Lines Extraction From Indian Documents", *IEEE Transactions On Systems, Man, And Cybernetics - Part B: Cybernetics, Vol. 34, No. 4, August 2004.*

[9]M. K. Jindal, R. K. Sharma, G. S. Lehal, "A Study of Different Kinds of Degradation in Printed Gurmukhi Script", *Proceedings of the International Conference on Computing: Theory and Applications, 2007.*

[10] A Bharath and Sriganesh Madhvanath, "HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 4, April 2012.*

[11]M. K. Mohahmed Althaf, M. Baritha Begum, "Handwritten Characters Pattern Recognition using Neural Networks", *International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.*

[12]U. Pal and B. B. Chaudhari, "Indian Script Character Recognition: a Survey", *Pattern Recognition, vol. 37, p. 1887-1899, 2004.*

[13]Shailendra Kumar Shrivastava and Pratibha Chaurasia, " Handwritten Devanagari Lipi using Support Vector Machine*", International Journal of Computer Applications (0975 – 8887) Volume 43– No.20, April 2012.*

[14]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, pp. 215-219, Nov. 16-18, 2010.*

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

[16]Sushama Shelke, Shaila Apte, "Multistage Handwritten Marathi Compound Character Recognition Using Neural Networks", *Journal of Pattern Recognition Research, pp. 253-268, August 2011.*

[17] Vikas J Dongre, Vijay H Mankar, "Devnagari Document Segmentation Using Histogram Approach", *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 1, No. 3, August 2011.*

[18] C. Y. Suen, "Character recognition by computer and applications", in Handbook of Pattern Recognition and Image Processing, New York: Academic, pp. 569-586, 1986.

[19] 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.

[20] Shailendra Kumar Shrivastava and Pratibha Chaurasia, " Handwritten Devanagari Lipi using Support Vector Machine",International Journal of Computer Applications (0975 – 8887) Volume 43– No.20, April 2012.

[21] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier and C. Olivier, structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, Vol. 19(7), pp. 629-641, 1998.

[22]M. K. Hu, "Visual pattern recognition by moment invariants", IRE Transactions on Information Theory, Vol. 8(2), pp. 179-187, 1962.

[23]S. S. Reddi, "Radial and angular moment invariants for image identification", IEEE Transactions on PAMI, Vol. 3(2), pp. 240-242, 1981.

[24] C. H. Teh and R. T. Chin, "On image analysis by the method of moments", IEEE Transactions on PAMI, Vol. 10(4), pp. 496-513, 1988.

[25] C. Y. Suen, M. Berthod and S. Mori, "Automatic recognition of hand printed characters- the state of the art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487, 1980.

[26] A. B. S. Hussain, G. T. Toussaint and R. W. Donaldson, "Results obtained using a simple character

recognition procedure on Munson's handprinted data", IEEE Transactions on Computers, pp. 201-205, 1972.

[27] R. Tarling and R. Rohwer, "Efficient use of training data in the n-tuple recognition method", Electronics Letters, Vol. 29(24), pp. 2093-2094, 1993.

[28] D. Trier, A. K. Jain, T. Taxt, "Feature Extraction Method for Character Recognition – A Survey*", Pattern recognition*, vol.29, no.4, pp.641-662, 1996.

[29] Santanu Chaudhury, Geetika Sethi, Anand Vyas, Gaurav Harit, "Devising Interactive Access Techniques for Indian Language Document Images", *(ICDAR 2003).*

[30] U. Pal, T. Wakabayashi, F. Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers", *10th Intl. Conf. on Document Analysis and Recognition,* pp. 1111-1115, 2009.

[31] Tapan K Bhowmik, Swapan K Parui Utpal Roy, "Discriminative HMM Training with GA for Handwritten Word Recognition", *IEEE,* 2008.

[32] B.V.Dhandra, Mallikarjun Hangarge, "Global and Local Features Based Handwritten Text Words and Numerals Script Identification", *Intl. Conf. on Computational Intelligence and Multimedia Applications*, PP 471-475. 2007.

[33] Latesh Malik, P.S. Deshpande , "Recognition of printed Devnagari characters with regular expression in finite state models", *International workshop on machine intelligence research,* GHRCE Nagpur, India, 2009.

[34] M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madasu, "Fuzzy Model based recognition of handwritten Hindi characters*", Digital Image Computing Techniques and Applications 0-7695-3067-IEEE*. Feb-07.

[35] Reena Bajaj, Lipika Dey , Santanu Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", *Sadhana* Vol. 27, Part 1, pp. 59–72, February 2002.

[36] Canasai Kruengkrai, Virach Sornlertlamvanich, Hitoshi Isahara, " Language, Script, and Encoding Identification with String Kernel Classifiers", *Thai Computational Linguistics Laboratory*, Thailand.