IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE)
e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 9, Issue 2 Ver. II (Mar – Apr. 2014), PP 80-87
www.iosrjournals.org

Handwritten Devanagari Script Recognition: A Survey

Aradhana A Malanker [1], Prof. Mitul M Patel [2]
[1](EC Department, Gujarat Technological University, India)
2(EC Department, Gujarat Technological University , India)

Abstract : *In India, many people use Devanagari script for documentation. There has been a significant improvement in the research related to the recognition of printed as well as handwritten Devanagari text in the past few years. Basically Character recognition techniques associate a symbolic identity with the image of a character. Since creating an algorithm with a one hundred percent correct recognition rate is quite probably impossible in our world of noise and different font styles, it is important to design character recognition algorithms with these failures in mind so that when mistakes are inevitably made, they will at least be understandable and predictable to the person working with the program. An attempt is made to address the most important results reported so far and it is also tried to highlight the beneficial directions of the research till date.*

Keywords: *Devanagari Optical Character Recognition (OCR), handwritten character recognition, offline character recognition, Segmentation, Feature Extraction, Classification.*

I. INTRODUCTION

Character recognition is an art of detecting segmenting and identifying characters from image. More precisely Character recognition is process of detecting and recognizing characters from input image and converts it into ASCII or other equivalent machine editable form. It contributes immensely to the advancement of automation process and improving the interface between man and machine in many applications. Character recognition is getting more and more attention since last decade due to its wide range of application. Conversion of handwritten characters is important for making several important documents related to our history, such as manuscripts, into machine editable form so that it can be easily accessed and preserved.

Application of Offline Handwritten Character Recognition System: Some of the important applications of offline handwritten character recognition are listed in the following section:

(1) Recognition Of Ancient Document

The historical documents are currently being digitalized for prevention purpose and to make them available worldwide through large on-line digital libraries.

(2) Cheque Reading

Offline handwritten Character recognition is basically used for cheque reading in banks. Cheque reading is very important commercial application of offline handwritten character recognition. Handwritten character recognition plays very important role in banks for signature verification and for recognition of amount filled by user.

(3) Postcode Recognition

Handwritten character recognition system can be used for reading the handwritten postal address on letters. Offline handwritten character recognition system used for recognition handwritten digits of postcode. HCR can be read this code and can sort mail automatically.

(4) Form processing

HCR can also be used for form processing. Forms are normally used for collecting public information. Replies of public information can be handwritten in the space provided.

(5) Signature verification

HCR can also used for identify the person by signature verification. Signature identification is the specific field of handwritten identification in which the writer is verified by some specific handwritten text.

Basically Character recognition techniques associate a symbolic identity with the image of a character. To make the complicated process of online handwritten character recognition easier and more robust, focus should be on salient features of character. After pre-processing feature extraction is done. Feature extraction is a vital phase of character recognition. Features extracted from character encode the structural characteristics of character shape.

This field is broadly divided into two parts,

(1) Online character recognition
(2) Offline character recognition.

Off-line Character recognition further divided into two parts,
a) Machine printed character recognition
b) Handwritten character recognition

(1) Online Character Recognition:
In online character recognition, characters are recognized at real time as soon as it is written. Online systems perform better than offline recognition as they have timing information and since they avoid the initial search step of locating the character. Online systems obtain the position of the pen as a function of time directly from the interface. This is usually done through pen-based interfaces where the writer writes with a special pen on an electronic tablet.

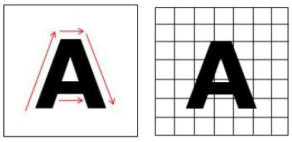


Fig 1.1: (a) Online Character Recognition (b) Offline Character Recognition

(2) Offline Character Recognition:
Offline character recognition can be further classified to printed characters and handwritten character recognition. In offline character recognition, the typewritten/handwritten character is typically scanned in form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm. Offline character recognition become more challenging due to shape of characters, great variation of character symbol and document quality. Therefore offline character recognition is considered as a more challenging task then its online counterpart.

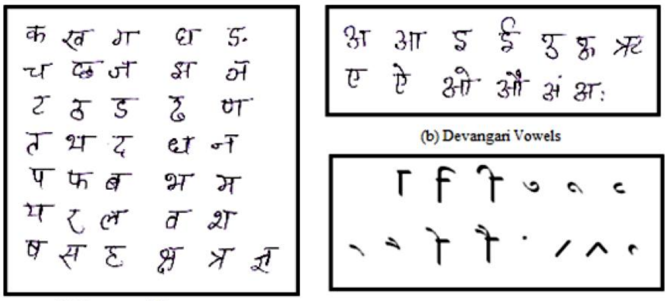
Table I: Comparison between online and offline handwritten characters

Sr. No.	Comparisons	Online characters	Offline characters
1.	Availability of strokes	Yes	No
2.	Data requirement	Samples/second	Dots/inch
3.	Way of writing	Digital pen on LCD	Paper document
4.	Recognition rates	Higher	Lower
5.	Accuracy	Higher	Lower

Technologist faces challenges in developing offline character recognition. Firstly, for reading a page in unknown language, anyone may be unable to recognize the various characters. But on the same page, numerical statements can be easily interpreted because the symbols for numbers are universally used. This explains why many OCR systems recognize numbers only, while relatively few understand the full alphanumeric character range.

Introduction of Devanagari Script

Devanagari also called Nagari (Nāgarī, नागरी, the name of its parent writing system), is an abugida alphabet of India and Nepal . It is written from left to right, does not have distinct letter cases, and is recognizable by a horizontal line that runs along the top of full letters. Devanagari is being used for writing not only Sanskrit and Hindi but also Marathi, Koknani, Rajasthani, Nimadi etc. Devanagari is the main script used to write Standard Hindi, Marathi, and Nepali. Since the 19th century, it has been the most commonly used script for Sanskrit. Devanagari is also employed for Bhojpuri, Gujarati, Pahari, (Garhwali and Kumaoni), Konkani, Magahi, Maithili, Marwari, Bhili, Newari, Santhali, Tharu, and sometimes Sindhi, Dogri, Sherpa and by Kashmiri-speaking Hindus. It was formerly used to write Gujarati. The script has a complex composition of its constituent symbols. Devanagari script (Hindi) has 13 vowels and 36 consonants shown in the Fig. 2. They are called basic characters. All the characters have a horizontal line at the upper part, known as Shirorekha or headline.


(a) Devanagari Consonants

(b) Devangari Vowels

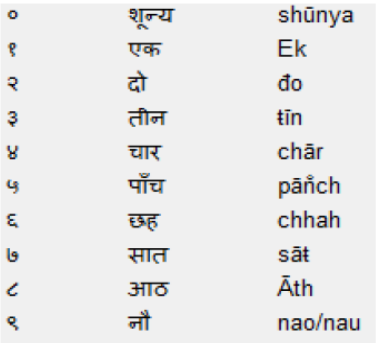
(c) Modifiers

Fig 1.2:Devanagari Isolated Handwritten Characters, Modifiers

Offline handwritten Character recognition is basically used for cheque reading in banks. Cheque reading is very important commercial application of offline handwritten character recognition. Handwritten character recognition plays very important role in banks for signature verification and for recognition of amount filled by user.

Follow are main reasons for difficulty in recognition of Devanagari Characters:-

- Some Devanagari character similar in shape.
- Different or even the same writer can write differently times depending upon pen or pencil.
- The character can be written at different location on paper or its window.
- Character can be written in different fonts .



०	शून्य	shūnya
१	एक	Ek
२	दो	đo
३	तीन	ŧīn
४	चार	chār
५	पाँच	pān̊ch
६	छह	chhah
७	सात	sāŧ
८	आठ	Āth
९	नौ	nao/nau

Fig 1.3: Sample of Devanagari numerals

II. Working Principle

Any character recognition system goes under following steps, i.e. Image acquisition, Preprocessing, Segmentation, Feature extraction, classification and post processing.[6] Block diagram of general character recognition system is shown in Figure 4

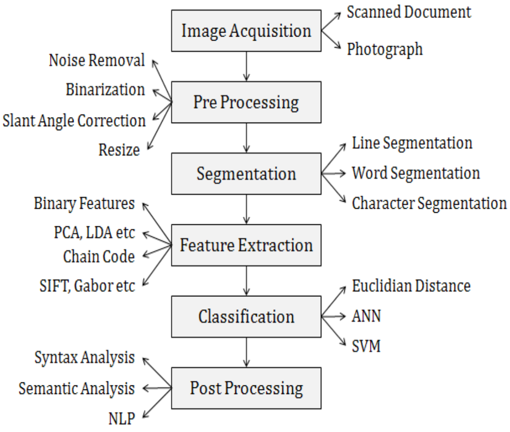


Fig 2.1: Block Diagram of Character Recognition System

1] Image Acquisition:

In Image acquisition, the recognition system acquires a scanned image as an input image. The image should have a specific format such as .jpeg, .bmp etc. This image is acquired through a scanner, digital camera or any other suitable digital input device.

2] Preprocessing;

Preprocessing involves series of operations performed to enhance to make it suitable for segmentation. Preprocessing step involves gray into binary conversion by using threshold value obtained by Otsu's method noise removal generated during document generation[3]. Proper filter like mean filter, min-max filter, Gaussian filter etc may be applied to remove noise from document. Binarization process converts gray scale or colored image to black and white image. Binary morphological operations like opening, closing, thinning, hole filling etc may be applied to enhance visibility and structural information of character. If document is scanned then it may not be perfectly horizontally aligned, so we need to align it by performing slant angle correction. Input document may be resized if it is too large in size to reduce dimensions to improve speed of processing. However reducing dimension below certain level may remove some useful features too.

3] Segmentation:

Generally document is processed in hierarchical way. At first level lines are segmented using row histogram. From each row, words are extracted using column histogram and finally characters are extracted from words. Accuracy of final result is highly depends on accuracy of segmentation.

4] Feature Extraction:

Feature extraction is the heart of any pattern recognition application. Feature extraction techniques like Principle Component Analysis (PCA), Linear Discriminates Analysis (LDA), Independent Component Analysis (ICA), Chain Code (CC), zoning, Gradient based features, Histogram might be applied to extract the features of individual characters. These features are used to train the system.

5] Classification:

When input image is presented to HCR system, its features are extracted and given as an input to the trained classifier like artificial neural network or support vector machine. Classifiers compare the input feature with stored pattern and find out the best matching class for input.

Types of Classifiers

Support Vector Machine (SVM): Support vector machine is supervised learning tool, which is used for classification and regression. The basic SVM takes a set of input data and predicts, for each given input. Given a set of training examples, each marked as belonging to one of two categories, SVM training algorithm builds a model that assigns new examples into one category or the other. More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class since in general the larger the margin the lower the generalization error of the classifier.

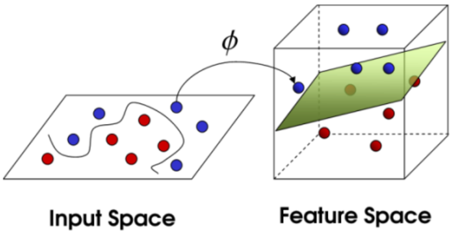


Fig 2.2: Feature transformation

Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

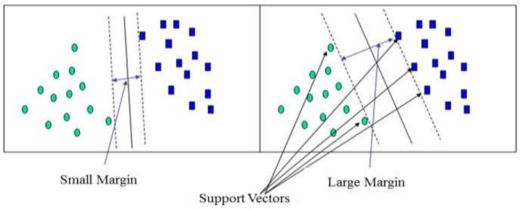


Fig 2.3: Maximum margin classification by SVM

Artificial Neural network(ANN): Artificial neural network is widely accepted classifier for diverse patterns. ANN works on phenomenon of biological neurons and learns to classify unseen data. Multilayer neural networks have been widely used in pattern recognition applications. Various paradigms have been used. The different network models are specified by:

Network topology: the number of neurons and how the neurons are interconnected.

Node characteristics: the type of non-linear transfer function used by the neuron for calculating the output value.

Training rules: specify how the weights are initially set and adjusted to improve performance of the network.

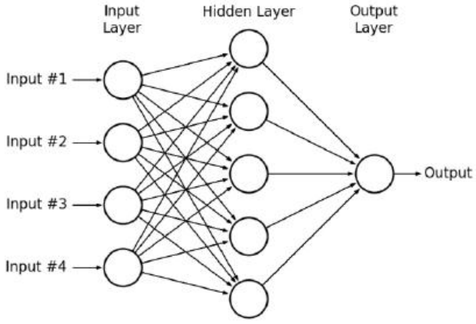


Fig 2.4: Model of ANN

Architecture of neural network depends on nature and complexity of applications. However multilayer neural network with proper choice of parameter is capable enough to classify almost any pattern. There are so many parameters that control the performance of neural network, like

- Number of layers
- Number of neurons in each layer
- Transfer function used between two layers
- Learning algorithm
- Number of epochs.

Hidden Markov Models (HMM): A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be considered the simplest dynamic Bayesian network. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden’. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, and bioinformatics. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other.

III. Recognition Of Handwritten Devanagari Script

Only during recent years, research toward Indian handwritten character recognition is getting increased attention although the first research report on offline handwritten Devanagari characters was published in 1977. Many approaches have been proposed toward handwritten Devanagari numeral, character, and word recognition in the past decade [7]. To have an idea about the shape of handwritten characters, Fig. 4 shows some handwritten numerals, characters, and sentences in Devanagari. Two approaches are mainly used in handwritten character recognition. First is segmentation-based approach and the other is segmentation-free approach (holistic approach). In the first approach, the words are initially segmented into characters or pseudo characters, and then, recognized. As a result, the success of the recognition module depends on the performance of the segmentation technique. The second approach treats the whole word as a single entity and it recognizes without doing explicit segmentation.

A. Pre-processing and Segmentation Techniques

Devanagari word image is first smoothed using a median filter, and then, binarized by Otsu’s [3] thresholding method. The binarized image is then smoothed using a median filter. Noise removal of the document is also an important step toward the recognition. Bajaj *et al.* [8] used a median filtering-based Approach for noise removal from the images of handwritten Devanagari numerals.

B. Features Considered

Even though researchers test different features, statistical and structural features are mostly used for handwritten numeral/ character recognition. Bajaj *et al*. [8] represented each handwritten Devanagari numeral using three types of features: 1) density features; 2) moment features of right, left, upper, and lower profile curves; and 3) descriptive component features. For extracting the features, a box approach is proposed by Hanmandlu *et al*. [8], [9] for handwritten numbers, which requires a spatial division of the numeral image into boxes. Ramteke and Mehrotra [10] evaluated the performance of various techniques based on moment invariants on handwritten Devanagari numerals. The features that have been extracted are based on moments, image partition, principal component axes, correlation coefficient, and perturbed moments. Kumar [11] compared performances of five feature-extraction methods on handwritten characters. The various features covered are Kirsch directional edges, distance transform, chain code, gradient, and directional distance distribution. From the experimentations, it is found that Kirsch directional edges are least performing and gradient is best performing with SVM classifiers. With multilayer perceptrons (MLP), the performance of gradient and directional distance distribution is almost same. The chain-code-based feature is better as compared to Kirsch directional edges and distance transform. A new feature is also proposed in the paper, where the gradient direction is quantized into four-directional levels and each gradient map is divided into 4 × 4 regions. This is combined with total distances in four directions and neighbourhood pixels weight In [15], the features are extracted from handwritten Devanagari characters using a box approach. Each character image is divided into 24 boxes. The features are represented using normalized vector distances for each character.

The shirorekha and spine in a handwritten character are detected using a differential-distance-based technique. Also features like crossing points, end points, and corners are also considered in the same work. The main features for handwritten Devanagari characters considered in [19] are the CH features, four side views based, and shadow-based features. The features used by Pal *et al.* [12] for handwritten characters are mainly based on directional information obtained from the arc tangent of the gradient and Gaussian filter. In [7], a comparative study of Devanagari handwritten character recognition using 12 different classifiers and four sets of features is presented. Feature sets used in the classifiers are computed based on curvature and gradient information obtained from binary as well as gray-scale images.

C. Recognition/Classification Methods Used

Bajaj *et al*. [8] combined decisions of multiple classifiers for handwritten Devanagari numerals. A neural network-based classification scheme is designed for this task. Three different neural classifiers have been used for classification. The outputs of the three classifiers are combined using a connectionist scheme. Hanmandlu *et al.* [9] proposed a fuzzy model-based scheme for recognition of handwritten Devanagari numerals by representing them in the form of exponential membership functions, which serve as a fuzzy model. Modifying the exponential membership functions fitted to the fuzzy sets does the recognition. These fuzzy sets are derived from features consisting of normalized distances obtained using the Box approach. The Gaussian distribution function has been adopted by Ramteke and Mehrotra [10] for classification of handwritten numerals. In [18], a method is proposed based on cubic spine interpolation for determining smooth and continuous edges in the images of handwritten Devanagari numerals.

Table II Comparison of Numeral Results by Researchers

Method	**Feature**	**Classifier**	**Data Set (Size)**	**Accuracy (%)**
Bajaj et al. [8]	Statistical	Neural Network	2,460	89.68
Ramteke et al.[10]	Moment Invariants	Gaussian Distribution	2,000	92
Lakshmi et al.[18]	Gradient	PCA	9,800	94.25
Hanmandlu et al.[9]	Box Approach	Fuzzy Model	3,500	95
Hanmandlu et al.[13]	Box Approach	Bacterial Foraging	3,500	96
Sinha et al.[3]	Zone Approach	SVM	4,832	99.11

Table III Comparison of Character Results by Researchers.

Method	Feature	Classifier	Data Set (Size)	Accurac y (%)
Arora et al.[19]	Combined	MLP	1,500	89.58
Aggarawal et al.[]	Gradient	SVM	7,200	94
Kumar et al.[11]	Gradient	SVM	25,000	94.1
Pal et al.[12]	Gradient & Gaussian Filter	Quadratic	36,172	94.24
Pal et al.[14]	Gradient	SVM & MQDF	36,172	95.13
Pal et al.[7]	Gradient	MIL	36,172	95.19

IV. CONCLUSION

With the advent of computer and information technology, there has been a dramatic increase in research in the field of Devanagari OCR since 1990. Different strategies using combination of multiple features, multiple classifiers, and several templates have been widely considered in the prior art. Only a few studies have been reported in the areas of restrictions Devanagari script recognition. Only a few papers are published in identifying scripts. Overall, researchers assume that a given document is written in a particular script. In countries like India where there are many languages and alphabets , script identification must be made prior to the recognition in applications such as reader address, the address where you can write in any script in India. More research in this direction in handwritten documents, in the near future is expected.

In India, large volumes of historical documents and books (handwritten or printed in Devanagari script) has not yet been digitized for easy access, sharing, indexing, etc. This will definitely be useful to other research communities in India in the areas of social sciences, economics and linguistics.

Handwritten character recognition is still a research area of burning pattern recognition. Each and every step that directly contributes to the accuracy of the system, as pre - processing, segmentation, feature extraction, training methods, etc. all. So all these area are open for independent research.

Indian national language Hindi (written in Devanagari script) is the third most popular language in the world after Chinese and English. Consequently, research on Devanagari script is gaining much attention due to its large market potential. Some of the leading institutes in India doing research on Devanagari OCR are the Indian Statistical Institute in Kolkata, International Institute of Information Technology in Hyderabad, the Indian Institute of Science in Bangalore and the Indian Institute of Technology in New Delhi.

Acknowledgements

I am sincerely thankful to my guide **Asst. Prof. Mitul M. Patel** whose constant help, stimulating suggestions and encouragement helped me a lot to reach up to this stage .I am highly indebted to him for his guidance and constant supervision as well as for providing necessary information regarding the project .We are thankful to IOSR Journal for the support to develop this document.

REFERENCES

[1] Ved Prakash Agnihotri, "Offline Handwritten Devanagari Script Recognition" I.J. Information Technology and Computer Science, July 2012

[2] Gita Sinha, Mrs. Rajneesh Rani, Prof. Renu Dhir, "Handwritten Devanagari Numeral Recognition Using Zonal Based Feature Extraction Method and SVM Classifier", : IJARCSSE, June 2012

[3] Yi-Kai Chen, Jhing-Fa Wang, "Segmentation of Single- or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis", IEEE , NOVEMBER 2000

[4] Ashutosh Aggarwal, Mrs.Rajneesh Rani, Prof. Renu Dhir,"Handwritten Devanagari Character Recognition Using Gradient Features [",IJARCSSE,, May 2012

[5] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey" , IEEE, NOVEMBER 2011

[6] MANSI SHAH AND GORDHAN B JETHAVA, "A LITERATURE REVIEW ON HAND WRITTEN CHARACTER RECOGNITION" Indian Streams Research Journal Vol -3 , ISSUE –2, March.2013

[7] U. Pal, T. Wakabayashi, and F. Kimura, "Comparative study of Devanagari handwritten character recognition using different features and classifiers," in Proc. 10th Conf. Document Anal. Recognit., 2009

[8] R. Bajaj, L. Dey, and S. Chaudhuri, "Devanagari numeral recognition by combining decision of multiple connectionist classifiers," Sadhana, vol. 27, 2002.

[9] M. Hanmandlu and O. V. R. Murthy, "Fuzzy model based recognition of handwritten numerals," Pattern Recognit., vol. 40,2007.

[10] R. J. Ramteke and S. C. Mehrotra, "Feature extraction based on moment invariants for handwriting recognition," in Proc. IEEE Conf. Cybern. Intell. Syst., 2006

[11] S. Kumar, "Performance comparison of features on Devanagari handprinted dataset," Int. J. Recent Trends, vol. 1, 2009.

[12] U. Pal, N. Sharma, T.Wakabayashi, and F. Kimura, "Off-line handwritten character recognition of Devnagari script," in Proc. 9th Conf. Document Anal. Recognit., 2007

[13] M. Hanmandlu, A. V. Nath, A. C. Mishra, and V. K. Madasu, “Fuzzy model based recognition of handwritten hindi numerals using bacterial foraging,” in Proc. Int. Conf. Comput. Inf. Sci., 2007.

[14] U. Pal, S. Chanda, T. Wakabayashi, and F. Kimura, “Accuracy improvement of Devnagari character recognition combining SVM and MQDF,” in Proc. 11th Int. Conf. Frontiers Handwrit. Recognit., 2008.

[15] M. Hanmandlu, O. V. R. Murthy, and V. K. Madasu, “Fuzzy Model based recognition of handwritten Hindi characters,” in Proc. Int. Conf. Digital Image Comput. Tech. Appl., 2007.

[16] S. Kumar, “An analysis of irregularities in Devanagari script writing: A machine recognition perspective,” Int. J. Comput. Sci. Eng., vol. 2,2010.

[17] P. B. Pati and A. G. Ramakrishnan, “A blind indic script recognizer for multi-script documents,” in Proc. 9th Conf. Document Anal. Recognit.,2007.

[18] C. V. Lakshmi, R. Jain, and C. Patvardhan, “Handwritten Devnagari numerals recognition with higher accuracy,” in Proc. Int. onfut.Intell. Multimedia Appl., 2007.

[19] S. Arora, D. Bhattacharjee, M. Nasipuri, D. K. Basu, M. Kundu, and L. Malik, “Study of different features on handwritten Devnagari character,” in Proc. 2nd Emerging Trends Eng. Technol., 2009.