Optical character recognition for printed text in Devanagari using ANFIS

1) Prof. Sheetal A. Nirve
Dept. Of Electronics & Comm.
DIEMS, Aurangabad,
Maharashtra, India
Sheetalnirve0504@gmail.com

2) Dr. G. S. Sable
Principal of savitribai phule women's engineering college, sharnapur
Aurangabad, Maharashtra,
India

Abstract: In India, more than 300 million 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. 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**.** In this paper we propose an efficient image retrieval technique which uses dominant color and texture features of an image. Though, Affine Moment invariant technique is well experimented by many researchers, an attempt is made to enhance the existing results by extracting various supportive features like moments invariant, vector Gradient, chain code(freeman chain code) image thinning, structuring the image in box format, noise removal, etc. A performance of approximately 90% correct recognition is achieved.

Keywords— Optical Hindi character recognition (OCR), Data set, Affine Moment invariants Rotation, neural network (NN) training of NN, Recoganisation.

1. INTRODUCTION

Character recognition is the process to classify the input character according to the predefine character class, with increasing the interest of computer applications, modern society needs the input text into computer readable form. This research is a simple approach to implement that dream as the initial step to convert the input text into computer readable form. Some research for hand written characters are already done by researchers with artificial neural networks. Digital document processing is gaining popularity for application to office and library automation, bank and postal services, publishing houses and communication technology. English Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. But same is not the case for Indian languages which are complicated in terms of structure and computations. In OCR domain, it is now widely accepted that a single feature extraction method and single classification algorithm can't yields better performance rate. Neural networks and fuzzy logic are two complimentary technologies which are used in pattern recognition process. There are two type of neural network, feedback and feed forward.It is therefore, a compound feature extraction approach based on soft computing for recognition of printed Marathi vowels and consonants is proposed. The OCR has been tested on samples from various magazines and newspapers.

2. LITERATURE SURVEY

The United States Postal Services has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rainbow. In 1965 it began planning an entire banking system, National Giro, using OCR technology, a process that revolutionized bill payment systems in the UK. Then in 1970's efforts were initiated by Sinha at Indian Institute of Technology, Kanpur. A syntactic pattern analysis system for Devanagari script recognition is presented in Sinha's Ph.D. thesis. AnotherOCRsystem development of printed Devanagari is by Palit and Chaudhuri as well as Pal and Chaudhuri. A team comprising Prof. B. B. Chaudhuri, U. Pal, M. Mitra, and U. Garain of Indian Statistical Institute, Kolkata, developed the first commercial level product for printed Devanagari OCR. The same technology has been transferred to Center for Development for the Advance Computing (CDAC) in 2001 for commercialization and is marketed as "Chitrankan". An approach based on the detection of "shirorekha" is proposed by Chaudhuri and Pal with the assumption that the skew of such header lines show the skew of the whole document. Initially the

connected component labeling in the document is done. The required estimate of skew angle is obtained by averaging of angles between the horizontal line and the lines joining the first pixel of the leftmost line segment and the last pixel of the rightmost line segment of each line. There are some documents called multi skew documents in which the text lines are not parallel to each other. An approach, which is an extension to that of is proposed for skew estimation in multi skew documents. Das and Chanda also proposed a fast and script-independent skew estimation technique based on mathematical morphology. Garain and Chaudhuri presented another technique for identification and segmentation of touching machine-printed Devanagari characters. Garain and Chaudhuri presented another technique for identification and segmentation of touching machine-printed Devanagari characters.

.

3. SYSTEM DEVELOPMENT

The letter order of Devanagari, like nearly all Brahmi scripts, is based on phonetic principles that consider both the manner and place of articulation of the consonants and vowels they represent. This arrangement is usually referred to as the varanmala "garland of letters".

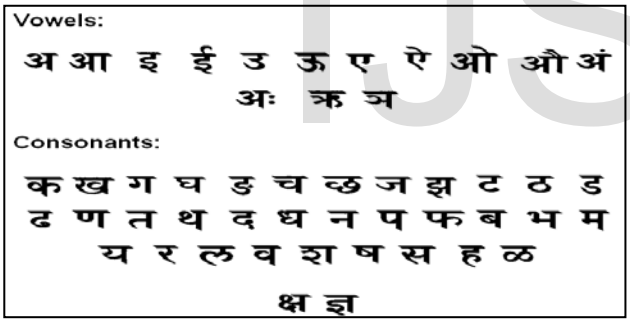


Fig.3.1 (a)

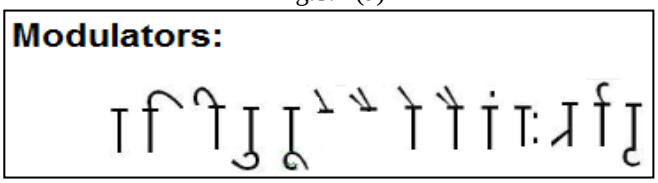


Fig. 3.1 (b)

Fig. 3.1 Vowels, consonants and modulators of Devanagari script

Devanagari script has about 14 vowels and 33 consonants. As well as some modifiers are used to modifie the vowels and consonants to form new words or sentences. The vowels and the consonants are shown in figure 3.1.l. (a) and (b) respectively. In English as well as in Marathi, the vowels are used in two ways:

1. They are used to produce their own sounds. The vowels shown in are used for this purpose in Devanagari.

2. They are used to modify the sound of a consonant.

A consonant in pure form always touches the next character, yielding conjuncts, touching characters, or fused characters. Fig3.1.2 (a) shows the consonant with modifiers and Fig. 3.1.2 (b) shows some of the conjuncts formed by writing pure form consonants followed by consonant ' य'. We can use almost any consonant in place of ' य' and write over 100 conjuncts.

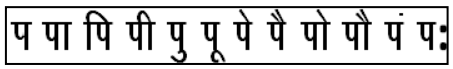


Fig. (a)

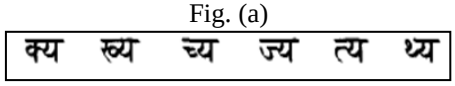


Fig. (b)

Fig. 3.2: Characters and Symbols of DevanagariScript; (a) the modifier symbols attached to the consonant to indicate 'प' their placing; (b) Some sample Conjunctions

3.1) Problem definition:

As we know that recognition of character is not easy task. Due to various font sizes and writing style it is difficult to recognize the character. Also in Devnagari script, many characters have similar shape, which creates trouble in recognition. There for instead of utilizing single feature we are using various features like GLCM, color dominant, affine movement invariant and Histogram.

3.2 Tools required for system:

3.2.1 MATLAB

MATLAB (matrix laboratory) is a numerical computing environment and fourth-generation programming language. Allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and Fortran. An additional package, Simulink, adds graphical multi-domain simulation and Model-Based Design for dynamic and embedded systems. For my project I am using MATLAB 2010b version is used.

3.2.2 GUI

i)GUI (Graphic user Interface):

MATLAB supports developing applications with graphical user interface features. It also has tightly integrated graph-plotting features. The structure of application m-files generated by the MATLAB GUI development environment and some Techniques for inclusion of Java components and Active X controls into MATLAB. In our project we use this GUI for reorganization of Devanagari character. In this GUI there are 7 buttons, 8 labels and 6 panels are used for different purpose.

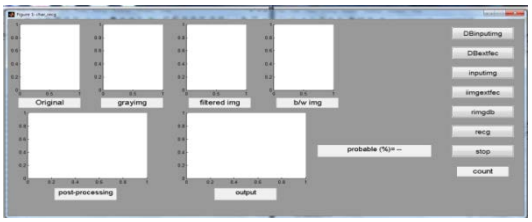


Fig 3.3: Used GUI

3.3 proposed character recognition system:

3.3.1 Data base generation:

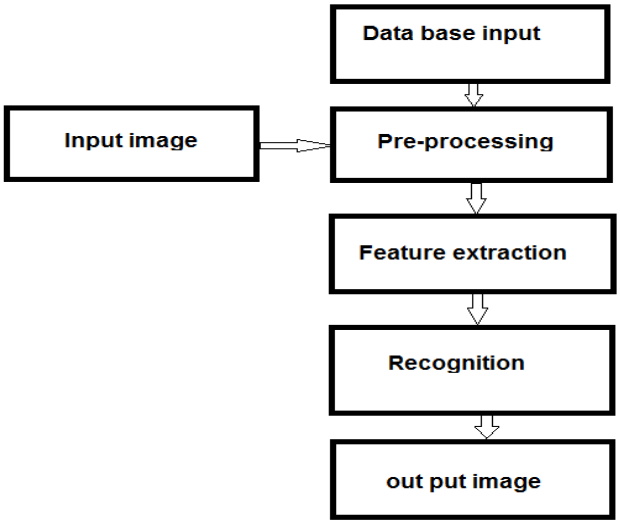


Fig. 3.4: Block dia. of proposed system

3.3.2 Image Pre-Processing:

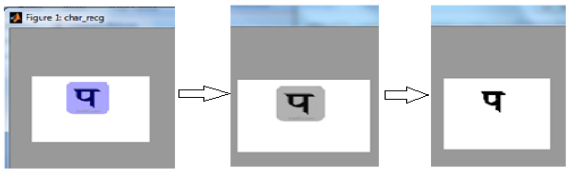
In imaging science, image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it.

i) Converting Color image to gray scale to binary image:

In present technology, almost all image capturing and scanning devices use colour. A colour image consists of a coordinate matrix and three colour matrices. Coordinate matrix contains x, y coordinate values of the image. The colour matrices are labelled as red (R), green (G), and blue (B). Techniques presented in this study are based on grey scale images, and therefore, scanned or captured colour images are initially converted to grey scale using the following equation:

Gray colour = 0.299*Red + 0.5876*Green +0.114*Blue

The scanned image was first converted from RGB scale to gray-scale. It was then splitted into individual character blocks using MATLAB script to obtain raw individual character samples. The following pre-processing and noise removal techniques were used on raw samples to obtain a clean dataset. For converting to binary threshold value is taken automatically.



i).Original image ii).Gray scale image iii).binary image

Fig. 3.5 conversion of original image

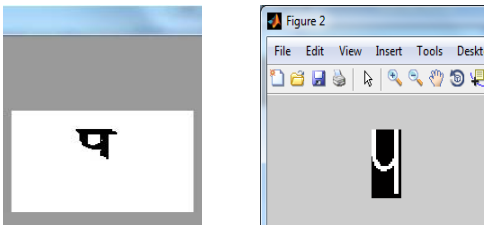
ii)Median Filtering :

Scanning process introduces irregularities such as speckle noise and salt and pepper noise in the output image. Noise reduction (also called smoothing or noise filtering) is one of the most important processes in image processing. Median Filter is used in this study due to its edge preserving feature.

iii)Removal of header line:

By using following command top line can be removed,

"[cut plan retimg] = remtopline(img,bwimg); "



i).Binary image ii).Top line removed

Fig. 3.7 Top line removed image

3.4 Feature extraction:

3.4.1 Exact computation of geometric moments:

Regular or geometric moments of order (p + q) for image intensity function f(x,y) are defined as m_{pq}=

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)dxdy, \quad (1)$$

with p,q >= 0. A digital image of size M × N is an array of pixels. Centers of these pixels are the points (xi,yj), where the image intensity function is defined only for this discrete set of points
(xi,yj) ϵ [0,M _ 1] × [0,N _ 1].
Δxi = xi+1 _ xi, Δyj = yj+1 _ yj are sampling intervals in the x- and y-directions, respectively. In the literature of digital image processing, the intervals Δxi and Δyj are fixed at constant values Dxi = 1, and Δyj = 1, respectively. Therefore, the set of points (xi,yj) will be defined as follows:

$x_i=(i-\frac{1}{2})\Delta x,$ (2.1)

$y_j=(j-\frac{1}{2})\Delta y,$ (2.2)

with i = 1, 2, 3, . . . , M and j = 1, 2, 3, . . . , N. For the discrete-space version of the image, Eq. (1) is usually approximated as

$\bar{M}_{[x]}=\sum_{i=1}^{M}\sum_{j=1}^{N} x_i^p y_j^q \Delta x\Delta y$ (3)

Eq. (3) is the so-called direct method for geometric moment's computations, which is the approximated version using zeroth-order approximation (ZOA). Eq. (3) is not a very accurate

approximation of Eq. (1).The set of geometric moments can be computed exactly by

$M_{[x]}=\sum_{i=1}^{M}\sum_{j=1}^{N} Ip(i)Iq(j)f(xi,yj)$, (4)

Where,

$Ip(i)=\frac{1}{p+1}[U_{i+1}^{p+1}-U_i^{p+1}]$, (5.1)

$Iq(j)=\frac{1}{q+1}[V_{i+1}^{p+1}-V_i^{p+1}]$, (5.2)

And

$U_{i+1}=x_i+\frac{\Delta xi}{2}$, (6.1)

$U_i=x_i-\frac{\Delta xi}{2}$, (6.2)

$Vj_{+1}=yj+\frac{\Delta yi}{2}$, (6.3)

$Vj=yj-\frac{\Delta yi}{2}$ (6.4)

3.4.2 Affine moment invariants:

Moments are one of the parameters that describe the image or object of interest. Moment invariants are moments which do not change under a group of transformations. Image normalization means bringing the image to a position in which the effect of transformation is eliminated. Affine transformation is represented by the following matrix form:

$$\begin{matrix} x' \\ y \end{matrix} = \begin{pmatrix} a11 & a12 \\ a21 & a22 \end{pmatrix} \begin{matrix} x \\ y \end{matrix} + \begin{matrix} b1 \\ b2 \end{matrix} \quad (7)$$

To achieve normalization, affine transformation decomposed into a group of simple one-parameter transforms. This group consists of translation, uniform scaling, first rotation, stretching, and second rotation

x'=x–x0,y'=y–y0 (8.1)

x'=αx,y=αy, (8.2)

x'=xcosθ–ysinθ,y'=xsinθ+cosθ, (8.3)

x'=δx,y'=1/δy, (8.4)

x'=xcosΦ–ysinΦ,y'=xsinΦ+ycosΦ (8.5)

where (x0,y0) is the centroid; α , δ > 0; θ, Φ are the rotation angles. The image function is invariant under the group of transformations (8) if and only if it is invariants under the general affine transformation (7).

3.5 GLCM (gray level co-occurrence matrix):

Following properties are calculated by GLCM :

`1)Contrast :` Returns a measure of the intensity contrast between a pixel and its neighbor over the whole image.

Range = [0 (size(GLCM,1)-1)^2]. Contrast is 0 for a constant image.

2)Correlation : Returns a measure of how correlated a pixel is to its neighbor over the whole image.

`Range = [-1 1].` Correlation is 1 or -1 for a perfectly positively or negatively correlated image. Correlation is `NaN` for a constant image.

`3)Energy :` Returns the sum of squared elements in the GLCM.

Range = [0 1]. Energy is 1 for a constant image.

`4)Homogeneity :` Returns a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Range = [0 1]. Homogeneity is 1 for a diagonal GLCM.

Now we will store all this feature in .mat file, to generate the database. This features are useful at the time of recognition of character.

3.6 Recognition algorithms :

3.6.1 ANFIS (Artificial neuro fuzzy interference system):

1)Neural Network

Neural network is also known as Artificial Neural Network (ANN), is an artificial intelligent system which is based on biological neural network. Neural networks able to be trained to perform a particular function by adjusting the values of the connections (weight) between these elements.

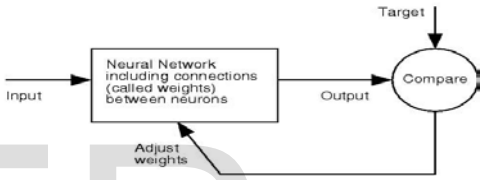


Fig. 3.10: Neural Network Block Diagram

Neural network is adjusted and trained in order the particular input leads to a specific target output. Example at Figure 3.5 the network is adjusted, based on a comparison of the output and the target until the network output is matched the target. Now a days, neural network can be trained to solve many difficult problems faced by human being and computer.

4. PERFORMANCE ANALYSIS

4.1 Database of characters :

Character recognition is done by generating database of devnagri character, extracting feature of each character and saving them for further recognition process. I have selected characters of different font size and font face for generating database. The following figure shows the data base used. Fig 4.1 shows the image used to generate database.

- Used Font size is 14.
- Used font face is kruti dev 014.

Fig. 4.1 Image used to generate database

Now each line is separated from complete figure as shown in the following figure. Then each character is separately cropped from single line. And simultaneously, centroid of each character is find out which is denoted by red cross and box around the character is represented by blue colour.

Fig. 4.2 separation of each character

Following MATLAB windows represents the samples taken to generate data base. The value of each character is stored in structure files and denoted by <1×1struct>. The structure consist of Area occupied by character, Centroid, Bounding Box, Eccentricity, Orientation, Binary values of image, and perimeter. Each character have unique value of this perimeters. All this values are stored in findchardata file. Format of Findchardata file is .mat file.

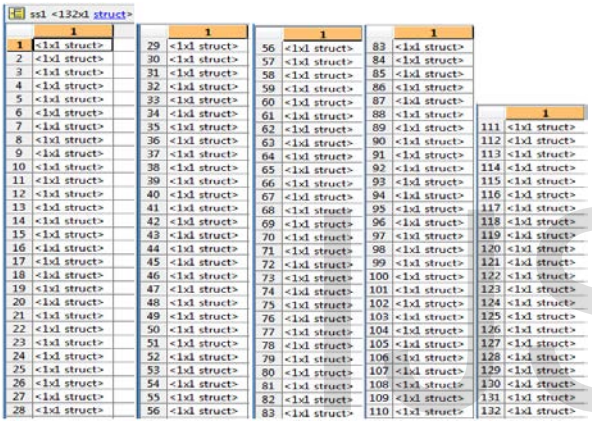


Fig.4.3 Sample taken to generate database

4.2 Feature extraction of character :

Each <1×1> struct present in the table consist of values of co-efficient of binary image, Area, centroid, eccentricity, orientation, bounding box, perimeter of single character. It is not possible to represent feature of all character in this report therefore we will represent two characters with their details of parameters.

Table 4.1 Parameters of character प

ss1(43,1) <1x1 struct>

Field	Value	Min	Max
Area	187	187	187
Centroid	[34.9412,89.6417]	34.9412	89.6417
BoundingBox	[24.5000,74.5000,17,34]	17	74.5000
Eccentricity	0.8014	0.8014	0.8014
Orientation	-67.3618	-67.36...	-67.36...
Image	<34x17 logical>		
Perimeter	132.0416	132.04...	132.04...

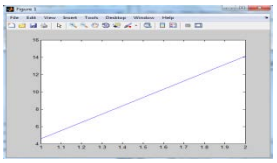
4.2.1 GLCM of character :

Now the following graph shows the GLCM . In GLCM the present texture of character is texture correlation as function of offset. The gray level co-occurrence matrix represents the values of 'contrast', 'correlation', 'energy', 'homogeneity' of each character which is stored in datafile.mat file. We took example of two characters there for, features of only two characters are shown here.

Table 4.2 Value of contrast of प

stats.Contrast <1x2 double>

1	2
4.5776	14.1278

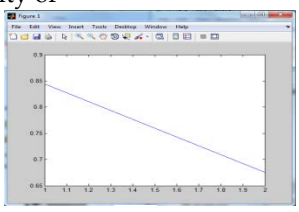


Graph 4.1 Graph of contrast of प

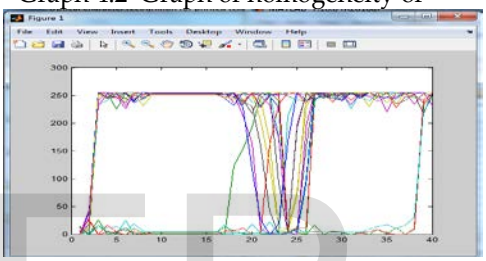
Table 4.3 value of homogeneity of प

stats.Homogeneity <1x2 double>

1	2
0.8442	0.6751



Graph 4.2 Graph of homogeneity of प



Graph 4.3 Extracted Feature of character प

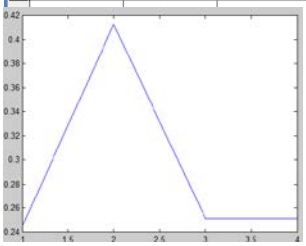
Output of NN :

Output of neural network can be represented by following graph. As well as recognized output is also shown in following figure. (a), (b).

Table 4.4 output of NN

outputs <1x4 double>

1	2	3	4
-0.0577	0.2819	-0.3532	-0.3532



Graph 4.4 Output of NN

Fig. 4.5 Recognized character

5. CONCLUSION

Character recognition is one of the difficult task , because verity font size and font faces are present now a days. So it's a try to achieve maximum accuracy and reduce time duration required in recognition of character. The proposed method hopefully can inspire a new thinking and new way to tackle the face recognition problem. Extensive training and testing experiments are carried out in order to demonstrate the effectiveness of the proposed method for devnagri character recognition. The performance of the proposed method in terms of recognition accuracy is obtained. Features used in character recognition i.e. GLCM, Colour dominant, Histogram, Affine moment invariant, gives good results compare to others, and for recognition process ANFIS (Artificial neuron fizzy interference system) tech. is used which gives the best result compare to other technique.

Talking about single characters i.e.  it gives 100% accuracy. But when talking about all devnagri character it shows mistake in recognising some character. Recognition rate of all devnagari character is near about 95%.

6. REFERANCES

[1] Kailash S. Sharma,A. R. Karwankar, Dr. A.S.Bhalchandra," Devnagari Character Recognition Using Self Organizing Maps" ICCCCT'10

[2] http://www.heatonresearch.com/articles/series/1

[3] R.M.K. Sinha, and Veena Bansal, "On Automating trainer for construction of prototypes for Devnagari text recognition", Technical report TRCS-95-232, IIT Kanpur, India 1995.

[4] http://en.wikipedia.org/wiki/Handwriting_recognition

[5] R.M.K. Sinha, and Veena Bansal, "On Devanagari documentation processing", IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada 1995.

[6] Veena Bansal, R.M.K. Sinha, "On How to Describe Shapes of Devanagari Characters and Use Them for Recognition," icdar, pp.410, Fifth International Conference on Document Analysis and Recognition (ICDAR'99), 1999

[7] Veena Bansal & R.M.K. Sinha, "Segmentation of Touching Characters in Devanagari", http://www.iitk.ac.in/ime/veena/PAPERS/stwo.pdf

[8] M.Babu Rao, Dr.B.Prabhakara Rao, Dr.A.Govardhan, "Content Based Image Retrieval using Dominant Color and Texture features" (IJCSIS) International Journal of Computer Science and Information Security,Vol. 9, No. 2, February 2011

[9] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal "Offline Recognition of Devanagari Script: A Survey", ieee transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 6, november 2011.

[10] Mohanad Alata — Mohammad Al-Shabi," TEXT DETECTION AND CHARACTER RECOGNITION USING FUZZY IMAGE PROCESSING", Journal of electrical engineering, VOL. 57, NO. 5, 2006, 258–267

[11] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", Second Edition, John Wiley & Sons Inc, New York, 2006, pp. 576- 579, 582.

[12] H.Ma and D. Doermann, "Adaptive Hindi OCR using generalized Hausdorff

image comparison," *ACM Trans. Asian Lang. Inf. Process.*, vol. 2, no. 3, pp. 193–218, 2003.

[13] U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 444–457, Mar. 2009.

[14] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," *Pattern Recognit.*, vol. 37, pp. 1887–1899, 2004.