International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 2, May 2013, 93-100

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

DEVANAGARI OCR USING KNN AND MOMENT

PRASHANT S. KOLHE & S. G. SHINDE

TPCT's College of Engineering Osmanabad, Maharashtra, India

ABSTRACT

We have built a comprehensive machine for Marathi Number recognition system. The number and vowels are recognized using an classifier which is KNN along with the feature set which is extracted as a moments features.

KEYWORDS: Devanagri Script, OCR System, KNN

INTRODUCTION

OCR work on printed Devnagari script started in early 1970s. Among the earlier pieces of work, some of the efforts on Devanagri character recognition are due to Sinha [1,7,8] and Mahabala [1]. Sethi and Chatterjee [5] also have done some earlier studies on Devanagri script and presented a Devanagri hand-printed numeral recognition system based on binary decision tree classifier. They [6] also used a similar technique for constrained hand-printed Devanagri character recognition. They did not show results of scanning on real document pages. The first complete OCR system development of printed Devanagri is perhaps due to Palit and Chaudhuri [4] as well as Pal and Chaudhuri [3]. For the purpose some standard techniques have been used and some new ones have been proposed by them. The method proposed by Pal and Chaudhuri gives about 96% accuracy. A survey for hand-written recognition of character is proposed [2]. A few of these work deals with handwritten characters of Devanagri. Because of the complexities involved with Devanagri script, already existing methods can not be applied directly with this script report on handwritten Devanagri characters was published in 1977 [9] and not much research work is done after that. Some research work are available towards Devanagri numeral recognition [10-12] but to the best of our knowledge there are only two reports on Devanagri off-line handwritten character recognition [13,14] after the year 1977. An excellent survey of the area is given in [15]. Devanagari is the script for Hindi which is official language of India.

The OCR techniques can be broadly classified into two methods Feature Mapped Recognition and Image Mapped Recognition. In the Feature Mapped Recognition, the recognition task is accomplished by Extracting certain primitives or distinctive features. The individual characters are recognized based on a decision function that decides the presence and absence of different primitive components in the character. In the Image Mapped approach the identification and the extraction of features are implicit processes within the recognition process.

DEVANAGARI FEATURE EXTRACTION

We will now briefly review the few important works done towards feature extraction techniques used for devanagri. R.M.K. Sinha et. al. [1,7,8,17,18,19] have reported various feature extraction and recognition aspects of devanagari script.In his work N. Sharma et. al.[14] used 64 dimensional feature vector and the features are obtained from the directional chain code information of the contour points of the characters. S. Basavaraj Patil et.al. [20], R Bajaj et. al. [21] and s. kumar used neural network successfully. K. Jaynathi et.al[22] used structure analysis for feature extraction. U.pal et. al. [23] features used are obtained from the directional information of the contour points of the numerals. A Modified Quadratic Discriminant Function (MQDF) has been used for the recognition of the numerals. Devanagri

characters recognition based on segmentation using various operators and converting image into a set of characters having definite prerequisite relationship is reported in [24,25,26,27,28]. Padma et. al. [29] have proposed a method based on visual discriminating features to identify characters. Hanmandlu and Murthy [10] proposed a Fuzzy model based recognition of handwritten Hindi numerals.

For recognition of handwritten Devanagri numerals, Ramakrishnan et al. [30] used independent component analysis technique for feature extraction from numeral images. Ramteke et al [31] proposed an isolated Marathi handwritten numeral scheme based on invariant moments. They employed a Gaussian Distribution Function for classification. Bajaj et al [11] employed three different kinds of features namely, density features, moment features and descriptive component features for classification of Devanagri Numerals.

They proposed a multi-classifier connectionist architecture for increasing the recognition reliability. Kumar and Singh [13] proposed a Zernike moment feature based approach for Devanagri handwritten character recognition. They used an artificial neural network for classification. In an attempt to develop a bilingual handwritten numeral recognition system, Lehal and Bhatt [32] used a set of global and local features derived from the right and left projection profiles of the numeral images for recognition of handwritten numerals of Devanagri and Roman scripts.Sethi and Chatterjee [6] proposed a decision tree based approach for recognition of constrained hand printed Devnagari characters using primitive features. R.Kapoor et al.[33] extracted nodal features from Devanagri characters.

Bhattacharaya et al [34,35] proposed a Multi-Layer Perceptron (MLP) neural network based classification approach for the recognition of Devanagri handwritten numerals. Recently, significant contributions towards the improvement of recognition rates have been made by means of different combination strategies [36,37,38],and the use of ANN, support vector machines & HMM [12][39] [40] .

DATASET

We have created different dataset with different ISM software fonts.

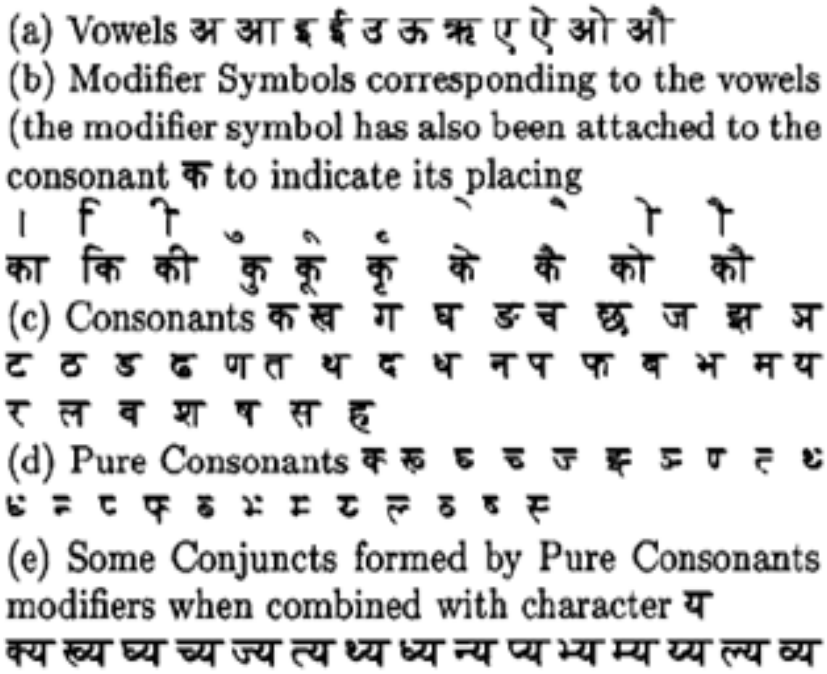

Figure 1: Devnagari Vowels, Consonants, Modifier, Conjuncts & Pure Consonants.[47]

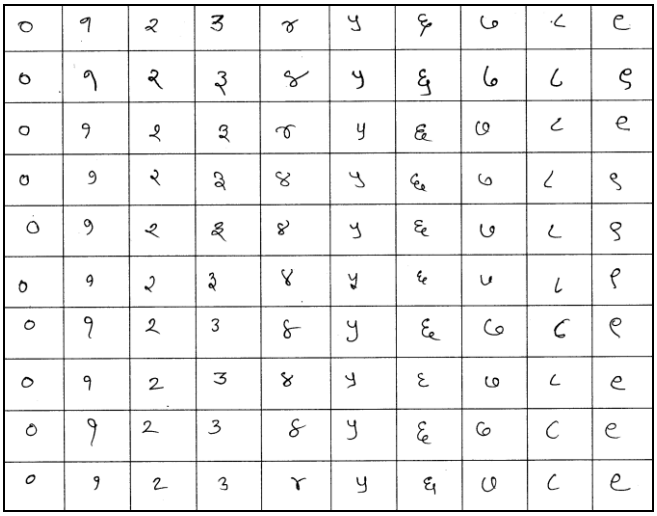

Figure 2: Devnagari Numbers

FEATURE EXTRACTION

A Moment Functions

Moment functions are defined on images as the weighted sums of the image intensity function. Moment functions of order (p+q) are generally defined as

$$\phi_{pq} = \int_x \int_y \psi_{pq}(x,y) f(x,y)\, dxdy,$$

Where $\psi_{pq}(x,y)$ is called the moment weighting kernel.

When applying moment functions to digital images it is often desirable to write them out using the following discrete notation:

$$\phi_{pq} = \sum_x \sum_y \psi_{pq}(x,y) f(x,y).$$

Some properties of the weighting kernel are passed onto the moments themselves, such as invariance features, ad orthogonality. Depending on the function chosen for the weighting kernel, the calculated moments can capture different aspects of the input image[43].

Zernike Moments

As Opposed to geometric moments, Zernike Moments are defined over the unit disk instead of the real plane and exhibit the orthogonality property.

Zernike polynomials are mainly used in optometrics, where they arise as the expension of a wavefront function in optical systems with circular pupils [5]. Zernike introduced a set of complex polynomials which form a complete orthogonal set over the interior of the unit circle, i.e., $x^2 + y^2 = 1$. Let the set of these polynomials be denoted by $\{V_{nm}(x,y)\}$. The form of these polynomials is:

$$V_{nm}(x,y) = V_{nm}(\rho\ \theta) = V_{nm}(\rho)\exp(jm\theta)$$

$R_{nm}(\rho)$ Radial Polynomial defined as

$$R_{nm}(\rho) = \sum_{s=0}^{n-|m|/2} (-1)^s \cdot \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \rho^{n-2a.}$$

Note that $R_{n.-m}(\rho) = R_{nm}(\rho).$

These polynomials are orthogonal and satisfy

$$\iint_{x^2+y^2\le 1} \left[V_{nm}(x,y)\right]^{*} V_{pq}(x,y)\, dx\, dy = \frac{\pi}{n+1}\delta_{np}\delta_{mq}$$

with $\delta ab = \begin{cases} 1 & a=b \\ 0 & otherwise \end{cases}$

Zernike moments are the projection of the image function onto these orthogonal basis functions. The zernike moment of order n with repetition m for a continuous image function f(x,y) that vanishes outside the unit circle is

$$A_{nm} = \frac{n+1}{\pi} \iint_{x^2+y^2\le 1} f(x,y)\, V_{nm}^{*}(\rho,\theta)\, dx\, dy$$

For a digital image, the integrals are replaced by summations to get.

$$A_{nm} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f(x,y)\, V_{nm}^{*}(\rho,\theta), \quad x^2 + y^2 \le 1.$$

To compute the zernike moments of a given image, the center of the image is taken as the origin and pixel coordinates are mapped to the range of unit circle, i.e. $x^2 + y^2 \le 1.$ Those pixels falling outside the unit circle are not used in the computation. Also note that $A_{nm}^{*}\, A_{n.-m}$. Therefore |*Anm*| can be used as a rotation invariant feature of the image function. Since *An,–m= Anm*, and therefore |*An,–m*|=|*Anm*|, we will use only |*Anm*| for features. Details can be found in [42]. We are Zernike movement of order two and movement4.for every dataset. The feature vector size is of 1x4,means 4 feature per character image are extracted.

CLASSIFICATION

Euclidian Distance-Based K-NN Classification

In *K*NN classification, training patterns are plotted in d-dimensional space, where d is the number of features present. These patterns are plotted according to their observed feature values and are labeled according to their known class. An unlabelled test pattern is plotted within the same space and is classified according to the most frequently occurring class among its *K*-most similar training patterns; its nearest neighbors. The most common similarity measure for *k*nn classification is the Euclidian distance metric, defined between feature vectors $\vec{x}$ *and* $\vec{y}$ as :

$$euc(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{f} (x_i - y_i)^2}$$

Where *f* represents the number of features. Smaller distance values represent greater similarity

RESULTS

Table 1: Result of Vowels

Sr.No.	Moment Order	Number of Moment	Recognition Rate%
1	4	7	65.44
2	5	10	68.66
3	6	14	69.06
4	7	18	71.63
5	8	23	73.98
6	9	28	75.65
7	10	34	80.55

Table 2: Result of Number

Sr.No.	Moment Order	Number of Moment	Recognition Rate%
1	4	7	68
2	5	10	72
3	6	14	73.44
4	7	18	75.89
5	8	23	77.90
6	9	28	80.45
7	10	34	81.55

Here we have extracted feature of dataset and classifier are used for classification.The results are given above .

REFERENCES

1. R.M.K. Sinha, H. Mahabala,,"Machine recognition of Devanagri script", IEEE Trans. System, Man Cybern. 9(1979) 435-441.
2. Plamondon, R. Srihari, S.N. ,Ecole Polytech.,Montreal, Que.; Online and Offline HandwritingRecognition : A comprehensive Survey,1EEE Transactions On Pattern Analysis And Machine Intelligence. VOL. 22, NO. 1. JANUARY 2000 63
3. U. Pal , B.B. Chaudhuri , "Printed Devanagri script OCR system", Vivek 10 (1997) 12-24.
4. S. Palit, B.B. Chaudhuri,,"A feature-based scheme for the machine recognition of printed Devanagri script", P.P. Das, B.N. Chatterjee (Eda.) Pattern Recognition, Image Processing and Computer Vision, Narosa Publishing House: New Delhi, India 1995, pp. 163-168.
5. I.K. Sethi, B. Chatterjee, "Machine recognition of constrained hand-printed Devanagri numerals", J. Inst.Electron. Telecom. Eng. 22 (1976) 532-535.
6. I.K. Sethi, B. Chatterjee,"Machine recognition of constrained hand-printed Devanagri characters", Pattern Recognition 9 (1977) 69-76
7. R.M..K. Sinha, "A syntactic pattern analysis system and its application to Devanagri script recognition", Ph.D. Thesis , Electrical Engineering Department, Indian Institute of Technology, India, 1973.
8. V. Bansal, R.M.K. Sinha, "Partitioning and searching dictionary for correction of optically read Devanagri characters strings", Proceedings of the Fifth International Conference on Document Analysis and Recognition , 1999, pp. 653-656.

9. S. Arora, D.Bhattacharya, M. Nasipuri, L.Malik, "A Novel Approach for Handwritten Devanagari Character Recognition" in IEEE –International Conference on Signal And Image Processing, Hubli, Karnataka, Dec 7-9, 2006.

10. M. Hanmandlu and O.V. Ramana Murthy, "Fuzzy Model Based Recognition of Handwritten Hindi Numerals", In Proc. Intl. Conf. on Cognition and Recognition, pp. 490-496, 2005.

11. R. Bajaj, L. Dey, and S. Chaudhury, "Devanagri numeral recognition by combining decision of multiple connectionist classifiers", Sadhana, Vol.27, pp.-59-72, 2002.

12. U. Bhattacharya, S. K .Parui, B. Shaw, K. Bhattacharya, "Neural combination of ANN and HMM for handwritten Devanagri Numeral Recognition", In Proc. 10th IWFHR, pp.613-618, 2006.

13. S. Kumar and C. Singh, "A Study of Zernike Moments and its use in Devanagri Handwritten Character Recognition", In Proc. Intl. Conf. on Cognition and Recognition, pp. 514-520, 2005.

14. N. Sharma, U. Pal, F. Kimura and S. Pal, "Recognition of Offline Handwritten Devanagri Characters using Quadratic Classifier", In Proc. Indian Conference on Computer Vision Graphics and Image Processing, pp- 805-816, 2006.

15. U. Pal and B.B. Chaudhuri, ''Indian script character recognition: A Survey'', Pattern ecognition, Vol. 37, pp. 1887-1899, 2004.

16. Petr Somol; Jana Novovičová; Pavel Pudil,''Notes on the Evolution of Feature Selection Methodology'',Kybernetika, Vol. 43 (2007), No. 5, 713–730

17. V. Bansal And R.M.K. Sinha,"On How To Describe Shapes of Devanagari Characters and Use Them for Recognition", Proc. 5 th Conf. Document Analysis and Recognition, Banglore, India, Sept. 1999,pp. 410-413.

18. S. S. Marwah, S. K. Mullick and R. M. K. Sinha , Recognition of Devnagri Character using hierachical binary decision trace classifier ", IEEE international conference on system, Man, Cybernete oct. 1994.

19. Veena Bansal and R.M.K. Sinha, "Segmentation of touching and fused Devanagari characters, Technical Report", TRCS-97-247, I.I.T. Kanpur, India, 1997.

20. S. Basavaraj Patil and N.V.Subbareddy, "Neural network based system for script identification in Indian documents," Sadhana, vol. 27, part-1, pp. 83-97, 2002.

21. R Bajaj and S Chaudhary. 'Devanagari Numeral Recognition using Multiple Neural Classifiers.' Indian Conference on Pattern Recognition, Image Processing and Computer Vision (ICPIC), December 1995.

22. K. Jaynathi, A.Suzuki, H. Kanai,Y. Kawazoe, M.Kimura, K. Kido, " Devanagari Character Recognition Using Structure Analysis",IEEE Trans,1989,pp 363-366.

23. U. Pal, S. Chanda, T. Wakabayashi and F. Kimura, "Accuracy Improvement of Devnagari Character Recognition Combining SVM and MQDF", In Proc. 11th ICFHR, pp.367-372, 2008.

24. P. Deshpande, L.Malik, S. Arora," Character Recognition with Histogram Band Analysis of Encoded String and Neural Network", Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, December 16-18, 2005 , pp354-359.

25. P.S.Deshpande, Latesh Malik, Sandhya Arora "Characterizing Hand Written Devnagari Charaters Using Regular Expressions" IEEE TENCON, Hong Kong November 14-17 2006

26. Latesh Malik ,P.S. Deshpande & Sandhya Bhagat ,Character recognition using relationship between connected segments and neural network. Issue 1, Volume 5 ,January 2006 ISSN 1109-2750

27. U. Garain, B.B. Chaudhuri," Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis", Proceedings of the 6th International Conference on Document Analysis and Recognition. (ICDAR '01).

28. B. Shaw, S. K. Parui and M. Shridhar, "Off-line Handwritten Devanagari Word Recognition: A Segmentation Based Approach", IEEE , 2008.

29. M.C.Padma and P. Nagabhushan," Script Identification and separation of text words of Kannada Hindi and English languages through discriminating features," Proc. of NCDAR-2003, pp. 252-260. 2003.

30. K.R. Ramakrishnan, S.H. Srinivasan and S. Bhagavathy, "The independent components of characters are 'Strokes', Proc. of the 5th ICDAR, 1999, pp. 414-417.

31. R.J. Ramteke, P.D.Borkar, S.C. Mehrotra, "Recognition of Marathi Handwritten Numerals: An Invariant Moments Approach", Intl.Conf. on Cognition and Recognition, pp. 482-489, 2005.

32. G.S. Lehal and Nivedan Bhatt, " A recognition system for Devnagri and English handwritten numerals", Advances in Multimodal Interfaces– ICMI 2001, T. Tan, Y. Shi and W. Gao (Editors), LNCS, Vol. 1948, 2000, pp. 442-449.

33. R. Kapoor, D. Bagai and T. S. Kamal, "Representation and Extraction of Nodal Features of DevNagri Letters", Proc. of ICVGIP, 2002.

34. U. Bhattacharya, T.K. Das, A. Datta, S.K. Parui and B. B. Chaudhuri, "A hybrid scheme for handprinted numeral recognition based on a self-organizing network and MLP classifiers", IJPRAI, Vol. 16(7), 2002, pp. 845-864.

35. U. Bhattacharya, B. B. Chaudhuri, R. Ghosh and M. Ghosh, "On Recognition of Handwritten Devnagari Numerals", In Proc. of the Workshop on Learning Algorithms for Pattern Recognition (in conjunction with the 18th Australian Joint Conference on Artificial Intelligence), Sydney, pp.1-7, 2005.

36. J. Cao, M. Ahmadi, and M. Shridhar, "Handwritten numeral recognition with multiple features and multistage classifiers," in IEEE International Symposium on Circuits and Systems, vol. 6, (London), pp. 323-326, May 30-June 2 1994.

37. K.V. Prema, N.V. Subba Reddy, "Two Tier Architecture For Unconstrained Handwritten Character Recognition",Spandan, Vol 27, Part 5, October 2002 , pp 585-594.

38. S. Kumar, "Recognition of Pre-Segmented Devanagari Handwritten Characters using Multiple Features and Neural Network Classifier", Ph.D. Thesis, Punjabi University, Patiala, Punjab, India, 2008, unpublished.

39. C. V. Jawahar, M. N. S. S. K. Pavan Kumar, and S. S. Ravi Kiran, "Recognition of Indian Language Characters using Support Vectors Machines," Technical Report TR-CVIT-22, International Institute of Information Technology,Hyderabad, 2002.

40. S. K. Parui and B. Shaw, "Off-line Devanagari Handwritten Word Recognition: An HMM based approach", Proc. PReMI-2007(Springer), LNCS-4815, pp. 528-535, Dec. 2007.

41. Eric W. Weisstein. Zernike polynomial. From MathWorld A Wolfram Web Resource. http://mathworld.wolfram.com/ZernikePolynomial.html.

42. **S**. Maitra, "Moment invariants," Proc. IEEE, vol. 67, no. 4, pp. 697-699, Apr. 1979.

43. Y. S. Abu-Mostafa and D. Psaltis, "Image normalization by complex moments," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-7, no. 1, pp. 46-55, Jan. 1985.

44. T.M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol. IT-13, pp. 21-27, Jan. 1967.