Spatial Features for Handwritten Kannada and English Character Recognition

B.V.Dhandra
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Karnataka, India

Mallikarjun Hangarge
Karnatak Arts, Science and Commerce College, Bidar
Karnataka, India

Gururaj Mukarambi
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Karnataka, India

ABSTRACT
This paper presents a handwritten Kannada and English Character recognition system based on spatial feature**s.** Directional spatial features viz stroke density, stroke length and the number of stokes are employed as potential features to characterize the handwritten Kannada numerals/vowels and English uppercase alphabets. KNN classifier is used to classify the characters based on these features with four fold cross validation. The proposed system achieves the recognition accuracy as 96.2%, 90.1% and 91.04% for handwritten Kannada numerals, vowels and English uppercase alphabets respectively.

General Terms
Pattern Recognition, Document Image Analysis

Keywords
OCR, Spatial Features, K-Nearest Neighbor.

1. INTRODUCTION
Automatic recognition of handwritten characters and numerals is an important and challenging problem due to its diversified applications like revenue records processing, form processing, reading of postal zip codes, passport number, employee codes, bank cheques and many more. Therefore, the problems of automatic recognition of handwritten characters/numerals are being attempted for decades. Methods like dynamic programming, hidden Markov modeling, neural network, expert system and combinations of these techniques have been proposed for solving these problems [1]. Extensive work has been carried out for recognition of characters and numerals in foreign languages like English, Chinese, Japanese, and Arabic. With respect to the Indian scripts, a major work can be found in [2, 3, and 24] on Tamil and Bengali scripts, where as the work on handwritten Kannada numerals/vowels recognition is in still infant stage. Recognition of handwritten Kannada characters is a complex task due to the unconstrained shapes, variation in writing style and different kinds of noise that break the strokes primitives in the characters or change in their topology. Designing handwritten character recognition system for multilingual documents is another challenging problem. Every Indian State has a three language formula for official communication as per the Indian constitution (i.e., English, Hindi and regional language). However, in almost all the states we find most of the documents in English and regional languages mixed together. Therefore, it is the need of the hour to design a bilingual/multilingual OCR system for automatic processing of the bilingual/multilingual documents and it is an extension of [19].

1.1 Literature Review
Recognition accuracy of the underlying image depends on the sensitivity of the selected features. Hence, number of feature extraction and selection methods can be found in the literature such as template matching, projection histogram, zoning, geometric moments, and invariant moments [4]. Most of the methods have employed fuzzy features [5, 6], templates and deformable templates [11, 12], structural and statistical features [7, 6]. Dinesh Acharya *et al* [8] have used the 10-segment string, water reservoir, horizontal/vertical strokes, and end point as the potential features for recognition and have reported the recognition accuracy of 90.50%. Draw back of this procedure is that, it is not free from thinning. U. Pal *et al* [14] have proposed zoning and directional chain code features and considered a feature vector of length 100 for handwritten numeral recognition, achieved reasonably high accuracy, but the time complexity of their algorithm is more. Work on English handwritten character recognition can be found in [20, 21, 22, and 23]. Recently, a piece of work on bi-lingual and tri-lingual numeral recognition of Indian scripts have been proposed in [25, 26]. From the literature survey, it is evident that still handwritten character recognition is a fascinating area of research to design a robust optical character recognition (OCR) system. This has motivated us to design a simple and robust handwritten Kannada and English OCR algorithm independently as a first step. In future we aim to work on designing of a bilingual OCR system for Kannada and English language, since most of the public and government documents exist in Karnataka state (one of the states of India) are in bilingual form(i.e., mixer of Kannada and English). For example, application forms, railway reservation slips, bank withdrawal slips and cheques have Kannada and English characters.

This paper is organized as follows: Section 2 contains the preprocessing of the images and data collection. Feature extraction procedure is discussed in Section 3. The experimental details and results obtained are presented in Section 4. Section 5 contains analysis of the affect of length of Structuring Element on proposed recognition system and conclusion is the subject matter of Section 6.

2. DATA SET AND PREPROCESING
The standardized database for Kannada handwritten characters is not available therefore, our own database is created along with English dataset. We have collected handwritten data from different professionals belonging to schools, colleges, business

organization etc. Each person has been asked to write 10 samples of each character of Kannada and English. These are scanned through a flat bed HP scanner at 300 DPI and using Otus's method they are binarized. Noise removal is performed by employing morphological area opening operation. A sample of Kannada handwritten alphabets and numerals are presented in Fig. 1, 2 and English alphabets are presented in Fig. 3.

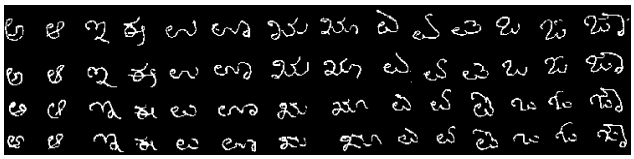


Figure 1: A sample handwritten Kannada vowels

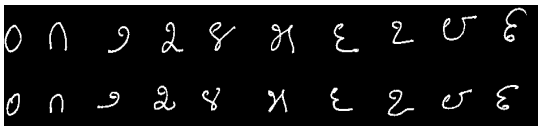


Figure 2: A sample data set of handwritten Kannada digits

Figure 3: A sample data set of handwritten uppercase English alphabets

3. FEATURE EXTRACION

The proposed feature extraction technique and its flow of execution is described below:

To perform basic morphological transformations, a line structuring element (SE) is used. The length of structuring element is computed as threshold value multiplied by the height of the input image. The threshold values are experimentally fixed as 20%, 30%, 40%, 50% and 60% of the height of the input image. The variation in the recognition rate of the proposed system for English Characters is observed with different thresholding values and it is graphically depicted in Fig. 7. Following operations are features obtained for classification of input character image.

1. Perform directional opening of a input character image I by

$$\gamma_{(\theta,\mu)}(I) = \varepsilon_{(\theta,\mu)}[\delta_{(\theta,\mu)}(I)], \qquad (1)$$

2. Compute the stroke length which is defined as the number of pixels in a stroke as the measure of its length [12]. Average stroke length is defined as the average of the length of the individual strokes obtained in an image I. Thus, average stroke length (ω_θ) is given by of an image.

$$\omega_\theta = \frac{1}{n}\sum_{i=1}^{n} length(stroke_i) \qquad (2)$$

for $\theta = 0^0, 45^0, 90^0$ and 135^0, where ω_θ is a feature vector of size 1x 4, where n is number of strokes in an image.

3. Obtain the average stroke density (ν),which is defined as the number of strokes per unit length (x-axis) of the input image I, which is computed by using the formula

$$\nu(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n_i}{width}\right) \qquad (3)$$

for $\theta = 0^0, 45^0, 90^0$ and 135^0, where the size of ν is 1x4. The sum of this row vector is used as average stroke density.

4. Compute the on-pixel ratio (η) using

$$\eta = \sum_{i=1}^{M}\sum_{j=1}^{N} f(i,j) \qquad (4)$$

In other words, on-pixel ratio is the ratio of on-pixels remaining after hole-filling of the input image f to its size.

5. Compute the aspect ratio (β) using

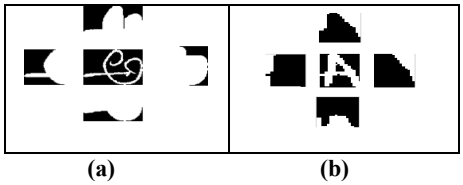
$$\beta = \frac{width(f)}{height(f)} \qquad (5)$$

That is, aspect ratio is the ratio of the width to the height of an input image [12].

6. **Eccentricity:** It is the length of major axis divided by the length of the minor axis of an input image f.

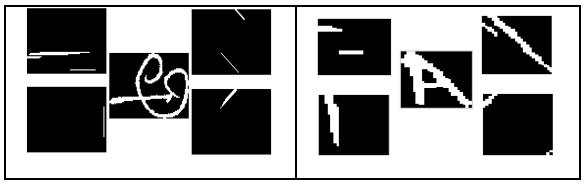
7. **Extent:** It is the proportion of the pixels in the bounding box that are also in the region. It can be computed as area (the number of pixels in a region) divided by the area of the bounding box.

8. **Directional profiles**: off pixels count of left, right, top and bottom of an input character is referred as directional profiles. A sample of Four directional profile images of Kannada and English characters are shown in Fig. 4. Further, four directional strokes extracted from a Kannada and English characters also shown in Fig. 5.



(a) **(b)**

Figure 4: Four Directional Profile of (a) Kannada Character (b) English Character



(a) **(b)**

Figure 5: Four Directional Strokes of (a) Kannada Character (b) English Character

Algorithm: Recognition of Handwritten Characters.

Input: Preprocessed handwritten character image.

Output: Recognized character.

Begin

1. Compute the four directional openings of a character.
2. Compute stroke length using (2).
3. Compute the average stroke density using (3).
4. Compute on pixels ratio using (4).
5. Compute the aspect ratio, eccentricity, extent and directional profiles using equations (5), (6), (7) and (8) respectively.
6. Store a feature vector for further analysis.
7. Use KNN algorithm with k=1and 5 to recognize vowels and numerals of Kannada script and English alphabets based on the feature vector obtained in step 6.

End

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Based on the KNN classifier 1000 numerals and 1400 vowels of Kannada character are classified. Further, proposed algorithm is experimented with 2600 English uppercase alphabets for their classification. The performance of KNN classifier is observed for different values of K i.e. K=1, 3, 5, 7 and it is graphically shown in Fig. 6. The classification results of Kannada numerals and vowels are presented in Table 1 and 2. The recognition rate of English uppercase alphabets is presented in Table 4. From Table 1, it is clear that the classification confusion is more between the numeral six and seven. It is due to their similarity structures. Table 2, reveals more error rate in classification of the vowels of having similar structural shape. To overcome this limitation, one may need to add some more dominating features that could discriminate the most confusing characters. In this direction, effort for identification of dominate features is in progress. Table 3 presents the comparison between proposed method and other methods found in the literature for handwritten Kannada numerals. English uppercase character recognition rate reported in [21, 22] are on an average 98% and 99.7% respectively. However, we have obtained 91.04% with KNN classifier. The performance evaluation of the proposed method with [21 22] is difficult, because the experimental conditions like number of features, classifiers, training dataset size, testing dataset size and datasets used by [21, 22] are different. Therefore, comparison of the performance of the proposed method with [21, 22] is just numeric. To realize the efficacy of the proposed features, we extended our experimentation with multiclass SVM classifier to recognize the English uppercase alphabets with different parameter values for C and Sigma, and obtained 95.32%. It indicates that further modifications and extensions to the proposed technique will give encouraging results.

Table 1 Percentage of numerals recognition accuracy using KNN classifier with k=1

Training samples =750, Test samples =250 and Number of features = 13			
Kannada Numeral	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy
೦	75	25	100.0
೧	75	25	98.9
೨	75	25	100.0
೩	75	25	98.1
೪	75	25	97.3
೫	75	25	95.4
೬	75	25	90.0
೭	75	25	84.0
೮	75	25	98.0
೯	75	25	100.0
Average Percentage of Recognition accuracy = 96.2			

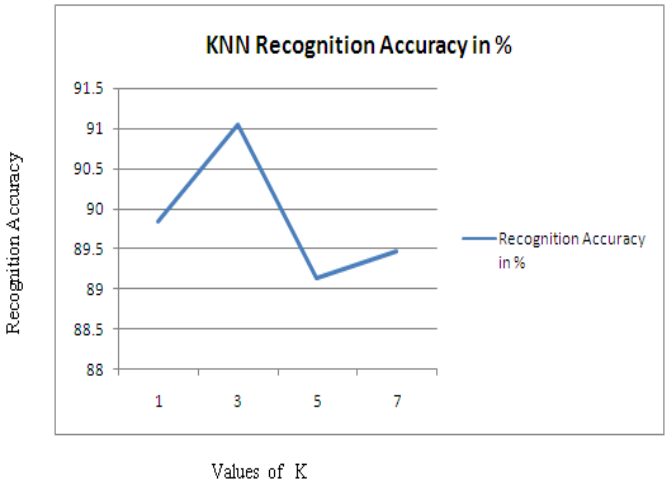


Figure 6: Graphical presentation of recognition accuracies with different K values

Table 2 Percentage of vowels recognition accuracy using KNN classifier with k=1

Training samples =1050, Test samples =350 and Number of features = 13			
Kannada Vowels	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy
ಌ	75	25	92.0

	75	25	98.9
	75	25	91.3
	75	25	93.1
	75	25	94.2
	75	25	92.3
	75	25	90.0
	75	25	84.0
	75	25	93.1
	75	25	92.1
	75	25	83.0
	75	25	87.3
	75	25	88.5
	75	25	90.8
Average Percentage of Recognition Accuracy = 90.1			

Table 3 Comparative results of handwritten Kannada numerals with other methods

Methods	Features and Classifier used	Data/ feature set	% of Acc.
[14]	Structural features k- means classifier	500	90.50
[15]	Image Fusion Nearest Neighbour	1000	91.20
[16]	Radon transform Nearest Neighbour	1000	91.20
[17]	Template matching, similarity-dissimilarity, binary distance transform, majority voting.	1000	91.00
[18]	Structural features Nearest Neighbour	2500	95.40
Proposed	Spatial features	1000	96.2

Table 4 Percentage of English Uppercase alphabets recognition accuracy using KNN classifier with k=5

Training samples = 1950, Test samples = 650 and Number of features= 13			
English Alphabet	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy
	75	25	83.67
	75	25	63.82
	75	25	84.21
	75	25	94.87
	75	25	100.00
	75	25	91.83
	75	25	97.87
	75	25	100.00
	75	25	100.00
	75	25	100.00
	75	25	83.33
	75	25	97.85
	75	25	96.22
	75	25	89.28
	75	25	84.61
	75	25	100.00
	75	25	88.88
	75	25	90.56
	75	25	94.44
	75	25	100.00
	75	25	100.00
	75	25	97.61
	75	25	75.00
	75	25	81.13
	75	25	80.43
	75	25	96.00
Average Percentage of Recognition Accuracy = 91.04			

5. STRUCTURING ELEMENT LENGTH V/S RECOGNITION ACCURACY

In mathematical morphology, structuring element type and size plays an important role in extracting the intended objects from an image. The variation in the length of the structuring element and its affect on the proposed recognition system is depicted in Fig. 7. It can be observed from Fig.7 that this algorithm is dependent on the length of the structuring element. However, in this paper an attempt is made to vary the length of the structuring element

automatically, when the input image changes. Further, different threshold values are chosen as mentioned in Section 3 for fixing the length of the structuring element and observed the recognition accuracies as shown in Fig. 7. The role of fixing and optimizing the length of the structuring element is a crucial job. But, selecting the type of the structuring element, determining its size and its optimization is application dependent. Hence no formula can be derived for determining the optimum size of structuring element. It can be noticed that when the threshold value is 30%, the recognition rate is high i.e., 91.04% with respect to English uppercase alphabets. However, recognition rate of Kannada vowels and numerals are 90.1% and 96.2% when the threshold value is 70%.

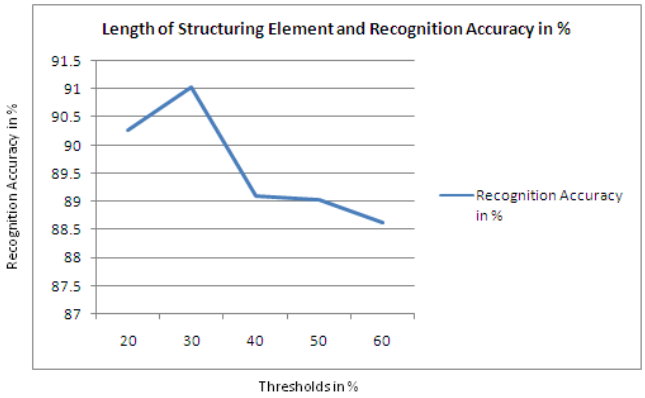


Figure 7: structuring element length v/s recognition accuracy plot

6. CONCLUSION

For independent recognition of handwritten Kannada numerals, vowels and English uppercase alphabets only 13 global spatial features are considered. The proposed directional spatial features shown quite encouraging performance with respect to handwritten Kannada and English characters. The results obtained are quite encouraging and comparable with the existing techniques. This algorithm is independent of image normalization, thinning, noise and slant of the characters. However, it is dependent on the size and type of the structuring element used for feature extraction. The aim of this work is to design the bilingual handwritten OCR system for Kannada and English languages. In future, we will extend it for bilingual recognition and it will be modifed to overcome the different complexities like writing style, quality of the paper, mood of the writer etc..

ACKNOWELDGEMENT

This work is supported by UGC, New Delhi under Major Research Project grant in Science and Technology, (F.No-F33 -64/2007 (SR) dated 28-02-2008). Authors are grateful to UGC for their financial support.

7. REFERENCES

[1] A.L.Koerich, R. Sabourin, C.Y.Suen, "Large off-line Handwritten Recognition: A survey", Pattern Analysis Application 6, 97-121, 2003.

[2] A. F. R. Rahman, R.Rahman, M.C.Fairhurst, "Recognition of handwritten Bengali Characters: A Novel Multistage Approach", Pattern Recognition, 35,997-1006, 2002.

[3] R. Chandrashekaran, M.Chandrasekaran, Gift Siromaney, "Computer Recognition of Tamil, Malayalam and Devanagari characters", Journal of IETE, Vol.30, No.6, 1984.

[4] Oivind Trier, Anil Jain, Torfiinn Taxt, "A feature extraction method for character recognition-A survey", pattern Recognition, vol 29, No 4, pp-641-662, 1996.

[5] Shamic Surel, P. K. Das, "Recognition of an Indian Scripts Using Multilayer Perceptrons and fuzzy Features" Proceedings Of 6th International Conference on Document Analysis and Recognition (ICDAR), Seattle, pp 1220-1224, 2001.

[6] P.Nagabhushan, S.A.Angadi, B.S.Anami, "A fuzzy statistical approach of Kannada Vowel Recognition based on Invariant Moments", Proceedings of NCDAR-2003, Mandy, Karnataka, India, pp275-285, 2003.

[7] L.Heutte, T.Paquest, J.V.Moreau, Y.Lecourtier, C.Oliver, "A structural/ statistical feature based vector for handwritten character recognition", Pattern Recognition, p.629-641, 1998.

[8] Dinesh Acharya U, N V Subba Reddy and Krishnamurthy, "Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster", IISN-2007, pp-125 - 129.

[9] B.V.Dhandra, V.S.Mallimath, Mallikargun Hangargi, "Multi-font Numeral recognition without Thinning based on Directional Density of pixels", Proceedings of first IEEE (ICDIM-2006) Bangalore, pp.157-160, Dec-2006.

[10] U Pal and P.P.Roy, "Multi-oriented and curved text lines extraction from Indian documents", IEEE Trans on system, Man and Cybernetics-Part B, vol.34, pp.1667-1684, 2004.

[11] J.D. Tubes, A note on binary template matching. Pattern Recognition, 22(4):359-365, 1989.

[12] Anil K.Jain, Douglass Zonker, "Representation and Recognition of handwritten Digits using Deformable Templates", IEEE, Pattern analysis and machine intelligence, vol.19, no-12, 1997.

[13] R.C.Gonzal, R.E.Woods, "Digital Image Processing", Pearson Education, 2002.

[14] N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", 9th International Conference on Information Technology (ICIT'06), ICIT, pp. 133-136.

[15] Rajput, G.G., Mallikarjun Hangarge, "Recognition of Isolated Handwritten Kannada Numeral based on Image fusion method", PREMI07, LNCS, Vol. 4815, Springer Kolkata, pp153-160, 2007.

[16] V. N. Manjunath Aradhya, G. Hemanth Kumar and S.Noushath, Robust Unconstrained Handwritten Digit Recognition Using Radon Transform, Proceedings of IEEE-ICSCN 2007, pp-626-629, (2007).

[17] B.V. Dhandra, R.G.Benne and Mallikargun Hangargi, "Isolated Handwritten Kannada Numeral recognition based on Template matching", IEEE-ACVIT -07, pp.1276-1282, Dec-2007.

[18] B.V. Dhandra, R.G.Benne and Mallikargun Hangargi, "Handwritten Kannada Numeral recognition based on structural features", IEEE International conference on

Computational Intelligence and Multimedia Application", ICCIMA-07, pp.157-160, Dec-2007.

[19] B.V.Dhandra, Mallikarjun Hangarge and Gururaj Mukarambi "Handwritten Kannada Numerals and Vowels Recognition Based on Spatial Features", National Seminar on Recent Trends in Image Processing and Pattern Recognition", RTIPPR-2010, pp.139-142, February 2010.

[20] Manish Mangal, Manu Pratap Singh, "Handwritten English vowels recognition using hybrid evolutionary feed-forward neural network, Malaysian Journal of Computer Science, Vol. 19(2), Pp.No.169-187,2006.

[21] L. Heutte a, T. Paquet a, J.V. Moreau b, Y. Lecourtier a, C. Olivier, A Structural/statistical feature based vector for handwritten Character Recognition, Pattern Recognition Letters 19, 1998 629–641.

[22] Subhangi D.C, P.S.Hiremath, "Handwritten English Character and Digit Recognition Using Multiclass SVM Classifiers and Using Structural Micro Features, International Journal of Recent Trends in Engineering, Vol. 2, No.2, November 2009, Pp. 193-195.

[23] Vellappa Ganapathy, Kok leong liew, "Handwritten Character Recognition Using Multi-scale Neural Network Training Techniques, World Academy of Science, Engineering and Technology, Pp. 32-37, 2008.

[24] S. Hewavitharana H. C. Fernando, "A Two Stage Classification Approach to Tamil Handwriting Recognition", Tamil Internet, California, USA, Pp.No.118-127, 2002.

[25] Benne R.G., Dhandra B.V. and Mallikarjun Hangarge,"Tri-scripts handwritten numeral recognition: a novel approach", Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, 2009, pp-47-51.

[26] G S Lehal and Nivedan Bhatt,"A Recognition System for Devnagri and English Handwritten Numerals" Springer Berlin vol-1948/2000,pp.no.442-449.