Zone Based Features for Handwritten and Printed Mixed Kannada Digits Recognition

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

Gururaj Mukarambi
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.

ABSTRACT

In the field of Optical Character Recognition (OCR), zoning is used to extract topological information from patterns. In this paper we propose Zone based features for recognition of the mixer of Handwritten and Printed Kannada Digits. A digit image is divided into 64 zones and pixel density is computed for each zone. This procedure is sequentially repeated for entire zone. Finally 64 features are extracted for classification and recognition. There could be some zone column/row having empty foreground pixels. Hence the feature value of such particular zone column/row in the feature vector is zero. The KNN and SVM classifiers are used to classify the mixed handwritten and printed Kannada digits. We have obtained 97.32% & 98.30% recognition rate for mixed handwritten and printed Kannada digits by using KNN and SVM classifiers respectively.

General Terms

Pattern Recognition, Document Image Analysis.

Keywords: OCR, Zone Features, KNN, SVM.

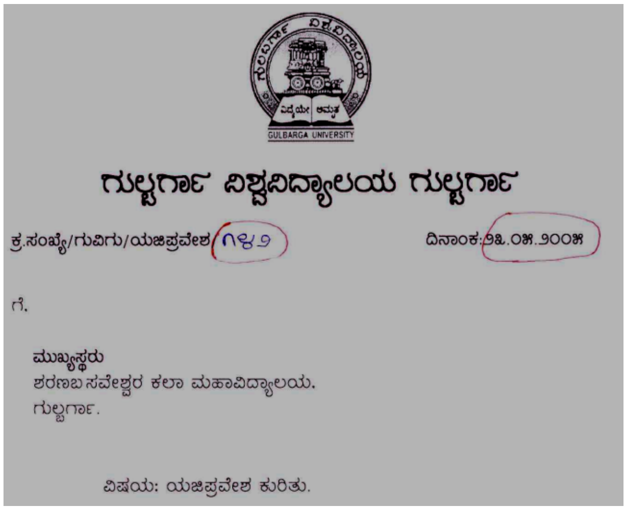
1. INTRODUCTION

Recent advances in Computer technology has made every organization to implement the automatic processing systems for its activities. For example, automatic recognition of vehicle numbers, postal zip codes for sorting the mails, ID numbers, processing of bank cheques etc. These are all the applications of digit recognition system. Hence, there is a need to develop an OCR to recognize the digits from a document containing Kannada handwritten and printed digits. In multilingual country like India, it is common that many documents consist of both handwritten and printed digits and characters in local languages. Mixture of handwritten and printed digits and characters in Indian context usually appears in a single document such as official letters. The most of the letter documents in Karnataka will have printed and handwritten Kannada digits. For example outward number in handwritten digits and date and other matters in printed digits. This addresses the need for development of single (Handwritten and Printed Mixed Kannada Digits) recognition system. In this direction many researchers have developed the numeral recognition systems by using various feature extraction methods such as geometrical features, global transformation, statistical features, topological and series expansion features like Fourier Transform, Wavelets, Hough Transform, Moments, etc. Extensive work has been carried out for recognition of characters and Digits in foreign languages like English, Chinese, Japanese, and Arabic.

With respect to the Indian scripts, a major work can be found in [1, 2] on Bengali and Tamil scripts, where as the work on handwritten Kannada Digits recognition is in still infant stage. Recognition of handwritten Kannada characters is complex task due to the unconstrained shapes, variation in writing style, etc.

U. Pal et al. [3] have proposed zoning and directional chain code features and considered a feature vector of length 100 for handwritten Kannada numeral recognition, achieved reasonably high accuracy, but the time complexity of their algorithm is more. Dinesh Acharya et al [4] have used the 10-segment string, water reservoir, horizontal/vertical strokes, and end points as potential features and have reported the recognition accuracy of 90.50%, which is relatively low accuracy and requires additional thinning algorithm. S.V. Rajashekararadhya et al. [5] have proposed zone based angle feature extraction system for handwritten numeral recognition of Kannada script and have reported the recognition accuracy of 96.05%. Anil.Jain et al. [6] a survey on feature extraction methods for character recognition is reviewed. Feature extraction methods includes Template matching, Deformable templates, Project Histograms, Zoning, Contour profiles, Geometric moment invariants, Zernike moments, Spline curve approximation, Fourier descriptors, Gabor and Gradient features. Dhandra et al. [7] have proposed spatial features and considered a feature vector of length 13 for handwritten Kannada numeral recognition and they have reported overall recognition accuracy of 96.2%. Hanmandlu et al. [8] have proposed a Fuzzy based approach for recognition of Hindi Multi-Font Digits. Ashwin et al. [9] have proposed a font and size independent OCR system for printed Kannada document and have formed the three basic Zones for the underlying character image. Each zone is divided into a number of circular tracks and sectors, on pixels in each angular region is used as a feature. Support vector machine was employed for the classification of characters and achieved an accuracy of 86.11%. A modified region decomposition method and optimal depth decision tree for the recognition of Kannada characters was used by Nagabhushan et al. [10]. Sanjeev Kunte et al. [11] have developed an OCR system for the recognition of basic characters of printed Kannada text, which works for different font size and font style. Each image was characterized by using Hu's invariant and Zernike moments. They have achieved the recognition accuracy as 96.8% with neural network classifier. Dhandra et al. [12] have proposed spatial features for Multi-font/Multi-size Kannada numerals and have reported an overall accuracy of 98.45%. Rajput et al. [13] have proposed chain code and fourier descriptors features for printed and handwritten numerals recognition and have reported the overall recognition accuracy of 97.76%. From the literature survey, it is evident that still handwritten character/numeral recognition is a fascinating area of research to design a single optical character recognition (OCR) system for bilingual, tri-lingual and multi-lingual numeral/character recognition. This has motivated us to design a recognition system for Kannada handwritten and printed mixed digits.

The sample Figure of the Kannada handwritten and printed mixed digits is shown in Fig 1.



GULBARGA UNIVERSITY

ಗುಲ್ಬರ್ಗಾ ವಿಶ್ವವಿದ್ಯಾಲಯ ಗುಲ್ಬರ್ಗಾ

ಕ್ರ.ಸಂಖ್ಯೆ/ಗುವಿಗು/ಯಜಿಪ್ರವೇಶ/೧೪೨ ದಿನಾಂಕ:೨೩.೦೫.೨೦೦೫

ಗೆ,

ಮುಖ್ಯಸ್ಥರು
ಶರಣಬಸವೇಶ್ವರ ಕಲಾ ಮಹಾವಿದ್ಯಾಲಯ.
ಗುಲ್ಬರ್ಗಾ.

ವಿಷಯ: ಯಜಿಪ್ರವೇಶ ಕುರಿತು.

Fig 1: Document of Mixture of handwritten and printed Kannada digits

This paper is organized into five sections. Section 2 contains the preprocessing of the images and data collection, Feature extraction procedure is discussed in Section 3, the experimental results and details are presented in Section 4. Conclusion is the subject matter of Section 5.

2. DATASET AND PREPROCESSING

The standard database for South Indian numeral script is neither available freely or commercially. Hence, we have created our own printed and handwritten digits database. Handwritten numeral data set is collected from different professionals belonging to Schools, Colleges, Doctors, Lawyers, etc. We are successful in collecting 1000 unconstrained handwritten Kannada digits and we are also successful in collecting 1100 printed Kannada digit samples by using Nudi and Bharha softwares. The collected data set containing multiple lines of isolated handwritten digits and printed digits are scanned through a flat bed HP scanner at 300 DPI and binarized using global threshold (i.e. Otsu's Method) and is stored in bmp file format. The scanned and segmented isolated digit images quite often contains noise that arises due to printer, scanner, print quality, etc. Noise removal is performed by employing morphological opening operation. Size normalization of an image is a crucial preprocessing stage in the development of robust digit recognizers. Normalization is the process of converting the random size image into standard size image. All isolated handwritten Kannada and printed Kannada digit images are normalized into a common height and width (i.e. 32 x 32 pixels) using bilinear technique. The normalized digit image is used for extracting the features. A sample data set of the Kannada handwritten and printed digits are shown in Fig.2 and Fig.3 respectively. The Normalization of handwritten Kannada digit is shown in Fig. 4.

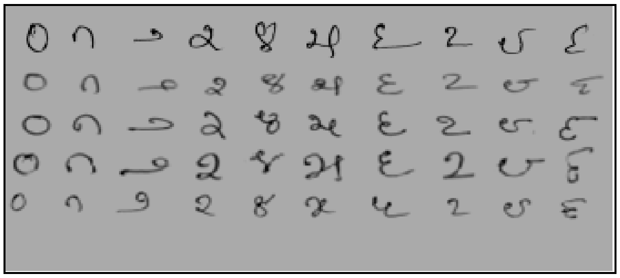


Fig 2: Sample dataset of handwritten Kannada digits

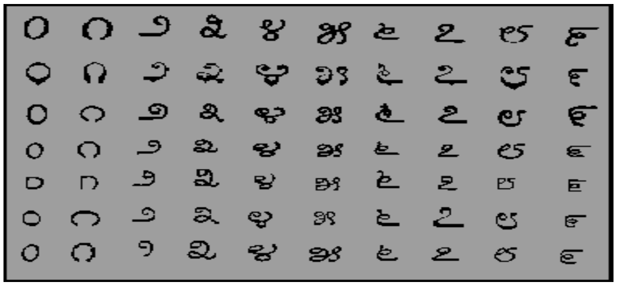
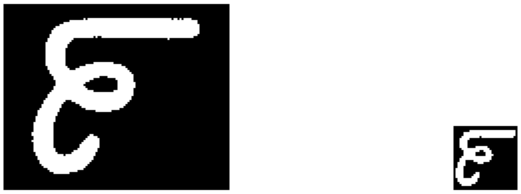


Fig 3: Sample dataset of printed Kannada digits



(a) **(b)**

Fig 4 : (a) Binary Image: Before Size normalization, (b) Image after Size normalization

3. FEATURE EXTRACTION

Feature extraction is a process of generating the relevant information from the preprocessed data for classification of underlying digits/characters. The preprocessed digit image is used as an input for feature extraction system. For extracting the potential features from the handwritten and printed Kannada digit image, the frame containing the preprocessed / normalized image is divided into non-overlapping zones of size 8 x 8. Thereby generating 64 zones. For each zone, the pixel density is computed and are used as a feature vector for classification. The optimum zone size is found to be 8 x 8. Hence, 64 features are used for experimentation.

```
Algorithm 3.1: Zone based pixel density feature extraction
               System
Input: Preprocessed handwritten/printed Kannada digit
       Image.
Output: Features for Classification and Recognition.
Begin

1. Divide the input digit image into 64 zones of size 8 x 8.
2. Compute the pixel density for each zone.
3. Repeat this procedure sequentially for all zones.
4. Finally, 64 features will be obtained for classification and
   recognition of digits by using k-nn and svm classifiers.
End
```

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Based on the KNN and SVM classifiers 1000 Kannada handwritten and 1100 Kannada printed Digits are classified. The experimental results are obtained by using 1050 training samples and 1050 testing samples for mixed Kannada handwritten and printed digits (see Table 3). The results obtained are encouraging for mixed handwritten and printed Kannada Digits. The Table 1 and Table 2 are presents the recognition accuracy of Kannada handwritten and printed Digits separately. The Table 3 shows the overall recognition accuracy for mixed handwritten and printed Kannada Digits. The Table 4 shows the average percentage of recognition accuracy.

Table 1. Percentage of handwritten Kannada Digits recognition accuracy using KNN (K=3) and SVM classifiers

Training samples =500, Test samples =500 and Number of features = 64				
Handwritten Kannada Digit	No. of sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
೦	50	50	100.00	97.62
೧	50	50	94.55	100.00
೨	50	50	100.00	100.00
೩	50	50	89.13	91.93
೪	50	50	100.00	100.00
೫	50	50	95.74	98.21
೬	50	50	91.67	91.30
೭	50	50	89.36	84.90
೮	50	50	96.42	100.00
೯	50	50	98.11	98.21
Average Percentage of Recognition accuracy			95.50	96.22

From the above table, the handwritten Kannada digits 3 and 7 have approximately same accuracy, because the digits 3 and 7 have similar in shape.

Table 2. Percentage of printed Kannada Digits recognition accuracy using KNN (K=3) and SVM Classifiers

Training samples =550, Test samples =550 and Number of features = 64				
Printed Kannada Digit	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
೦	55	55	100.00	100.00
೧	55	55	100.00	100.00
೨	55	55	100.00	100.00
೩	55	55	100.00	100.00
೪	55	55	100.00	100.00
೫	55	55	100.00	100.00
೬	55	55	100.00	100.00
೭	55	55	100.00	100.00
೮	55	55	100.00	100.00
೯	55	55	100.00	100.00
Average Percentage of Recognition accuracy			100.00	100.00

Table 3. Percentage of Mixed handwritten and printed Kannada Digits recognition accuracy using KNN (K=3) and SVM Classifiers

Training samples =1050, Test samples =1050 and Number of features = 64				
Mixed Handwritten and Printed Kannada Digit	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
೦	105	105	96.30	100.0000
೧	105	105	98.99	100.0000
೨	105	105	100.00	100.0000
೩	105	105	95.88	96.9388
೪	105	105	98.97	98.7500
೫	105	105	96.90	98.9583
೬	105	105	96.55	92.4528
೭	105	105	89.59	95.8763
೮	105	105	100.00	100.0000
೯	105	105	100.00	100.0000
Average Percentage of Recognition accuracy			97.32%	98.30%

Table 4. Average Percentage of Recognition accuracy

Kannada Digit	**Average Percentage of Recognition accuracy With KNN (K=3)**	**Average Percentage of Recognition accuracy With SVM**
Handwritten	95.50%	96.22%
Printed	100.00%	100.00%
Mixed(Both)	97.32%	98.30%

5. CONCLUSION

In this paper, a zone based features are proposed for recognition of mixed handwritten and printed Kannada Digits. The proposed method has shown the encouraging results for recognition of mixed handwritten and printed Kannada Digits by using KNN and SVM Classifiers. We have obtained recognition accuracy of 97.32% and 98.30% for mixed handwritten and printed Kannada Digits with KNN and SVM classifiers respectively. The novelty of this method is independent of thinning and slant of the Digits/characters with high recognition accuracy. The feature plan is to design a recognition system for mixed handwritten and printed digits for bilingual.

6. 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 and also to Dr. P. Nagabhushan, Dr. G. Hemanth Kumar, and Dr. D. S. Guru of University of Mysore for their discussion and critical comments.

7. REFERENCES

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

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

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

[4] Dinesh Acharya U, N VSubba Reddy and Krishnamurthy, “Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster”, IISN-2007, pp-125 -129.

[5] S.V.Rajashekararadhya and P.V.Vanaja Ranjan,”Neural network based handwritten numeral recognition of Kannada and Telugu scripts”, TENCON 2008.

[6] 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.

[7] B.V.Dhandra, Mallikarjun Hangarge, Gururaj Mukarambi, ”Spatial Features for Handwritten Kannada and English Character Recognition”, Special Issue on RTIPPR-10, International Journal of Computer Applications, pp.146-150, Aug-2010.

[8] M.Hanmandlu, Grover, V. K.Madasu, S.Vasikarla, "Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals", International Conference on Information Technology (ITNG'07), 2007.

[9] T V Ashwin, P S Sastry: “A Font and Size-Independent OCR System for Printed Kannada Documents using Support Vector Machine.Sadhana, vol. 27, pp.35-38, Feb. 2002.

[10] Nagbhushan P, Pai.Radhika “Modified region Decomposition Method and Optimal Depth Decomposition Tree in the Recognition of non–uniform Sized Characters experimentation with Kannada characters”. Pattern Recognition.Letters-20:1467-1475

[11] Kunte Sanjeev R, Sudhaker Samuel, “Hu’s invariant moments& Zernike moments approach for the recognition of basic symbols in Printed Kannada text”. Sadhana Vol .32, Part 5, October 2007, pp. 521-533.

[12] B.V.Dhandra, Mallikarjun Hangarge, Gururaj Mukarambi, "Spatial Features for Multi-Font/Multi-Size Kannada Numerals Recognition”, International Conference on Communication, Computation, Control and Nano Technology (ICN-2010),Bhalki, Bidar, Karnataka.

[13] G.G.Rajput, Rajeswari Horakeri, Sidramappa Chandrakant," Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM", International Journal on Computer Science and Engineering Vol. 02, No. 05, 2010,pp.1622-1626.

[14] A.Majumdar and B.B.Chaudhuri," A MLP Classifier for Both Printed and Handwritten Bangla Numeral Recognition", ICVGIP 2006, LNCS-4338, pp.796 – 804, 2006.

[15] Rafael C Gonzalez, Richard E Woods and Steven L Eddins, “Digital Image Processing using MATLAB”, Year 2008.