Kannada and English Numeral Recognition System

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 this Paper, zone based features are used for recognition of handwritten and printed Kannada and English numerals. The handwritten and printed Kannada and English numeral images are normalized into 32 x 32 dimensions. Then normalized images are divided into 64 zones and their pixel densities are used as feature vector. Thus, the dimension of feature vector is 64. The handwritten and printed Kannada and English numerals are tested for classifications on 4,000 sample images as an experiment and obtained an accuracy of 95.25% for KNN classifier and 97.05% for SVM classifier for mixed numeral inputs with 2-Fold cross validation for handwritten and printed Kannada and English numerals. A total of 40 classes have been reduced to 19 classes pertaining to handwritten and printed Kannada numerals and handwritten and printed English numerals to enable to increase the recognition accuracy. The novelty of the proposed algorithm is thinning free, independent of slant of the characters.

General Terms
Document Image Analysis

Keywords
OCR, Zone Features, KNN, SVM.

1. INTRODUCTION
The increasing advancement of electronic technology is promoting for automatic processing of the events in an organization in general and office automation in particular. The office automation is depending on object/optical character recognition (OCR). For optical character recognition document images are the basic inputs. These document images may be unilingual, bilingual or multilingual. Hence, a multilingual OCR system is required to process the multilingual document images. Multilingual documents will come across in India, since India is multilingual country. Automatic recognition of vehicle numbers, ID numbers, postal zip codes for sorting the mails and bank cheques etc. These are all the applications of multi-numeral recognition system. Further, the regional documents may contain numerals in regional language and in English (international language) language. Hence, there is a need to develop an OCR system that recognizes the handwritten and printed regional language numerals and English numerals from a document. The most of the documents in Karnataka will have printed and handwritten Kannada and English numerals. For example in a document outward number may be in handwritten numerals (It may be in Kannada/ English) and date and other matters in printed numerals. This addresses the need for development of single (Handwritten and Printed Bilingual Kannada and English Mixed Numerals) recognition system. In this direction many researchers have developed numeral recognition systems by using various feature extraction methods such as template matching, spatial, fourier descriptors and shape descriptors, Invariant moments, central moments and modified moments, structural / statistical, Zoning etc. Extensive work has been carried out for recognition of characters and Digits in foreign languages like English, Chinese, Japanese, and Arabic. For Indian scripts, a major work can be found in [1, 2] on Bengali and Tamil scripts, where as the work on handwritten Kannada numeral recognition is still in infant stage. But recognition of handwritten Kannada characters is a 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 amounting to 100 features for handwritten Kannada numeral recognition, and achieved reasonably high accuracy, but the time complexity of their algorithm is high due to large dimension feature set. Dinesh Acharya et al. [4] have used the 10-segment string, water reservoir, horizontal/vertical strokes, and end points as potential features for handwritten Kannada numerals and have reported the recognition accuracy of 90.50%, and further requires it additional thinning algorithm. Dhandra et al.[5] have proposed spatial features and considered a feature vector of length 13 for handwritten exclusively for Kannada numeral recognition and have reported overall recognition accuracy of 96.2%.

S.V. Rajashekararadhya et al. [6] have proposed zone centriod and image centriod based angle feature extraction system for isolated Kannada numerals recognition and reported 97.3% accuracy. Dhandra et al. [7] have proposed spatial features for Multi-font/Multi-size Kannada numerals and have reported an overall accuracy of 98.45%. Dhandra et al. [8] have proposed pixel density features for handwritten and printed Kannada mixed numerals recognition and have reported the overall recognition accuracy of 98.70%.

From the literature survey, it is evident that still handwritten numeral/ character recognition is carried out for single language and little work is carried out for bilingual, tri-lingual and multi-lingual numeral/character recognition. This has motivated us to design a single recognition system for mixture of handwritten and printed Kannada and English numerals.

The paper is organized as follows: Section 1 contains introduction part. Section 2 is about Bilingual approach. Section 3 contains the preprocessing and data collection details. Feature extraction algorithm is described in Section 4. The experimental details and results obtained are discussed in Section 5. Section 6 contains the conclusion part.

The sample document of Kannada handwritten and printed mixed numerals and English handwritten and printed mixed numerals are shown in Fig1and Fig2.

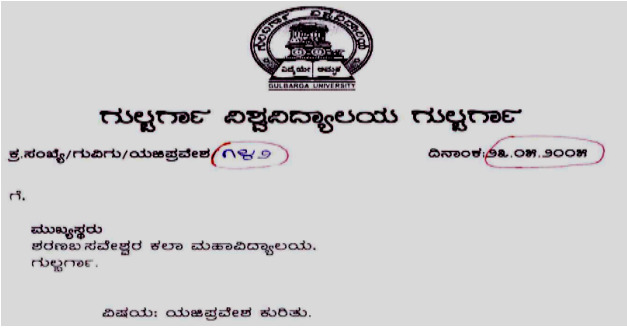

Figure 1: Sample document of Kannada handwritten and printed mixed numerals

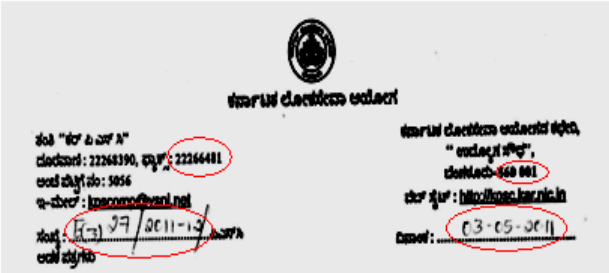

Figure 2 Sample document of English handwritten and printed mixed numerals

2. BILINGUAL APPROACH

Recognition of bilingual documents can be classified in two ways (1) Through script identification (2) Bilingual approach, in this approach; the OCR to be employed for the recognition of the bilingual documents (Kannada / English can be activated based on the script recognition of the input word/character. This approach reduces the search space in the database and allows for the Kannada and English characters recognition to be handled independently from each other.

In bilingual approach, characters are handled in the same manner, irrespective of the script they belong to. In any classification problem, the feature dimension is very much dependent on the number of classes. In the proposed work, the total number of classes to be classified is 40 (handwritten and printed Kannada script numerals -20 and handwritten and printed English script numerals - 20). However, as the number of classes increases, it is prudent to divide the classification problem. Hence, the classification problem of Kannada/English bilingual numeral recognition is reduced to 19 classes based on the observation that handwritten and printed Kannada numerals have the similar shape and similarly for handwritten and printed English numerals.

3. PREPROCESSING / DATA COLLECTION

The standard database for South Indian numeral script is neither available freely nor commercially. Hence, handwritten Kannada and English numerals database has been created. Handwritten Kannada and English numeral data set of size 2000 collected from different professionals belonging to Schools, Colleges, Doctors, Lawyers, etc. A total of 2000 printed Kannada and English numeral samples are created by using Nudi and Bharha softwares. The collected data set containing multiple lines of handwritten numerals and printed numerals 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 numeral image quite often contains noise that arises due to scanner, printer, print quality, etc. Noise removal is performed by employing morphological opening operations. The binary image is normalized to a resized size (32 x 32) dimension. All handwritten and printed Kannada and English numeral images are normalized into a common height and width (i.e. 32 x 32 pixels) using bilinear technique. A sample data set of the Kannada handwritten and printed numerals are shown in Fig.3 and Fig.4 respectively. The sample data set of handwritten and printed English numerals are shown in Fig 5 and Fig 6 respectively.

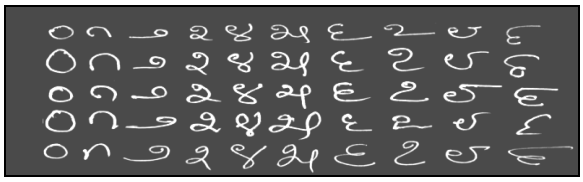

Figure 3 Sample data set of handwritten Kannada numerals

Figure 4 Sample data set of Printed Kannada numerals with multi font and multi style

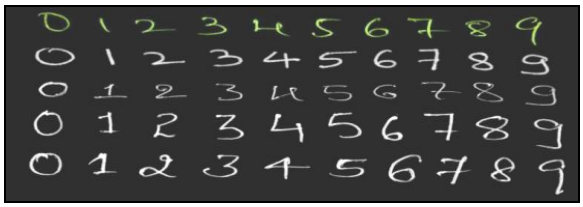

Figure 5 Sample data set of handwritten English numerals

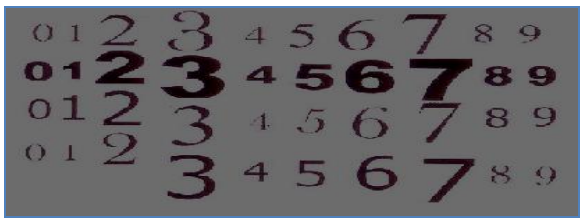

Figure 6 Sample data set of printed English numerals with Multifont and Multisize

4. FEATURE EXTRACION

The feature extraction method is based on distinguishing characteristics of an image. Here we have used Zone based feature extraction method for handwritten and printed Kannada and English mixed numerals recognition. The normalized images i.e.

(32 x 32) are divided into 64 non overlapping zones. For each zone pixel density is calculated and considered as a feature vector of size 64 for classification. There can be some empty zones (rows/columns), for such zones zero values are assigned. The zones 8 x 8 was used to generate 64 features and then these features are used for classification. The large zone size has failed to capture the essential local information of a numeral image. Hence zone size 8 x 8 is considered as optimum size for experimentation.

Feature extraction algorithm is the following:

Algorithm: Handwritten and Printed Kannada and English Mixed Numerals Recognition Based on Zone Features.
Input: Preprocessed Mixed numeral Images.
Output: Recognition of numerals.
Start

1. The normalized image is divided into 64 non overlapping zones of size 8 x 8.
2. Pixel density is calculated for each zone.
3. Generated 64 features are used for classification.
4. The KNN and SVM classifiers are used to recognize the numerals.

Stop

The sample image is divided into 64 zones is shown in Figure 7.

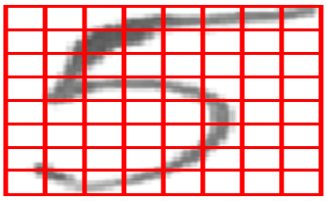

Figure 7 sample image of zone size 8 x 8

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

A Total of 4000 mixed handwritten and printed bilingual Kannada and English numeral images are classified using KNN and SVM classifiers. The recognition accuracy of 95.25%, 97.05% was achieved and the results are encouraging for handwritten and printed bilingual numeral images. The Table 1 and Table 2 presents the recognition accuracy of handwritten and printed Kannada numerals for input of Kannada numerals. The Table 3 and Table 4 show the recognition accuracy for independent input of handwritten and printed English numerals. The Table 5 and 6 shows the recognition accuracy for mixed handwritten and printed Kannada numerals and English numerals respectively. The Table 7 presents the recognition accuracy of the handwritten and printed bilingual mixed numerals with KNN and SVM classifiers respectively.

Table 1. Average Percentage of Recognition Accuracy for Handwritten Kannada Numerals input with KNN (K=3) and SVM Classifiers

Training samples =500, Test samples =500 Number of features = 64				
Handwritten Kannada Numeral	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

Table 2. Average Percentage of Recognition Accuracy for Printed Kannada Numerals input with KNN (K=1) and SVM Classifiers

Training samples =500, Test samples =500 Number of features = 64				
Printed Kannada Numeral	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	100.00
೧	50	50	100.00	100.00
೨	50	50	100.00	100.00
೩	50	50	100.00	100.00
೪	50	50	100.00	100.00
೫	50	50	100.00	100.00
೬	50	50	100.00	100.00
೭	50	50	100.00	100.00
೮	50	50	100.00	100.00
೯	50	50	100.00	100.00
Average Percentage of Recognition accuracy			100.00	100.00

Table 3. Average Percentage of Recognition Accuracy for Handwritten English Numerals input with KNN (K=1) and SVM Classifiers

Training samples =500, Test samples =500 Number of features = 64				
Handwritten English numerals	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
0	50	50	100.00	100.00
1	50	50	96.00	98.00
2	50	50	99.00	99.00
3	50	50	98.00	99.00
4	50	50	95.00	99.00
5	50	50	98.00	97.00
6	50	50	100.00	100.00
7	50	50	97.00	95.00
8	50	50	90.00	97.00
9	50	50	98.00	95.00
Average Percentage of Recognition accuracy			97.10	97.90

Table 4. Average Percentage of Recognition Accuracy for Printed English Numerals input with KNN (K=1) and SVM Classifiers

Training samples =500, Test samples =500 Number of features = 64				
Printed English Numeral	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
0	50	50	100.00	100.00
1	50	50	100.00	100.00
2	50	50	100.00	100.00
3	50	50	100.00	100.00
4	50	50	100.00	100.00
5	50	50	100.00	100.00
6	50	50	100.00	100.00
7	50	50	100.00	100.00
8	50	50	100.00	100.00
9	50	50	100.00	100.00
Average Percentage of Recognition accuracy			100.00	100.00

Table 5. Average Percentage of Recognition Accuracy for Handwritten and Printed Kannada Numerals mixed input with KNN (K=1) and SVM Classifiers.

Training samples =1000, Test samples =1000 Number of features = 64				
Handwritten and Printed Kannada Mixed Numerals input	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
	100	100	97.00	100.00
	100	100	99.00	99.00
	100	100	100.00	100.00
	100	100	95.50	97.50
	100	100	99.00	99.00
	100	100	96.50	99.00
	100	100	94.50	93.50
	100	100	85.00	94.00
	100	100	97.50	100.00
	100	100	99.50	99.50
Average Percentage of Recognition accuracy			96.35	98.15

Table 6. Average Percentage of Recognition Accuracy for Mixed Handwritten and Printed English numerals input with KNN (K=1) and SVM Classifiers

Training samples =1000, Test samples =1000 Number of features = 64				
Handwritten and Printed English Mixed Numerals	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
0	100	100	100.00	100.00
1	100	100	99.00	99.50
2	100	100	99.00	99.50
3	100	100	99.50	100.00
4	100	100	97.50	99.50
5	100	100	98.00	99.00
6	100	100	100.00	100.00
7	100	100	98.00	98.00
8	100	100	96.50	98.00
9	100	100	98.00	99.50
Average Percentage of Recognition accuracy			98.55	99.30

Table 7. Average Percentage of Recognition Accuracy for Mixed Handwritten and printed Kannada and English numerals input with KNN (K=3) and SVM Classifiers

Training samples =2000, Test samples =2000 Number of features = 64				
Handwritten and Printed Mixed Kannada and English numerals	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy With KNN	Percentage of Recognition Accuracy With SVM
	200	200	99.75	100.00
	100	100	96.50	99.50
	100	100	100.00	96.00
	100	100	89.00	100.00
	100	100	93.50	94.00
	100	100	95.50	96.50
	100	100	94.00	99.50
	100	100	83.00	97.50
	100	100	99.00	96.50
	100	100	94.50	98.00
1	100	100	99.00	98.50
2	100	100	94.00	100.00
3	100	100	99.00	95.00
4	100	100	93.50	95.50
5	100	100	94.00	98.00
6	100	100	100.00	93.50
7	100	100	97.00	88.00
8	100	100	92.50	100.00
9	100	100	96.00	98.00
Average Percentage of Recognition accuracy			95.25	97.05

From the above Table 7, it is clearly shown that the Kannada and English zero numerals are mixed into one class because the zero numeral is in similar shape. We have formed 9 classes for mixed handwritten and printed Kannada numerals. Then handwritten and printed English numerals are mixed together into 19 classes by considering zero numeral as one class for both languages. Here an attempted is made for bilingual recognition of handwritten Kannada and English numerals as well as mixed printed Kannada and English numerals.

The following Table 8 presents the average percentage of Recognition accuracy for handwritten Kannada, printed Kannada, handwritten English, printed English, mixed handwritten and printed Kannada, mixed handwritten and printed English and mixed Kannada and English numerals and Figure 8 shows the scatter plot of average percentage of recognition accuracy (in %) with respect to KNN and SVM classifiers.

Table 8 Average Percentage of Recognition accuracy

Numerals	Average Percentage of Recognition Accuracy in (%)	
	KNN	SVM
Handwritten Kannada	95.50	96.22
Printed Kannada	100.00	100.00
Handwritten English	97.10	97.90
Printed English	100.00	100.00
Mixed handwritten and printed Kannada	96.35	98.15
Mixed handwritten and printed English	98.55	99.30
Mixed handwritten and printed Kannada and English	95.25	97.05

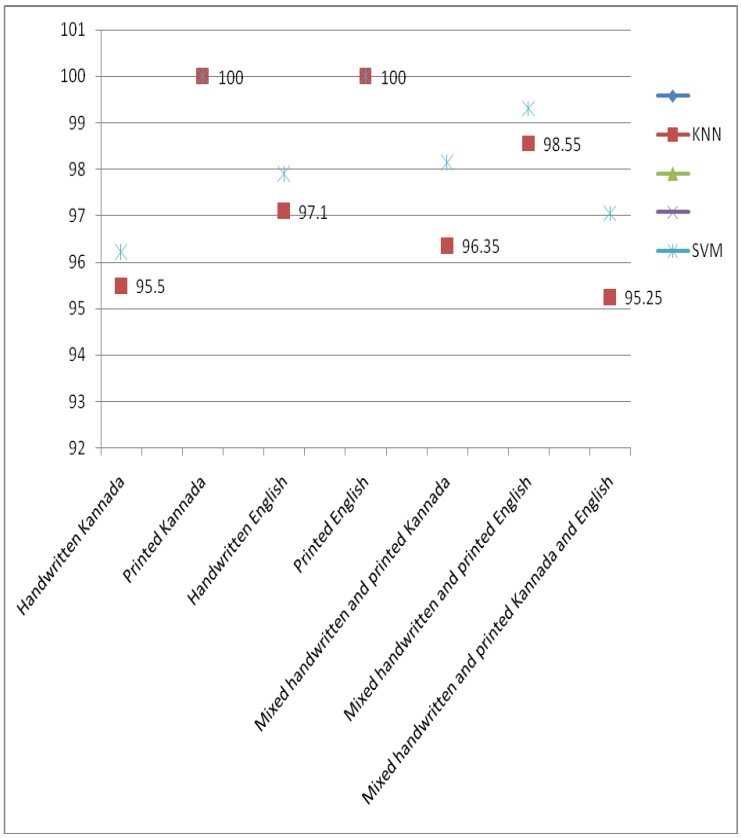

Figure 8 Scatter plot of Numeral Recognition

6. CONCLUSION

In this paper, a zone based features are proposed for recognition of mixed handwritten and printed Kannada and English numerals. The proposed method has shown the encouraging results for recognition of mixed numerals for Kannada and English. A recognition accuracy of 95.50%, 96.22%, 100%, 100%, 97.10%, 97.90%, 100% and 100% are achieved for handwritten

Kannada, Printed Kannada, handwritten English and printed English numerals independently by using KNN and SVM classifiers with 2-Fold cross validation. Further, recognition accuracy of 95.25% and 97.05% are obtained for mixed handwritten and printed Kannada and English numerals with 2-Fold cross validation by using KNN and SVM classifiers respectively. The novelty of this method is independent of thinning and slant of the numerals/characters with high recognition accuracy.

7. 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.

8. REFERENCES

[1] A.F.R.Rahman, M.C.Fairhurst, "Recognition of handwritten Bengali Characters: A Novel Multistage Approach", Pattern Recognition, pp. 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), pp. 133-136, 2006.

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

[5] 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.

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

[7] 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, India.

[8] B. V. Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge," Zone Based Features for Handwritten and Printed Mixed Kannada Digits Recognition", International Conference on VLSI, Communication & Instrumentation (ICVCI) 2011 Proceedings published by International Journal of Computer Application (IJCA), 2011.

[9] Basavaraj Patil, "Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents", International Journal of Computer Applications (0975 – 8887) Volume 13– No.8, pp. No 6-14, Jan-2011.

[10] B.V.Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge, "Handwritten Kannada Vowels and English Character Recognition System", International Conference on Computer Science and Information System (CSIT), Bangalore, 2011.

[11] B.V. Dhandra, R.G.Benne and Mallikarjun Hangargi, "Script Independent Handwritten Numeral Recognition with structural features", ICISP-2009, pp 431-434, Mysore.

[12] B. B. Chaudhuri and U. Pal. An OCR system to read two Indian language scripts: Bangla and Devanagari (Hindi). In Proceedings of ICDAR, pages 1011–1015, 1997.