I J M I

International Journal of Machine Intelligence
ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 4, 2011, pp-259-262
Available online at http://www.bioinfo.in/contents.php?id=31

RECOGNITION SYSTEM FOR HANDWRITTEN AND PRINTED KANNADA NUMERALS AND VOWELS

GURURAJ MUKARAMBI[1*], DHANDRA B.V.[1], MALLIKARJUN HANGARGE[2]

[1]Department of P.G. Studies and Research in Computer Science Gulbarga University, Gulbarga, Karnataka.

[2]Department of Computer Science, Karnatak Arts, Science and Commerce College Bidar, Karnataka.

*Corresponding Author: Email- gmukarambi@gmail.com

Received: November 06, 2011; Accepted: December 09, 2011

Abstract- In this paper, a recognition system for handwritten and printed Kannada numerals and vowels is proposed based on zone features. The Kannada numerals and vowels are circular in nature; therefore the pixel density feature is potential features for handwritten and printed Kannada numerals and vowels. The preprocessed image is divided into 64 zones. The pixel density is computed between each zone. The dimension of feature vector is 64. For classification, SVM classifier with five fold cross validation test is used to obtain average percentage of recognition accuracy of 97.40% and 95.90% for mixture of handwritten and printed Kannada numerals and vowels respectively. The total number of classes is 48, but in this experiment 24 classes are reduced due to mixture of handwritten and printed Kannada numerals and vowels. The novelty of the proposed algorithm is free from thinning and slant of the numerals and vowels**.**

Key words – Document Image Analysis, Zone, K-Fold, SVM.

Introduction

The advancement of computer technology encouraged, every organization to implement the automatic processing of their organization activities. E.g. Postal zip codes for sorting the mails, automatic reading of the vehicle registration numbers, ID numbers, processing of bank cheques etc. are the application areas of numeral recognition system. Hence, there is a need to develop an OCR to recognize the numerals and vowels from a document containing handwritten and printed script. In multilingual country like India, it is common that many documents consist of both handwritten and printed numerals and characters in local languages. Mixture of handwritten and printed numerals 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 numerals and characters. For example outward number in handwritten numerals and date and other matters in printed numerals and characters. This address the need for development of single (Handwritten and Printed Mixed Kannada numerals and vowels) recognition system. In this direction, many feature extraction methods are proposed in the literature such as spatial, structural, statistical and topological features. Several feature extraction methods for recognition of English, Chinese, Arabic and Latin script are reviewed in [1, 2, 3, and 4]. Only little work on handwritten and printed can be noticed on Indian scripts. The following is a brief account of the work carried out in the literature is presented.

Dhandra et al. [5] have proposed spatial features for handwritten Kannada and English Character Recognition, and have achieved the recognition accuracy of 90.01% for handwritten Kannada vowels and 91.04% for handwritten English uppercase alphabets. Leena R et al. [6] have proposed a Gabor filter features for handwritten Kannada Kagunita recognition and reported the recognition accuracy of 86% for vowels and 65% for consonants. Sangame et al. [7] have proposed zone based invariant moment features for handwritten Kannada vowel recognition and reported the average recognition accuracy of 85.53%. Aradhya et al. [8] have proposed fourier transform and principal component analysis techniques for handwritten Kannada character recognition and achieved the recognition accuracy of 68.89% for combined vowels and consonants. Dhandra et al. [9] have proposed zone based pixel density features for handwritten Kannada vowels and English uppercase alphabets and achieved the average recognition accuracy of 97.03% for combined vowels and uppercase alphabets. Acharya et al. [10] have proposed multi-level classifiers for recognition of handwritten Kannada numerals and reported the recognition accuracy of 98%. Dhandra et al. [11] have proposed zone based features for handwritten and printed Kannada digits recognition and achieved recognition accuracy of 97.32%. From the literature survey, it is clear that, still handwritten and printed character recognition is an active area of research to develop multilingual OCR System for the Indian languages. Hence in the proposed study an attempt is made to use the zone based features for the recognition of mixture of handwritten and printed Kannada numerals and vowels.

In this paper, Section1 contains the introduction and Section2 describes data collection and preprocessing details. Section3 deals with the feature extraction procedure. Experimental results and discussion are the subject matter of Section4 and Section5 contains conclusion.

Data Collection and Pre-Processing

To validate and verify the results of the proposed algorithm the standard database for handwritten and printed Kannada script is not available; hence we have created own Handwritten Kannada datasets from different professionals belonging to Schools, Colleges, Government Departments, Lawyers etc. The printed datasets are created by using Nudi and Baraha software's with five different font and styles. The printed dataset contains multi-font and multi-size numerals and vowels. The collected handwritten and printed Kannada document contains a multiple lines of Kannada scripts. These documents are scanned and characters of Kannada script are manually segmented, then stored in bmp file format. These segmented samples may contain noise due to print quality, scanner etc. To remove the noise, a morphological opening operation is used [11]. The noise free Kannada sample images are normalized into 32 x 32 dimensions. These normalized images are used for the feature extraction. A total of 4,800 handwritten and printed Kannada numeral and vowel samples are used for the experiment. The Figure 1 and Figure 2 shows the sample dataset of handwritten and printed Kannada numerals respectively and Figure 3 and 4 shows the sample dataset of handwritten and printed Kannada vowels respectively.

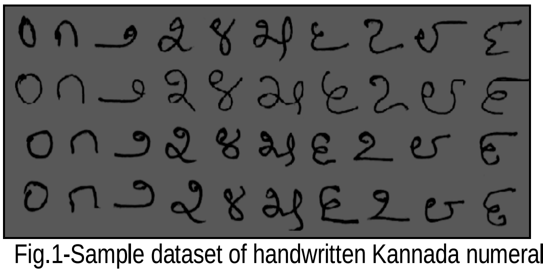


Fig.1-Sample dataset of handwritten Kannada numerals

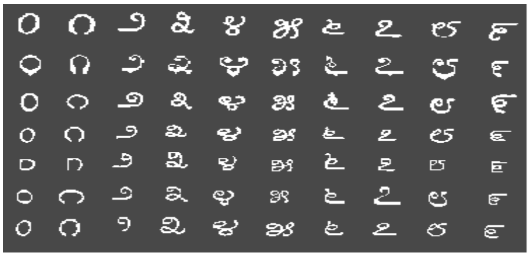


Fig.2 - Sample dataset of printed Kannada numerals with multi-font and multi-size

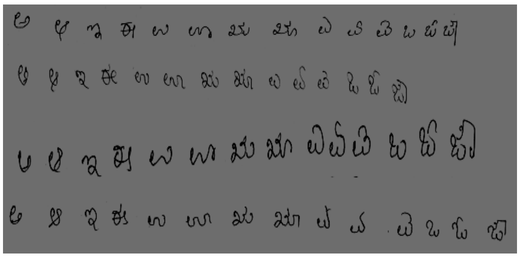


Fig.3 - Sample dataset of handwritten Kannada vowels

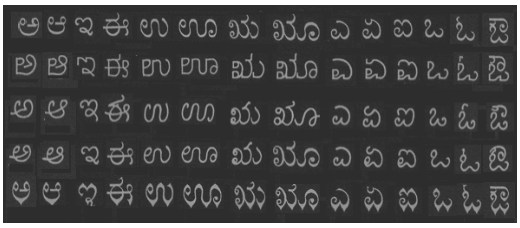


Fig.4- Sample dataset of printed Kannada vowels with multi-font and multi-style.

3. Feature Extraction

The feature extraction is described about the characteristics of an image. It is one of the important components for any recognition system, since the classification/recognition accuracy is depending on the features. Here zone based pixel density features are used for handwritten and printed Kannada numerals and vowels for recognition. All preprocessed images are normalized into 32 x 32 pixels. The normalization is carried out by using bilinear technique. Then normalized image is divided into 64 zones and pixel density is calculated for each zone, there by generating 64 features. These 64 features are fed to SVM classifier for classification of images. The experiment is also carried out by considering large size zones, but it failed to capture the local information of the image. Hence we have fixed the size as 8 x 8 for experimentation as the optimal size and it gives better results.

```
Algorithm: Handwritten and Printed Kannada Numerals
           and Vowels.
Input:     Mixture of Preprocessed numerals/vowels
Output:    Recognize numeral/vowel

Begin
1. The normalized numeral/vowel image is divided into 64
   zones.
2. The pixel density is computed between each of 64
   zones.
3. For empty zones, zero value is assigned.
4. Create a feature vector of size 64.
5. SVM Classifier with 5 Fold cross validation test is
   carried out to recognize an image as the Kannada
   Numeral/Vowel.
End
```

The Figure 5 shows the sample handwritten Kannada Vowel image divided into 8 x 8 zones.

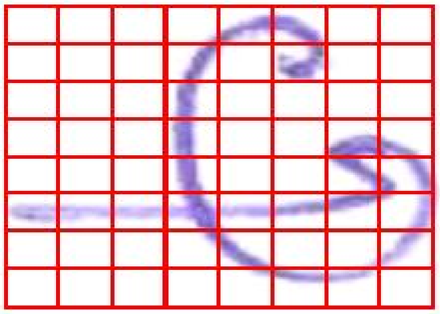


Fig. 5 - Sample Image of Zone Size 8 x 8.

Experimental Results and Discussions

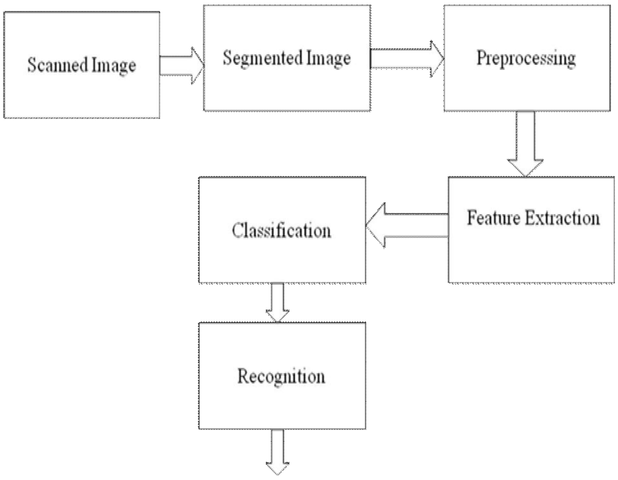
The SVM classifier with 5 fold cross validation test is used to classify 4,800 sample images of handwritten and printed Kannada numerals and vowels, 1000 samples for each of handwritten and printed Kannada numerals and 1400 samples for each of handwritten and printed Kannada vowels. When using the K-fold cross validation method, the dataset is randomly divided into K groups. Then the SVM classifier is used to train K times, using all of the training set data points except those in the Kth group. The proposed algorithm gives reasonable high recognition accuracy for mixture of handwritten and printed Kannada numerals is 97.40% and 95.90% for mixture of handwritten and printed Kannada vowels respectively.Table1 and Table2 shows the average percentage of recognition accuracy for mixture of handwritten and printed Kannada numerals and vowels respectively. The results exhibited in the table are encouraging for the handwritten and printed Kannada numerals and vowels recognition.

Table 1 Average percentage of recognition accuracy for mixture of handwritten and printed Kannada numerals using SVM classifier with five fold cross validation test

Handwritten and Printed Mixed Kannada numerals	K-Fold Cross Validation Method					Percentage of Recognition Accuracy in (%)
	First Fold	Second Fold	Third Fold	Fourth Fold	Fifth Fold	
೦	100.00	98.98	98.73	98.81	97.46	98.79
೧	98.90	99.48	98.18	97.75	95.72	98.00
೨	100.00	100.00	100.00	100.00	100.00	100.00
೩	97.92	96.46	95.70	96.29	94.82	96.23
೪	95.00	98.97	98.20	97.09	97.56	97.36
೫	99.06	98.98	98.98	97.79	98.59	98.68
೬	97.83	95.43	92.64	91.54	89.34	93.35
೭	94.84	93.40	93.65	92.72	90.10	92.94
೮	99.05	100.00	98.48	98.82	99.11	99.09
೯	100.00	100.00	99.29	99.68	98.82	99.55
Average Percentage of Recognition Accuracy in (%)						97.40

Table 2 Average percentage of recognition accuracy for mixture of handwritten and printed Kannada vowels using SVM classifier with five fold cross validation test

Handwritten and Printed Mixed Kannada vowels	K-Fold Cross Validation Method					Percentage of Recognition Accuracy in (%)
	First Fold	Second Fold	Third Fold	Fourth Fold	Fifth Fold	
ಅ	92.55	91.04	89.05	86.23	84.58	88.69
ಆ	98.91	99.00	95.52	92.70	89.92	95.21
ಇ	100.00	99.00	99.00	99.00	98.38	99.07
ಈ	98.86	95.52	98.25	96.01	94.53	96.63
ಉ	97.87	98.01	98.25	96.35	96.01	97.30
ಊ	100.00	98.50	98.01	97.18	97.51	98.24
ಋ	97.22	99.50	99.00	98.83	98.25	98.56
ೠ	100.00	97.50	95.25	95.00	93.62	96.32
ಎ	98.00	98.00	95.25	93.50	91.50	95.25
ಏ	98.00	98.43	93.97	88.48	81.67	92.11
ಐ	97.02	97.00	95.25	93.66	84.50	93.49
ಒ	99.15	100.00	99.50	99.33	98.88	99.37
ಓ	92.93	96.01	95.27	92.53	93.53	94.05
ಔ	100.00	99.00	98.50	98.34	96.27	98.42
Average Percentage of Recognition Accuracy in (%)						95.90



Result
Fig.6- Proposed framework for recognition

Conclusion

In this paper, the single Optical Character Recognition (OCR) algorithm works for both handwritten and printed Kannada numerals and vowels. The proposed algorithm gives encouraging results in an initial attempt to meet the objectives of handwritten and printed Kannada OCR system. The average percentage of recognition accuracy of 97.40% and 95.90% are achieved using SVM Classifier with 5 fold cross validation test for mixture of handwritten and printed Kannada numerals and vowels respectively. The specialty of this work is to combined handwritten and printed numerals and vowels with reduced 24 classes. The novelty of the algorithm is free from thinning and slant of the numerals and vowels. The future plan of this work is to build single OCR system for handwritten and printed Kannada characters.

References

[1] Liana M., Govindaraju V. (2006) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.5, pp 712-724.

[2] Arica N., Fatos T. (2001*) IEEE Transactions on System,* Vol.31, No.2, pp 216-233.

[3] Plamondon R., and Srihari S. N. (2000) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.1, pp 63-84.

[4] Nagy G. (1988) *Proceedings of International Conference on Pattern Recognition*, pp109-114.

[5] Dhandra B.V., Mukarambi G., Hangarge M. (2010*) IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition*, pp 146 -151.

[6] Leena R.R., Sasikumar M. (2011) *International Journal of Computer Theory and Engineering*, Vol.3, No1, pp 1793-8201.

[7] Sangame, Ramteke, Benne R., (2009) *Advances in computational Research*, Vol.1, pp 52-56.

[8] Aradhya M., Niranjan S.K., Hemantha Kumar (2010) *Special Issue of IJCCT*, Vol-1, Issue 2, 3, 4, pp 9-13.

[9] Dhandra B.V., Mukarambi G., Hangarge M. (2011) *Proceedings of International Conference on Computer Science and Information Technology (CSIT)*, pp 47-52.

[10] Acharaya D., Reddy S. (2008) *World Academy of Science, Engineering and Technology*, Vol.42, pp 278-283.

[11] Dhandra B.V., Mukarambi G., Hangarge M. (2011) *Proceedings published by International Journal of Computer Applications (IJCA)*, pp 5-8.