I J M I
International Journal of Machine Intelligence
ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 3, 2011, pp-155-159
Available online at http://www.bioinfo.in/contents.php?id=31

A SCRIPT INDEPENDENT APPROACH FOR HANDWRITTEN BILINGUAL KANNADA AND TELUGU DIGITS RECOGNITION

DHANDRA B.V.[1], GURURAJ MUKARAMBI[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. E-mail: dhandra_b_v@yahoo.co.in,gmukarambi@gmail.com,mhangarge@yahoo.co.in

Received: September 29, 2011; Accepted: November 03, 2011

Abstract- In this paper, handwritten Kannada and Telugu digits recognition system is proposed based on zone features. The digit image is divided into 64 zones. For each zone, pixel density is computed. The KNN and SVM classifiers are employed to classify the Kannada and Telugu handwritten digits independently and achieved average recognition accuracy of 95.50%, 96.22% and 99.83%, 99.80% respectively. For bilingual digit recognition the KNN and SVM classifiers are used and achieved average recognition accuracy of 96.18%, 97.81% respectively.
Keywords- OCR, Zone Features, KNN, SVM

1. Introduction

Recent advances in Computer technology, 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. To recognize such documents developing of handwritten optical character recognition system is essential.

In this direction, many researchers have developed the numeral recognition systems by using various feature extraction methods such as statistical features, topological and geometrical features, global transformation and series expansion features like Hough Transform, Fourier Transform, Wavelets, Moments, etc. A survey on different feature extraction methods for character recognition is reported in [1]. A review on different pattern recognition methods is given in [2]. 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 [3, 4] on Bengali and Tamil scripts, where as the work on handwritten Kannada and Telugu numerals recognition is in infant stage.

Recognition of handwritten Kannada and Telugu characters are 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. Most of the methods have employed fuzzy features [5, 6], templates and deformable templates [7, 8], Structural and Statistical features [9]. Dinesh Acharya et al. [10] have used the 10-segment string, water reservoir, horizontal/vertical strokes, and end point as the potential features for recognition of isolated handwritten Kannada numerals and have reported the recognition accuracy of 90.50%. Drawback of this procedure is that, it is not free from thinning. U. Pal et al. [11] 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.

Recently, a piece of work on bi-lingual and tri-lingual (i.e. English and Devanagri, Kannada, Telugu and Devanagri,) numeral recognition of Indian scripts has been proposed in [12, 13].

From the literature survey, it is evident that most of the authors have been attempted to recognize digits of a script. However, in day today life we come across with many documents which contain bi-script digits. For example, people who are living on the border area of two states like Karnataka and Andhra Pradesh (south Indian states) have a practice of using both Kannada and Telugu digits in a single document (i.e. postal zip codes, application forms, railway reservation slip and bank cheques etc.).

To read such documents monolingual OCR will not work. Therefore, the designing of bilingual and multilingual OCRs is essential for multi-script and multi-lingual country like India. This has motivated us to design a simple and robust bilingual handwritten (Kannada and Telugu) OCR system.

In this paper, Section 2 contains about Kannada/Telugu scripts. Section 3 is about bilingual OCR. The preprocessing of the images and data collection is presented in Section 4. Feature extraction procedure is discussed in Section 5, the experimental details and results obtained are presented in Section 6. Conclusion is the subject matter of Section 7.

2. About Kannada/Telugu Scripts

Kannada is one of the most well known Dravidian languages of India. It is the official language of the southern Indian state of Karnataka. It is spoken by about 44 million people in the Indian states of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. Its writing system is alpha syllabary in which all consonants have an inherent vowel. Other vowels are indicated with diacritics, which can appear above, below, before or after the consonants. When consonants appear at the beginning of a syllable, vowels are written as independent letters and when they appear together without intervening vowels, the second consonant is written as a special conjunct symbol, usually below the first. The direction of writing is left to right in horizontal lines. It has 14 vowels and 36 consonants and 10 digits.

The Telugu script is derived from the Telugu-Kannada script and developed independently at the same time as the Kannada script, as evidenced by their strong orthographical resemblance to one another.

3. Bilingual OCR

Recognition of bilingual documents can be approached 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/Telugu) 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 Telugu 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 20 (Kannada script digits -10 and Telugu script digits - 10). However, as the number of classes increases, it is prudent to divide the classification problem. Hence, the classification problem of Kannada/Telugu bilingual digit recognition is reduced to 11 classes based on the observation that all the digits have the similar shape except digit three of both the scripts.

4. Data Set and Preprocessing

There is no standard database available for south Indian scripts. Hence, we have created own database. Handwritten documents were collected from different professionals belonging to Schools, Colleges, Doctors and Lawyers etc. The collected documents were scanned through a flatbed HP scanner at 300 dpi and binarized using global threshold (i.e. Otus's) method. Unconstrained and isolated 1000 digits of each script were manually segmented from the scanned documents. The segmented digit images contain noise and that arises due to printer, scanner, print quality, etc. Noise removal is performed by employing morphological opening operation. All isolated handwritten Kannada and Telugu digit images are normalized into a common height and width (i.e.32 x 32 pixels).The normalized digit image is used for extracting the features. A sample dataset of handwritten Kannada and Telugu digits are shown in Fig.1 and Fig.2 respectively. As an example normalized Kannada digit is shown in Fig3.

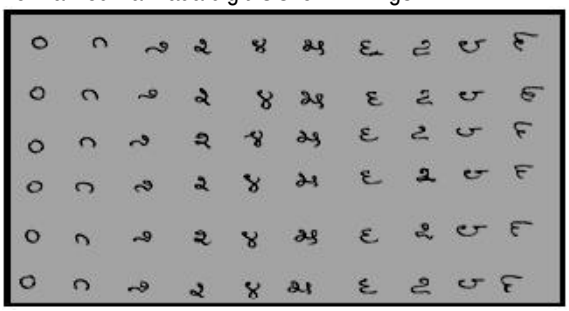

Fig.1 - Sample handwritten Kannada digits 0 to 9

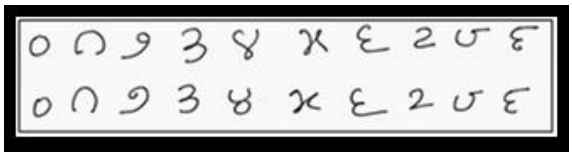

Fig. 2 - Sample handwritten Telugu digits 0 to 9

(a) (b)

Fig.3 - (a) Binary Image: Before Size normalization
(b) Image after Size normalized

5. Feature Extraction

Feature extraction is a problem of extracting the relevant information from the preprocessed data for classification of underlying objects/characters. The preprocessed digit image is used as an input for feature extraction. For extracting the potential feature from the handwritten digit image, the frame containing the preprocessed/normalized image is divided into non-overlapping zones of size 8 x 8 and obtained 64 zones. For each zone, the pixel density is computed and there pixel densities are used as a feature for recognition. Hence, 64 features vector is used for recognition of a digit.

Algorithm: Zone based pixel density feature extraction system

Input: Preprocessed digit Image.

Output: Features for Classification and Recognition.

Start

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 using KNN and SVM classifier.

Stop

6. Experimental Results and Discussions

For experimentation, each 1000 handwritten digits of Kannada and Telugu script were considered. In this paper, we have employed two approaches for experimentations one is script dependent digit recognition and another is bilingual digit recognition. Throughout the experimentation, 50% of the dataset is used for training and the rest is for testing with KNN and SVM classifiers. The experimental results of Kannada digits recognition accuracies are presented in Table 1 and Telugu digits recognition accuracies in Table 2 respectively.

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				
Kannada Digits	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 Table 1, it is clear that the Kannada digit 3 and 7 have the same accuracy approximately, because the digit 3 and 7 are similar in shape.

Table 2-Percentage of handwritten Telugu digits recognition accuracy using KNN (k=1) and SVM Classifiers

Training samples =500, Test samples =500 and Number of features = 64				
Telugu Digits	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy with KNN	Percentage of Recognition Accuracy with SVM
౦	50	50	100.00	100.00
౧	50	50	98.28	100.00
౨	50	50	100.00	100.00
౩	50	50	100.00	100.00
౪	50	50	100.00	100.00
౫	50	50	100.00	100.00
౬	50	50	100.00	100.00
౭	50	50	100.00	100.00
౮	50	50	100.00	98.00
౯	50	50	100.00	100.00
Average Percentage of Recognition accuracy			99.83	99.80

The experimental results have shown high recognition accuracies for Kannada and Telugu handwritten digits of zone size 8 x 8 as 96.22% and 99.80% with SVM classifier. It is to be noted that large size zones have failed to capture the essential part to distinguish the digits. The comparative analysis for handwritten Kannada and Telugu digits recognition are shown in Table 3 and Table 4. Further, to achieve our proposed bilingual system, we have fused both the dataset of Kannada and Telugu and hence the bilingual dataset size is doubled. In this experimentation, the input image may be a Kannada digit or Telugu digit without having any prior knowledge of the input script, digits are recognized. The bilingual digits recognition results are presented in Table 5. As we have discussed in Section 3, nine digits of both the scripts have similar shapes except the digit three, hence the bilingual classification problem has reduced to 11classes.

Table 3-Comparative results for handwritten Kannada digits with other methods

Methods Proposed by	Features and Classifier used	Data set	Percentage of Accuracy
N.Sharma	Structural features k- means classifier	500	90.50
G.G.Rajput	Image Fusion Nearest Neighbor	1000	91.20
G.Hemanth Kumar	Radon transform Nearest Neighbor	1000	91.20
B.V.Dhandra	Template matching, similarity-dissimilarity, binary distance transform, majority voting.	1000	91.00
Proposed	Zoning features K-Nearest Neighbor	1000	95.50

Table 4-Comparative results for handwritten Telugu digits with other methods

Methods Proposed by	Features and Classifier used	Data set	Percentage of Accuracy
R. Sanjeev Kunte	Wavelet descriptors Feed forward network classifier, Five script are considered	2500	92.30
S.V.Rajashekar aradhya	Zoning and Back propagation neural network	2000	96.50
S.V.Rajashek araradhya	Zoning and Nearest Neighbor	2000	97.5
B.V.Dhandra	Structural features and Probabilistic Neural Network	1250	99.60
Proposed	Zoning and Nearest Neighbor	1000	99.83

From Table 3 and 4, it is clear that our proposed method gives high recognition accuracy as compared to other methods found in the literature. Hence, zoning features with K-nearest neighbor classifier is a powerful tool for recognition of Kannada and Telugu handwritten digits independently. Table 5 reveals the high recognition rates of bilingual digits recognition system. Table 6 and 7 presents the confusion of classification with KNN and

SVM classifiers for bilingual recognition system (see Appendix Page). Here, we can notice that more confusion has occurred between the digit 6 and 7, because of both the digits have horizontal stroke at the right side.

Table 5-Average Percentage of handwritten bilingual (mixed digits of Kannada and Telugu Scripts) digits recognition accuracy using KNN and SVM classifiers

Training samples =1000, Test samples =1000 and number of features = 64				
Mixed Kannada and Telugu digits	No. of Sample Trained	No. of Sample Tested	Percentage of Recognition Accuracy with KNN (K=1)	Percentage of Recognition Accuracy with SVM
೦	100	100	100.0	100.0
೧	100	100	96.5	99.0
೨	100	100	99.0	99.5
೩	50	50	91.0	93.0
೪	50	50	100.0	100.0
೫	100	100	98.5	99.5
೬	100	100	96.5	98.5
೭	100	100	96.0	96.0
೮	100	100	82.5	93.5
೯	100	100	99.5	98.5
౬	100	100	98.5	98.5
Average Percentage of Recognition accuracy			96.18	97.81

7. Conclusion

In this paper, a single OCR system for handwritten Kannada and Telugu digits recognition is proposed. The proposed zoning features have shown quite encouraging performance with respect to handwritten Kannada and Telugu digits. We have obtained 95.50% and 99.83% recognition accuracies for handwritten Kannada and Telugu digits respectively for independent samples with KNN classifier. The results obtained are comparable with the existing techniques. Finally, the bilingual (both samples together) recognition accuracy of 96.18%, 97.81% is achieved with KNN, SVM classifier respectively. This algorithm is independent of thinning and slant of the digits. In future, we will extend it for south Indian bilingual digits recognition and also to develop new zone based feature extraction systems which provides the recognition accuracy chase to 100%. Also we plan to improve classifier to achieve still better recognition rate for handwritten characters.

Acknowledgment

This research work is supported by UGC, New Delhi, under Major Research Project grant in Science and Technology; videF.No-F33-64/2007(SR) dated 28-02-08.

References

[1] Oivind Trier, Anil Jain, Torfiinn Taxt (1996) *Pattern Recognition*, Vol 29, No 4, pp 641-662.

[2] Duda R.O., Hart P.E., Stork D.G. *Pattern Classification, 2nd Edition, Wiley-Newyork.*

[3] Rahman A. F. R., Fairhurst M. C. (2002) *Pattern Recognition*, vol.35, pp 997, 1006.

[4] Chandrashekaran R., Chandrasekaran M., Gift Siromaney (1984) *Journal of IETE*, Vol.30, No.6.

[5] Shamic Surel, Das P. K. (2001) *Proceedings Of 6th International Conference on Document Analysis and Recognition (ICDAR), Seattle*, pp 1220-1224.

[6] Nagabhushan P., Angadi S.A.,, Anami B.S. (2003) *Proceedings of NCDAR-2003, Mandy, Karnataka, India*, pp 275-285.

[7] Tubes J.D. (1989) *Pattern Recognition*, 22(4):pp 359-365.

[8] Anil K.Jain, Douglass Zonker (1997) *IEEE, Pattern analysis and machineintelligence*, vol.19, no-12.

[9] Heutte L., Paquest T., Moreau J.V., Lecourtier Y., Oliver C. (1998) *Pattern Recognition*, p.629-641.

[10] Dinesh Acharya U., Subba Reddy N.V. and Krishnamurthy (2007) *IISN-2007*, pp-125 -129.

[11] Sharma N., Pal U., Kimura F. (2006) *9th International Conference on Information Technology (ICIT'06), ICIT*, pp. 133-136.

[12] Benne R.G., Dhandra B.V. and Mallikarjun Hangarge (2009) *Advances in Computational Research*, Volume 1, Issue 2, 2009, pp-47-51.

[13] Lehal G.S. and Nivedan Bhatt (2000) *A Recognition System for Devnagri and English Handwritten Numerals, Springer Berlin vol-1948/2000*, pp.442-449.

[14] Dhandra B.V., Gururaj Mukarambi, Mallikarjun Hangarge (2010) *Special Issue on RTIPPR-10, International Journal of Computer Applications*, pp.146-150.

[15] Sharma N., Pal U., Kimura F. (2006*) ICIT*, pp.133-136, ICIT'06.

[16] Rajput G.G., Mallikarjun Hangarge (2007) *Recognition of Isolated Handwritten Kannada Numeral based on Image fusion method, PReMI-07, Vol. 4815, Springer Kolkata*, pp153-160.

[17] Manjunath Aradhya V. N., Hemanth Kumar G. and Noushath S. (2007) *Proc. of IEEE-ICSCN 2007*, pp-626- 629.

[18] Dhandra B.V., Benne R.G.and Mallikargun Hangargi (2007) *IEEE-ACVIT -07*, pp.1276-1282.

[19] Sanjeev Kunte R. and Sudhakar Samuel R.D., (2006) *VIE*, pp 94-98.

[20] Rajashekararadhya S.V. and Vanaja Ranjan P.V. (2008) *TENCON 2008*.

[21] Rajashekararadhya S.V. and Vanaja Ranjan P.V. (2008) *Journal of Theoretical and Applied Information Technology* (JATIT), pp.1171-1181.

[22] Dhandra B.V., Benne R.G.and Mallikargun Hangargi (2010) *Special Issue on RTIPPR-10, International Journal of Computer Applications*, pp.83-88.

[23] Dhandra B.V., Gururaj Mukarambi and Mallikarjun Hangarge (2011) *International Conference on VLSI, Communication & Instrumentation, (ICVCI – 2011) proceedings published by IJCA,* pp.no 5 – 9.

[24] Basavaraj Patil (2011) *International Journal of Computer Applications*, Volume. 13, No.8.

APPENDIX

Table 6 KNN (Confusion Matrix) for Bilingual OCR

Digits	೦	೧	೨	೩	೪	೫	೬	೭	೮	೯	೯
೦	200	0	0	0	0	0	0	0	0	0	0
೧	4	193	1	0	0	0	1	0	0	0	1
೨	0	0	198	2	0	0	0	0	0	0	0
೩	0	0	1	91	0	1	0	0	7	0	0
೪	0	0	0	0	100	0	0	0	0	0	0
೫	0	0	0	0	1	197	0	1	1	0	0
೬	0	1	0	1	0	0	193	1	3	1	0
೭	0	0	0	0	0	1	0	192	6	0	1
೮	0	0	0	3	1	0	1	29	165	0	1
೯	0	0	0	0	0	0	0	0	0	199	1
೯	0	0	0	0	0	2	0	1	0	0	197

Table 7 SVM (Confusion Matrix) for Bilingual OCR

Digits	೦	೧	೨	೩	೪	೫	೬	೭	೮	೯	೯
೦	200	0	0	0	0	0	0	0	0	0	0
೧	0	198	0	0	0	0	1	0	0	1	0
೨	0	0	199	1	0	0	0	0	0	0	0
೩	0	0	0	93	0	1	4	1	1	0	0
೪	0	0	0	0	100	0	0	0	0	0	0
೫	0	0	0	1	0	199	0	0	0	0	0
೬	0	0	0	2	0	1	197	0	0	0	0
೭	0	0	0	0	0	1	0	192	7	0	0
೮	0	0	0	0	0	1	2	10	187	0	0
೯	0	0	1	0	0	0	0	1	0	197	1
೯	0	0	0	0	0	3	0	0	0	0	197