

Kannada Characters and Numerical Recognition System using Hybrid Zone-Wise Feature Extraction and Fused Classifier

Kavya.T. N*, Pratibha.V, Priyadarshini. B. A
Dept. of Electronics and Communications Engineering
Dr.T.Thimmaiah Institute of Technology
Kolar Gold Fields-563120, Karnataka, India

Vijaya Bharathi. M, Vijayalakshmi.G.V
Dept. of Electronics and Communications Engineering
Dr.T.Thimmaiah Institute of Technology
Kolar Gold Fields-563120, Karnataka, India

Abstract— Character recognition is an important area of research in pattern recognition. Recognition of handwritten characters is difficult because of different writing styles, mood of a person, size of handwritten characters and aging of documents. Kannada characters are symmetric and curvy in nature, hence difficult to recognize in an offline system. In this project for Kannada character script, there is no commercially available database; hence we created our own database. The input image is preprocessed and features are extracted by dividing an image into zones and applying distance metric and pixel density algorithms on the zoned image. These features are combined to form a feature vector. This feature vector is given to the classifier for recognition. k-Nearest Neighbor (kNN) and Linear Discriminant Analysis (LDA) is used for classification. The classifier results are compared and the best result for each character is considered (fusion of classifiers). The overall accuracy of 94.6% for vowels, 84.7% for consonants and 98% for numbers were obtained.

Keywords— Kannada, Pre-processing, Feature Extraction, Zones, Distance Metric, Pixel Density, Classification, KNN, LDA.

I. INTRODUCTION

Pattern recognition and artificial intelligence have been evolving steadily, but only when they are harnessed together, machines will acquire the ability to exploit images like humans [1]. Character recognition is an extensively researched field in pattern recognition, artificial intelligence and machine vision, as it is used in various applications such as postal code identification, automatic plate number recognition, digital libraries, mails etc. Offline recognition of handwritten characters is difficult due to different writing styles, size and shape of the written character and in some cases the person's emotional state also affects the writing style of the character. In India people use more than one language in their day to day life. Each language has its own character set. But there are a lot of similarities between characters among the different languages. Hence it is necessary to design a recognition system for an individual language. Kannada is an official language of Karnataka state, it has 49 characters and its writing style is left to right. The few works done on character recognition so far are: M. Hanmandlu et al used zone / grid based feature extraction for handwritten Hindi numerals where they divided each character image into 24 zones and found distance of each pixel with respect to an absolute reference point [2]. Panyam Narahori sastry et al used zoning method for feature extraction and nearest neighbor classifier on Telugu character set and obtained 78% accuracy [3]. Prema K.V. et al used Gabor

transform to extract features of printed Kannada script and obtained 93.8% accuracy [4]. S. V. Rajashekaraaradhya et al used zone and distance metric based feature extraction method for handwritten Kannada digit recognition and obtained 97% accuracy [5].

In this paper we have detailed our work in different sections, section II gives the details about the proposed method for Kannada character recognition. In section III we have shown the results obtained by using the proposed system. Section IV gives the conclusion and the future work.

II. METHODOLOGY

The Fig. 1 below shows the flow chart for Kannada characters and numbers recognition. The handwritten Kannada character to be recognized is acquired in the form of a grey scale cropped images, this image is preprocessed using size normalization, binarization and thinning to make it more suitable for further analysis. The features are extracted from the preprocessed image by dividing it into zones, finding the distance from centroid of image and zone centroid to each foreground pixel in the zones and finding the density of foreground pixels in each zone. These features are combined together to form a feature vector. The classifier is priorly trained with the features of the training set images. The feature extracted from the testing set is given to the classifier for the recognition of the character.

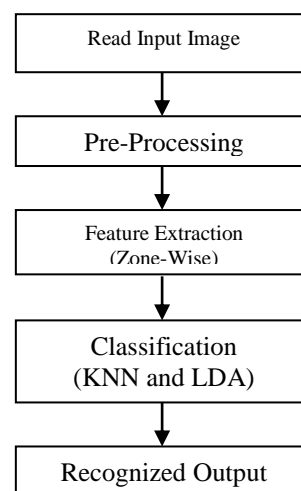


Fig. 1. Flow Chart for Kannada Character Recognition

A. Preprocessing

The pre-processing techniques used are size normalization, binarization and thinning.

Size Normalization: There was no size constraints considered while writing the character. Therefore to bring all the characters into uniform size of 129x129, size normalization is performed [6]. Fig. 2 shows Size Normalization.

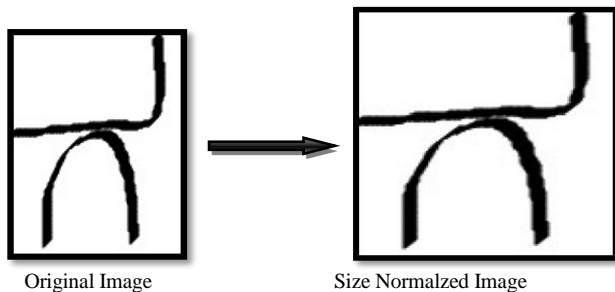


Fig. 2. Size Normalization

Binarization: Images acquired may be in different formats. So they are all converted into binary form for further processing [7]. The output image of binarization has been negated for further processing. Fig. 3 shows the output image of binarization.

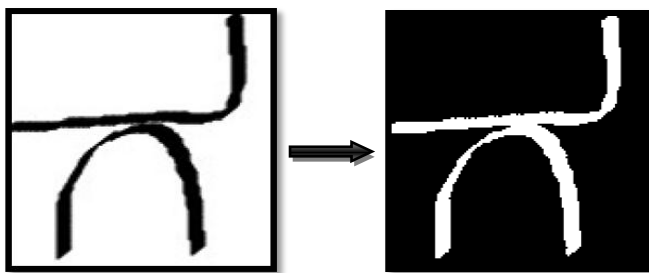


Fig. 3. Binarization

Thinning: When the character is written, the thickness level of pixel varies. So thinning [8] is performed on it to make it one pixel wide. The fig. 4 below shows thinned image.

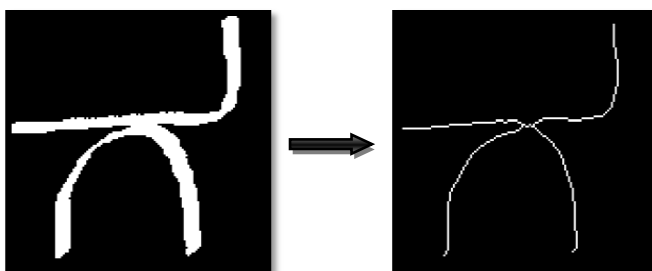


Fig. 4. Thinning

B. Feature Extraction

Feature extraction is an important step in pattern classification. It extracts the important information that characterises each class.

Proposed Algorithm: Finding image centroid, zone centroid [9] with its respective distances and density of the pixels.

1. Compute the input image centroid (shown in fig. 5).

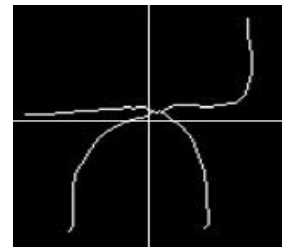


Fig. 5. Image Centroid

2. Divide the input image in to 9 equal zones (as shown in fig. 6).

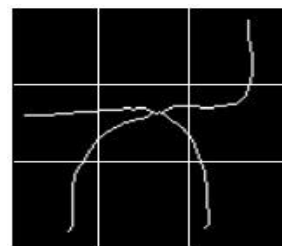


Fig. 6. Dividing Image into Zones

3. Compute the distance ED (Euclidean distance) between the image centroid to each pixel present in the zone.

$$ED(p,q) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2} \quad (1)$$

4. Repeat step 3 for the entire pixel present in the zone.
5. Compute average distance between these points.
6. Repeat this procedure sequentially for the entire zones.
7. Finally, 9 such features will be obtained for classification and recognition.
8. Compute the zone centroid.

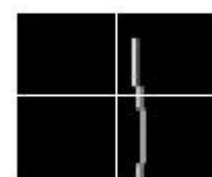


Fig. 7. Zone Centroid

9. Compute the distance between the zone centroid to each pixel present in the zone, compute average distance between these points.
10. Repeat this for all the zones in the image. Finally 9 such features are obtained.
11. Compute the density of pixels present in each zone.
12. Repeat step 11 for all the zones present in the image to obtain 9 such features.
13. Repeat step 11 to 12 for 4 equal zones to obtain 4 such features.
14. Combine all the features obtained to form a feature extraction vector (9+9+9+4).

The final feature vector is formed by combining the output of algorithm with the bounding box, and ratio of major axis length and minor axis length. The feature vector of $9+9+9+4+2$ (i.e.33 features) for a single character image is obtained.

C. Classification

Classification is the process of correctly assigning unknown patterns to their respective pattern class [1]. In our work we have fused LDA and KNN by comparing the classifier outputs to obtain a higher accuracy. The nearness between the test sample and every sample in the database is determined by KNN [24] using Euclidian distance as depicted in (1). The least distance between them is used to recognize the character. The degree of variability of classes is comparatively less in KNN. LDA recognizes the character by finding the difference between the classes. It is closely related to analysis of variance. It reduces the variance between same classes and increases the variance between different classes. LDA [24] is used because it gives good performance for multifont independent Optical Character Recognitions (OCR's), robust to noise and adaption across languages.

III. RESULTS AND DISCUSSIONS

A. Database

Kannada has 49 characters which include 15 vowels and 34 consonants. The number of theoretically possible combinations of vowel- consonant is 510 and vowel-consonant-consonant in Kannada script is 17340. As the number of classes is more, as an initial attempt we have considered only vowels, consonants and numbers separately for recognition. As there is no database available commercially we created our own database. For the purpose of training we collected 10 samples for each numbers, 25 samples each for vowels and 30 samples each for consonants. For testing purpose, 5 samples each for vowels, consonants and numbers is been collected. Therefore total of 1495 samples for training and 295 samples for testing was collected to create a database. Fig. 8 shows few samples of the database.

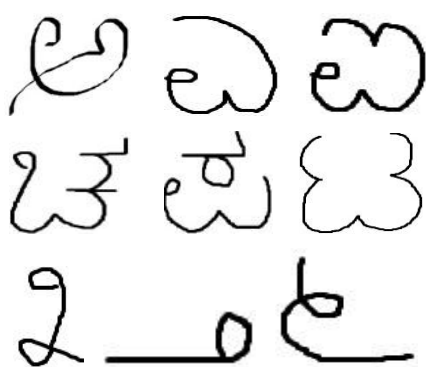


Fig. 8. Samples from the Database

As our work involved recognition of Kannada characters and numerals we initially considered only distance metric to extract features but accuracy obtained was less, therefore we also computed pixel density and included it in the feature vector.

B. Experiment on vowels

In our project we collected 30 samples for each character in vowels out of which 25 samples were taken for training and 5 samples were taken for testing purpose. The features are extracted from the pre-processed training set images, which were used to train the classifier. To extract features we initially considered only pixel distance which resulted in less accuracy, so in order to increase the accuracy we also considered pixel density, bounding box along with ratio of major axis and minor axis. We obtained an accuracy of 80% for KNN and 86.6% for LDA classifiers. But the characters that are easily recognized by one classifier were not recognized by the other classifier which resulted less accuracy, thereby to achieve a higher recognition rate in classifier level we have fused the classifiers. The classifiers are fused such that the recognition results obtained by both the classifiers are compared and one of which gives a higher accuracy is considered as the final result for the character. The overall accuracy obtained using the proposed algorithm after fusing the classifiers is 94.6%. Table 1 shows the experimental results for vowels using different zones and varying the number of training samples. Table 2 shows the result of fused classifier for vowels compared with individual classifiers. For testing purpose 5 samples for each character was used.

Table 1: Tabulation of Vowels Results

| | KNN | LDA |
|--|-----|-------|
| 16 ZONES, 10 TRAINING SAMPLES | 32% | 41% |
| 16 ZONES, 20 TRAINING SAMPLES | 38% | 53% |
| 9 ZONES, 20 TRAINING SAMPLES | 66% | 73% |
| 9 ZONES, PIXEL DENSITY 20 TRAINING SAMPLES | 69% | 74% |
| 9 ZONES, PIXEL DENSITY 25 TRAINING SAMPLES | 80% | 86.6% |

Table 2: Tabulation of Vowels Results for Fused Classifier.

| | |
|------------------|-------|
| KNN | 80% |
| LDA | 86.6% |
| FUSED CLASSIFIER | 94.6% |

C. Experiment on consonants

For consonants we collected 35 samples for each character, out of which 30 samples were taken for training and 5 samples were taken for testing purpose. The features are extracted from the pre-processed training set images, which were used to train the classifier. To extract features we initially considered only pixel distance which resulted in less accuracy, so in order to increase the accuracy we also considered pixel density, bounding box along with ratio of major axis and minor axis. We obtained an accuracy of 73% for KNN and 80% for LDA classifiers. But the characters that are easily recognized by one classifier were not recognized by the other classifier which resulted less accuracy, thereby to achieve a higher recognition rate in classifier level we have fused the classifiers. . The overall accuracy obtained using proposed algorithm after fusing the classifiers is 84.7%. Table 3 shows the experimental results for consonants using different zones and varying the number of training samples. Table 4 shows the result of fused classifier for consonants compared with individual classifiers. For testing purpose 5 samples for each character was used.

Table 3: Tabulation of Consonants Results

| | KNN | LDA |
|--|-----|-----|
| 16 ZONES, 10 TRAINING SAMPLES | 21% | 25% |
| 16 ZONES, 20 TRAINING SAMPLES | 48% | 57% |
| 9 ZONES, 20 TRAINING SAMPLES | 67% | 70% |
| 9 ZONES, PIXEL DENSITY 25 TRAINING SAMPLES | 71% | 76% |
| 9 ZONES, PIXEL DENSITY 30 TRAINING SAMPLES | 73% | 80% |

Table 4: Tabulation of Consonants Results for Fused Classifier

| | |
|------------------|-------|
| KNN | 73% |
| LDA | 80% |
| FUSED CLASSIFIER | 84.7% |

D. Experiment on numbers

For numbers we collected 15 samples for each number, out of which 10 samples were taken for training and 5 samples were taken for testing purpose. The features are extracted from the pre-processed training set images, which were used to train the classifier. The overall accuracy obtained using the proposed algorithm and fused classifier is 98%. Table 5 shows the result of fused classifier for numbers compared with individual classifiers. For testing purpose 5 samples for each character was used.

Table 5: Tabulation of Numbers Results for Fused Classifier

| | |
|------------------|-----|
| KNN | 96% |
| LDA | 98% |
| FUSED CLASSIFIER | 98% |

IV. CONCLUSION AND FUTURE WORK

In our work we have presented a Hybrid zone based feature extraction algorithm and fused classification for handwritten offline Kannada character recognition. For feature extraction the combination of distance metric and pixel density algorithms are considered. For fusion of classifiers KNN and LDA is used. We have obtained 94.6%, 84.7% and 98% recognition rate for Kannada vowels, consonants and numerical respectively. By using zone based feature extraction we have achieved good recognition result even when certain preprocessing steps like filtering, smoothing and slant removing are not considered. By using 9 zones a higher accuracy is obtained compared to other zones. Using pixel density along with distance metric gives a greater recognition rate than using only distance metric. The recognition rate of fused classifier, provide best results than individual KNN and LDA classifier. This recognition system can be used for both printed and handwritten characters. This system can be used to recognize characters of other languages given that the classifier is trained with database of those languages. Our results are comparable to those using zone wise

feature extraction methods in terms of accuracy for offline handwritten Kannada characters and numerical recognition system.

The future work aims to improve the accuracy (recognition rate) by improving the features and using different classifiers. Also we can extend the work for recognition of Kagunita, words and higher level of Kannada scripts and also for other languages.

REFERENCES

- [1] G.W. Awock and R.Thomas "applied image processing", McGraw hill international editions.
- [2] M. Hanmandlu, J. Grover, V. K. Madasu, and S. Vasikarla, "Input fuzzy for the recognition of handwritten Hindi numeral:", International Conference on Informational Technology, vol. 2, pp. 208-213., 2007.
- [3] Panyam Narahori Sastry, T. R. Vijaya Lakshmi, N. V. Koteswara Rao, T. V. Rajinikanth, Abdul Wahab, "Telugu Handwritten Character Recognition Using Zoning Features", IEEE, 2014.
- [4] Siddhaling Urologin, Dr. Prema K. V., Dr. N. V. Subba Reddy, "A Gabor Filter Based Method for Segmenting Inflected Characters of Kannada Script", Fifth International Conference on Industrial and Information Systems, 2010.
- [5] S.V. Rajashekararadhy, P. Vanaja Ranjan and V.N. Manjunath Aradhya, "Isolated handwritten Kannada and Tamil numeral recognition: A novel approach", First International Conference on Emerging Trends ICETET 08, 2008.
- [6] Chun Lei He, Ping Zhang, Jianxiong Dong, Ching Y. Suen, Tien D. Bui "The Role of Size Normalization on the Recognition Rate of Handwritten Numerals" Centre for Pattern Recognition and Machine Intelligence, Concordia University Montreal, Quebec, Canada H3G 1M8.
- [7] Aroop Mukherjee, Soumen Kanrar "Enhancement of Image Resolution by Binarization" International Journal of Computer Applications (0975 – 8887) Volume 10– No.10, November 2010
- [8] Kamaljeet Kaur, Mukesh Sharma "A Method for Binary Image Thinning using Gradient and Watershed Algorithm" IJARCSSE Volume 3, Issue 1, January 2013.
- [9] S.V. Rajashekararadhy, and P. Vanaja Ranjan, "Isolated handwritten Kannada digit recognition: A novel approach", Proceedings of the International Conference on Cognition and Recognition", pp.134-140, 2008.
- [10] Archana. N. vyas, Mukesh. M. Goswami, "Classification of Handwritten Gujarati Numerals" International Conference on Computer & Communication Technology (ICCT)-2015.
- [11] Dhandra B. V., Mallikarjun Hangarge, Gururaj Mukarambi, "Spatial Features For Handwritten Kannada And English Character Recognition", IJCA, Special Issue on Recent Trends on Image Processing, 2010.
- [12] G.G.Rajput, Rajeshwari Horakeri, "Shape Descriptors based Handwritten Character Recognition Engine with Application to Kannada Characters" International Conference on Computer & Communication Technology (ICCT)-2011.
- [13] Huiqin Lin, Wennuan Ou, Tonglin Zhu, "The Research of Algorithm for Handwritten Character Recognition in Correcting Assignment System" 2011 Sixth International Conference on Image and Graphics, china.
- [14] Leena R. Raghu, Sasi Kumar M., "Feature Extraction for Handwritten Kagunita Recognition", International Journal of Computing Theory and Engineering, 2011.
- [15] Leena Raghu, M. Sasikumar, "Adapting Moments for Handwritten Kannada Kagunita Recognition", 2010 Second International Conference on Machine Learning and Computing.
- [16] Madhavaraj A, A.G.Ramakrishnan, Nagaraj Bhat, "Improved Recognition of Aged Kannada Documents By Effective Segmentation Of Merged Characters" ACM - Proceedings of the International Workshop on Multilingual OCR, 2009.
- [17] Nafiz Arica, and Fatos T. Yarmar-Vural, "Optical Character Recognition for Cursive Handwriting" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 6, June 2002.

- [18] Niranjan. S. k., Vijaya Kumar, Hemanth Kumar G., Manjunath Aradhya V. N., " FLD Based Unconstrained Handwritten Kannada Character Recognition", Second International Conference on Future Generation Communication and Networking Symposis, 2008.
- [19] Padma M. C., Saleem Pasha, "Quad Tree Based Feature Extraction Technique for Recognizing Handwritten Kannada Character", Proceeding of International Conference on Emerging Research in Electronics, Computer Science and Technology, 2013.
- [20] Rajib Ghosh, Partha Pratim Roy, "A Novel Feature Extraction Approach for Online Bengali and Devanagari Character Recognition" 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN).
- [21] Rituraj Kunwar, Shashi Kiran K., Ramakrishna. A. G., "Online Handwritten Kannada Word Recognizer with Unrestricted Vocabulary", 12th International Conference on Frontiers in Handwriting Recognition, 2010.
- [22] Sangame. S. K., Ramteke. R. J., Rajkumar Benne, "Recognition of Isolated Handwritten Kannada Vowels", Advances in Computational Research, 2009.
- [23] Thungamani. M., Dr. Ramakhanth Kumar P., Keshava Prasanna, Shrivani Krishna Rau, "Offline Handwritten Kannada Text Recognition Using Support Vector Machine And Zernike Moments", International Journal of Computer Science and Networking Security, 2011.
- [24] http://en.wikipedia.org/k_nearest_neighbors_algorithm.
- [25] https://en.wikipedia.org/linear_discriminate_analysis.
- [26] Abubakar Muhammad Ashir, Gaddafi Sani Shehu, "Adaptive Clustering Algorithm for Optical Character Recognition", ECAI 2015 - International Conference – 7th Edition, Bucharest, Romania.
- [27] B.M. Sagar, Dr. ShobhaG, Dr. Ramakanth Kumar P, "Complete Kannada Optical Character Recognition with Syntactical Analysis of the script" Proceedings of the 2008 International Conference on Computing, Communication and Networking (ICCCN 2008).
- [28] Shwetha D, Mrs. Ramya S, "Comparison of Smoothing Techniques and Recognition Methods for Online Kannada Character Recognition System" IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), Unnao, India.
- [29] U. Pal, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition of six popular scripts", Ninth International conference on Document Analysis and Recognition ICDAR 07, Vol.2, pp.749-753, 2007.
- [30] Venkatesh Narasimha Murthy, Angarai Ganesan Ramakrishnan, "Choice of Classifier in Hierarchical Recognition of Online Handwritten Kannada and Tamil Akshara", Journal of International Computer Science, 2011.