A Comparative Study of Different Feature Extraction and Classification Methods for Recognition of Handwritten Kannada Characters

Mr. Y. C Kiran
Research Scholar, Jain University Bangalore, India
Associate professor, Department of ISE,
Dayananda Sagar College of Engineering

Ms. Roopa R. Tonashyal
M. Tech Student, Department of ISE,
Dayananda Sagar College of Engineering,
Bangalore, India

Abstract- **Handwritten Kannada Character Recognition has been a challenging research domain because handwritten is different for different persons. Handwriting is one of the most widely used communication technique. There is a need to convert these handwritten documents into an editable format which can be achieved by Handwritten Character Recognition systems. This significantly reduces the storage space. This paper focuses on offline handwritten kannada characters. Gives overview of the ongoing research in Optical Character Recognition systems. It gives brief overview of different methods for segmentation, feature extraction as well as for classification. There is mainly two methods for character recognition one is offline handwritten character recognition and another one is online handwritten character recognition. It mainly focuses on offline handwritten character recognition.**

Keywords:— Handwritten character recognition, Kannada Script, Directional chain code, Run Length Count, K-Nearest Neighbor, Linear classifier, Optical Character Recognition (OCR).

I. INTRODUCTION

A character is a system consisting of symbols that are used to represent information of a particular language. A character is represented as a entity using a number of writing tools so that the information present can be transmitted to the reader. Character recognition is a topic being studied in depth. Human's ability to recognize characters is being studied in its advanced form and systems are being developed to simulate this ability. Character recognition performs an important role in automating insertion and human-computer interface.

The Handwritten character recognition is an important area of research for its applications in banks, post office, and other organizations. An extensive attention has been received by Handwritten Character Recognition (HCR) in the fields of academics and production. Character Recognition Systems can be classified as on-line or off-line. The process of finding letters and words present in digital image of handwritten text is called as off-line handwriting recognition. Research in HCR has become popular because of many practical applications such as reading aid for the blind, recognition of numbers on bank cheques, automatic pin code reading for sorting of postal mail etc.

A lot of work has been done on the recognition of printed characters of Indian languages. On the other hand, attempts made on the recognition of handwritten characters are few. Most of the research in this area is mainly focusing on recognition of off-line handwritten characters for the scripts like kannada, Devanagari and Bangla . From the literature survey it can be seen that there is a lot of demand for character recognition systems for Indian scripts and an excellent review has been done on the Optical Character Recognition (OCR) for Indian languages.

II. DESCRIPTION OF THE KANNADA SCRIPT

Kannada is the official language of South Indian state Karnataka. Kannda language is mainly derived from Bramhi
script. It is one of the earliest languages evidenced epigraphically in India and spoken by about 50 million people in the Indian state of Karnataka ,Tamil Nadu, Andhra Pradesh and Maharashtra. Kannada script consist of set of 52 characters, comprise 16 vowels and 36 consonants. The scripts also include 10 different Kannada numerals of the decimal number system. Further there are distinct symbols that modify the base consonants called consonant and vowel modifiers. The numbers of these modifiers are same as that of the base characters. These characters are called as aksharas, are formed by graphically combining the symbols corresponding to consonants, consonant modifiers (optional) and vowel modifiers using well defined rules of combination.

The number of possible consonant-vowel combinations is 36*16=576 and number of possible consonant-consonant-vowel combinations is 36*36*16=20736, as there are 16 vowels and 36 consonants. Some kannada characters are Shown in figure 1

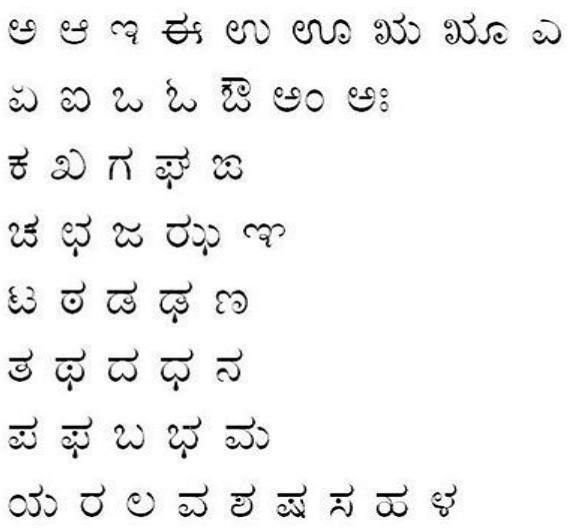

Fig. 1. Kannada characters

III. DATA SET AND PREPROCESSING

For OCR performance evaluation a standard database of character images is required which is lacking in Kannada script. In order to build the dataset for our experimentation, a dataset was created by collecting various handwritten samples from writers belonging to different categories comprising of different age group like officials, student, housewives etc. The samples were scanned through a flatbed HP scanner at 300 dpi. Isolated characters were obtained by manual cropping.

Initially the color images were converted to gray scale and in turn the gray scale images were converted to binary using global threshold method. Thinning is applied on the binary image. Thinning is an image preprocessing method performed to make the image crisper by reducing the binary valued image regions to lines that approximate the skeletons of the region. By using thinned binary image Region labelling is performed and a minimum rectangle bounding box is inserted over the character. The bounding box image would be of variable size due to different style and size of character. Hence this image is resized to desired size.

IV. FEATURE EXTRACTION

A. *Zonal based Feature Extraction*

The preprocessed image is resized to 60 x 60.The resized image is divided into zones or blocks of 5 x 5 to obtain the features. A feature vector is then computed by considering the number of on pixels in each zone. For each zone if the number of on pixels is greater than 5% of total pixels, then the value is stored for that block. There are totally 12*12 Zones in a image therefore The size of the feature vector is 144.The extracted features of characters are then used for training and classification.

B. *Image Fusion*

Here the extracted features of the several character images are fused to generate patterns, these patterns are stored in 8x8 matrices, irrespective of the size of character images. Zonal based feature extraction algorithm is used to extract the features of handwritten Kannada characters.

steps of feature extraction module

- Divide the normalized image into 64 zones of equal size.
- Create a Pattern Matrix (Mp_{tj}) of size 8x8.
- For each zone do this if the number of on pixels is greater than 5% of total pixels, store 1 in Mp_{tj}.
- Fusion the reduced image Mp_{tj} with Pm_j using the equation (1)

$$P_{Mj}^{New} = \left(\frac{1}{num+2}\right) * \left(num * P_{Mj}^{old} + M_{p_{tj}}\right) \qquad (1)$$

$$0 \le j \le 9$$

In equation (1) P_{Mj}^{New} is the fused pattern matrix obtained after fusing training images contained in Mp_{tj} and P_{Mj}^{old}(already stored in the pattern matrix table). $P_{Mj}{}^{New}$ is copied back to the table along with num increased by 1. The content of each cell of fused pattern matrix represents the probability of occurrence of a white pixel that is mapped with the test image to a typical Character.

C. *Radon Transform*

Radon transform is used as one of the feature extraction methods. In Radon transform, 50 diverging beams are used to compute the features. It is seen from the accumulator data
that the projections taken from 0 to 180 degree are exactly equal to the projections taken from 181 to 360 degree. For projection data average value is calculated is taken to build the feature vector. Average is taken for 180 degree rotation of angle theta that is once we get the projection vector, we take the average of the pixels at the co-ordinates from 0 degree to 179 degrees. Hence size of feature vector for one numeral is 703 x 1.

D. *Fan Beam Projection*

Fan beam projection is a variation of Radon transform. The fan-beam function computes projections of an image matrix along specified directions except that the projections are taken in a different way from that of Radon transform. Features were computed using fan-beam geometry. For Fan-beam, 55 diverging beams are calculated. First step is to determine the distance D from the fan-beam source to the center of rotation.

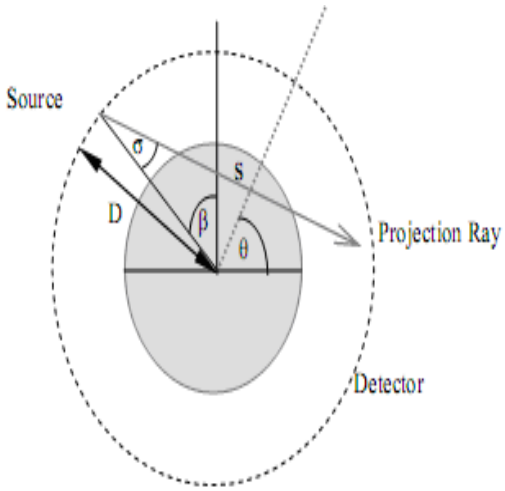

Fig. 2. Fan-beam Geometry

D must be large enough to ensure that the fan-beam source is outside of the image at all rotation angles. D is taken a few pixels larger than half the diagonal image distance, where the diagonal image distance is, d= $\sqrt{i^2 - j^2}$ here i and j are rows and columns of the image respectively. Fan-beam takes projections at different angles by rotating the source around the center pixel at θ degree intervals. These projection data is considered to be feature vector. It can be seen from the accumulator data of Fan beam that after 180 degree the signal repeats itself in the reverse direction. This is due to the reason that projections taken from 0 to 180 degree are exactly equal to the projections taken from 181 to 360 degree. The average value of the obtained projection data is computed in order to build the feature vector. For each Fan-beam the average of the projections of one direction is equal to the average of 55.

Parallel projections were computed. therefore the size of feature vector for one character is 1×360

E. Chain Code Features

Steps involved in this method are as follows

Step1: Starting point is determined for a character and store it in the database
Step2: All 8 neighbors are travelled
Step3: Find out the first nonzero value
Step4: Add this nonzero value to chain code list
Step5: Move to next position
Step6: Determine whether we reach to first point or not if not then go to step 2.

F. Discrete Fourier Transform

The discrete Fourier transform (DFT) converts a finite list of equally spaced samples of a function into the list of coefficients of a finite combination of complex sinusoids, ordered by their frequencies, that has those same sample values. It can be said to convert the sampled function from its original domain to the frequency domain.

The formula to compute the Discrete Fourier transform is in Equation (2).

$$X(k) = \sum_{\mathbb{N}}^{j=1} x(j)\omega_N^{(j-1)(k-1)} \qquad (2)$$

Where $\omega_N = \sum_{\mathbb{N}}^{j=1} e^{(-2\pi i)/N}$ and N is the total number of samples

For each block obtain the average pixel values of rows and columns and then extract the features by applying the Discrete Fourier Transform for each of the row and column vectors.

G. Run Length Count

In this method, First divide the entire image into 9 equal zones. In a image, whenever a pixel value changes from 0 to 1 or 1 to 0 it denotes that the information represents an edge. This information is very valuable as it denotes the geometry of the character and helps in identifying the character . To capture this information, we used Run Length Count (RLC) technique. In this method, for every zone, we find the Run Length count in horizontal and vertical direction. For each character totally 18 features will be extracted and this will serve as feature vector.

H. Diagonal Based Feature Extraction

Character image is converted into 90x 60 pixels. Every character is divided into 54 equal zones, each zone has 10x10 pixels . For each Zone features are extracted by moving along the diagonals of its respective 10x10 pixels. Each zone has19 diagonal lines .A line is summed to get a single sub-feature, thus totally 19 sub-features are calculated for each zone. A single feature value is calculated from 19 sub-features values and placed in the respective zone. This procedure is sequentially repeated for the all the zones. There may be some zones whose diagonals are empty. The feature values of these zones are zero. Totally, 54 features are extracted for each character.

V. CLASSIFICATION

A. K-Nearest Neighbor

The K-Nearest Neighbor Classifier is an efficient technique which is used when the classification problem has pattern classes that display a reasonably limited degree

of variability. It considers each input character given to it and classifies it to a certain class by calculating the distance between the input pattern and the training patterns. It takes into account only k nearest prototypes to the input pattern during classification. The decision is generally based on the class values of the k nearest neighbors. In the k-Nearest neighbor classification, Here compute the distance between features of the test sample and the feature of every training sample. Some measures of this method are described below

1. *Euclidean Distance Metric*

Euclidean distance between P and Q is computed using the equation (3) given below.

$$D=\|P-Q\| \qquad (3)$$

Where P represents the input test image and Q is the trained images of the classes in the database. The nearest neighbor method assigns the test character to a class that has the minimum distance. The corresponding character is declared as recognized character.

2. *Chebyshev Distance Metric*

The Chebyshev distance between two vectors or points p and q, with standard coordinates pi and qi, respectively is given by the equation (4)

$$(p,q) = (|p_i - q_i|) \quad 0 \le i \le n \qquad (4)$$

3. *Manhattan Distance Metric*

The Manhattan distance function computes the distance that would be travelled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two points is the sum of the differences of their corresponding components. The distance between a point P= (p1, p2 ...) and a point Q= (q1, q2 …), with standard coordinates pi and qi respectively is given by the following equation (5)

$$D(X) = \sum_{n}^{i=1} |p_i - q_i| \qquad (5)$$

B. *Clustering*

A cluster-based classification scheme is proposed to speed up the classification process. Several recent advances have been made in Computer Vision by incorporating clustering algorithms for the canonicalization of large data sets: selecting exemplars, building unsupervised object recognizers, learning in low-level vision and text on generation. Classification can be performed in two steps: The training mode, where the feature vectors of learning samples are clustered first based on a certain criterion. Then the classification mode, where the distance between a test sample and every cluster is calculated, and the clusters that are nearest to the test sample are chosen as candidate clusters. Then the classes within those candidate clusters are selected as the candidates of the test sample. Here, we have used two clustering methods- K-means and K-medoids for classification.

C. *Linear Classifier*

Linear classifier is a statistical classifier which makes a classification decision based on the value of the linear combination of the features. When the speed of classification is an issue a linear classifier is often used, since it is often the fastest classifier. Linear classifiers often work very well when the number of dimensions in feature vector is large as represented by equation (6).

$$y = f(\vec{w}.\vec{x}) = f\left(\sum_{j} w_j x_j\right) \qquad (6)$$

where w_j is weight vector, learned from a set of labeled training samples. X_j is the feature vector of testing sample. f is a simple function that maps the value to the respective classes based on a certain threshold.

D. *Artificial Immune System*

Artificial Immune System (AIS) based classification approach is relatively new in the field of pattern recognition (PR). Artificial Immune System (AIS) is inspired by mammalian immune system and has so far been applied in computer security, network intrusion detection, optimization, fraud detection, data analysis and machine learning, clustering, fault and anomaly detection, associative memories, control and scheduling, pattern recognition. The algorithm has two stages: (i) training and (ii) classification. Immune memory produced in training phase is used during classification stage.

E. *Classifier Fusion*

It is observed from the real life classification problems that usually the features are spread in many different ways and a common approach is to combine the selected features of different categories which are complementary into a single feature vector. However, when different types of features are combined into the same feature vector, some large-scaled features may dominate the distance, while the other features do not have the same impact on the classification. Otherwise, different classifiers can be used to classify based on each visual feature individually. The final classification can be obtained based on the combination of separate base classification results. Hence the non-homogeneous properties of individual features do not necessarily affect the final classification directly. Each feature has its own effect on the classification result. It has

been found that a consensus decision of several classifiers can give better accuracy than any single classifier. Thus, in the recent years combining classifiers has become a popular research area. The aim of combining classifiers is to form a consensus decision based on the opinions provided by different base classifiers. Combined classifiers have been applied to several classification tasks, for example to handwritten character identification, face recognition and fingerprint verification. Here, we have used two feature extraction methods to extract the features. These features are classified using k-NN and linear classifiers. The outputs of these classifiers are fused to get the final decision.

F. Neural Network

Neural Network is set of connected input and output units. Where each connection has a weight associated with it. During learning phase, the networks learns by adjusting the weights so as to be able to predict the correct class label of input samples. The input are fed simultaneously into a layer of units making up the input layer .The weighted outputs of these units are in putted simultaneously to a second layer of "neuron-like" units, known as hidden layer. The hidden layer's weighted outputs can be input to another hidden layer, and so on. The number of hidden layers are arbitrary. The weighted outputs of last hidden layer are input to units making up the output layer, which emits network's prediction for given samples.

VI. CONCLUSION

Thus here discussed so many feature extraction and classification methods here. As handwritings of various persons are different therefore each approach produces solution for few characters. For this work different handwritings are collected from different writers. challenges still prevails in the recognition of normal as well as abnormal writing, similar shaped characters, slanting characters, joined characters, curves and so on during recognition process.

REFERENCES

[1] Swapnil A. Vaidya, Balaji R. Bombade, "A Comprehensive Survey on Kannada Numerals and Character Recognition", ISSN: 2277 128X,March 2013.

[2] Rajashekararadhya S. V., Vanaja Ranjan P., Manjunath Aradhya V. N., "Isolated Handwritten Kannada and Tamil Numeral Recognition: A Novel Approach", First International Conference on Emerging Trends in Engineering and Technology- ICETET, pp.1192-1195,16-18 July 2008.

[3] Mamatha H. R., Karthik S., Srikanta Murthy K., "Feature Based Recognition of Handwritten Kannada Numerals – A Comparative Study", International Conference onComputing, Communication and Applications (ICCCA),22-24 Feb, 2012.

[4] S.V. Rajashekararadhya, p. Vanaja Ranjan, "Handwritten Numaral Recognition of kannada script",2009

[5] J.Pradeep, E.Srinivasan and S.Himavathi, "Diagonal based feature extraction for handwritten alphabets recognition system using neural network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[6] H. Imran Khan, Smitha U. V, Suresh Kumar D. S, "Isolated Kannada Character Recognition using Chain Code Features", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064.

[7] Mamatha H.R, Karthik S, Srikanta Murthy K, "Classifier Fusion Method to Recognize Handwritten Kannada Numerals".

[8] Thungamani.M, Dr Ramakhanth Kumar P, Keshava Prasanna, Shravani Krishna Rau, "Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments", IJCSNS International Journal of Computer Science and Network Security, VOL.11 No.7, July 2011.

[9] Mamatha.H.R, Sucharitha Srirangaprasad, Srikantamurthy K, "Data fusion based framework for the recognition of Isolated Handwritten Kannada Numerals", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, 2013.

[10] Jomy John, Pramod K. V, Kannan Balakrishnan, "Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011.

[11] M. Miciak, "Character recognition using Radon Transformation and Principal Component Analysis in postal applications", Proc. of International Multi conference on Computer Science and Information Technology, (2008) October 20-22, pp. 495-500.

[12] H. R. Mamatha, K. Srikanta Murthy, P. Vishwanath, T. S. Savitha, A. S. Sahana and S. Suma Shankari, "Evaluation of Similarity Measures for Recognition of Handwritten Kannada Numerals", CiiT International Journal of Digital Image Processing, ISSN 0974-9691 and Online: ISSN 0974-9586, DOI: DIP102011018, vol. 3,no. 16, (2011) October, pp. 1025-1029.

[13] A. Fitzgibbon and A. Zisserman, "On Affine Invariant Clustering and Automatic Cast Listing in Movies", Proceedings of 7th European Conference on Computer Vision, ECCV, vol. 3, (2002), pp. 304-320.

[14] G. G. Rajput, R. Horakeri and S. Chandrakant, "Printed and Handwritten Kannada Numeral Recognition Using Crack Codes and Fourier Descriptors Plate", IJCA Special Issue "Recent Trends in Image Processing and Pattern Recognition", RTIPPR, (2010), pp. 53-58.

[15] H. R. Mamatha, K. Srikanta Murthy, K. S. Amrutha, P. Anusha and R. Azeemunisa, "Artificial Immune System based Recognition of Handwritten Kannada Numerals", Advanced Materials Research, ©(2012) Trans Tech Publications, Switzerland, doi:10.4028/ www. Scientific.net/AMR. 433-440.900, vol. 433-440,(2012), pp. 900-906.