BPNN Approach on Graph - Based Printed Kannada Character Recognition System

Gurubasava[1], Rajesh Budihal[2]
[1]*Jain University, Banglore.*
[2] *Jain University, Banglore.*
(E-mail: shreeguru91@gmail.com, rajeshbudihal@gmail.com)

Abstract— Many commercial OCR systems are now available in the market, but most of those systems work for Roman, Chinese, Japanese and Arabic characters. There are no sufficient number of works on Indian language character recognition especially Kannada script which is one among 12 major scripts in India. The Kannada script consists of vowels and consonants which have different shapes and pattern. The graph based techniques have shown efficient results in other fields so it is decided to graphs for each Kannada character and used as matching parameter while recognizing it.

Keywords— *OCR, Pre-processing, Normalization, Feature Extraction, Classification.*

I. INTRODUCTION

Optical recognition system (OCR) scans the input printed Kannada character and recognize the character present in the database model to form a separate text, which can be edited or processed. There are many OCR systems available for handling the printed English document with reasonable levels of accuracy. Such systems are also available for many Europe languages as well as some of the Asian languages such as Japanese, Chinese, etc. However there are not many reported efforts at developing OCR system for Indian language especially for south Indian languages like Kannada. Most of the work have been carried out the OCR for Devangiri, Bangla and Telugu scripts and not many works are reported for Kannada languages.Kannada is the official language of the South Indian state of Karnataka. It has its own script derived from Bramhi script. Modern Kannada alphabet has a base set of 52 characters, comprising 16 vowels (called as Swaragalu) and 36 consonants (Vyanjanagalu).

There are 2 more consonants used in old Kannada, namely "lla" and "rra", taking the total number of consonants to 38. Further, there are consonant modifiers (vattaksharas or conjuncts) and vowel modifiers. The number of these modifiers is the same as that of base characters, namely 52. Compound characters called Aksharas are formed by graphically combining the symbols corresponding to consonants, consonant modifiers or vowel modifiers using well defined rules of combination. The script has its own numerals too. In addition to the base set of characters and numerals, the script includes special symbols used in poetry, Shlokas (prayer chants), Kannada grammar, etc. Thus, the number of possible consonant-vowel combination Aksharas is 38 X 16 = 608. Similarly the number of possible consonant-consonant-vowel Aksharas is 38 X 38 X 16 = 23104. While designing a character recognition system, if we consider each Akshara as a separate class, the number of classes becomes prohibitively high. However, in Kannada, consonant-modifiers and some of the vowel modifiers are mostly printed separately from the base character. So, if we treat each connected component as a different class, the number of classes in recognition can be reduced by a great extent.

There are lot of challenging issues present in recognizing the Kannada basic characters. As a consequence the methodology for Kannada OCR should be font and size independent. It must also be scalable for including variety of fonts for training with little effort. Hence a large amount of research is going on for the development of an efficient and robust OCR system for different languages and scripts containing diverse font styles and sizes. In order to overcome the above mentioned complexities, many methods have been proposed for OCR of Indian scripts like Bangla, Devanagiri, Telugu and Tamil. & Kannada. In many of the existing system, moment based features are widely used for character recognition. Classical moment invariants were introduced by Hu (1962) which is invariant under translation, rotation and scaling. Neural networks have fast training/learning rate because of their local-tuned neurons (Moody & Darken 1989). They also exhibit universal approximation property and have good generalization ability (Park & Sandberg 1991).

II. LITERATURE REVIEW

The basic Kannada characters such as consonants and vowels are identified by the optical character recognition (OCR) [1], [6], [11] system, which also handles several fonts, sizes and styles of printed Kannada characters. The system, first pre-processes [3], [5] the input document containing the complex Kannada characters and converts it into binary form. Then the system extracts the lines from the document image and segments the lines, Morphological Operations [12] and Projection Profiles based Segmentation [8] of handwritten Kannada Document into character and sub-character level pieces. Here histogram technique and connected component

method can be used for character segmentation and correlation method is used to recognize the characters. The complex Kannada characters are the combination of vowel or consonants or "Vathu" or vowel modifier. The system first segments the document into character level pieces [13] and it is compared with the sample characters. It recognized more than hundred Complex Kannada character and more than hundred complex Kannada words. In order to extract the features from printed Kannada characters, Hu's invariant moments and Zernike moments [9] can be used in pattern identification. Identification of Kannada characters can be done by using support vector machine (SVM) [2], [7]. Neural networks have to be trained with many handwritten samples. Identification rate can be enhanced by training the neural networks. An attempt [4] can be made to develop an algorithm for recognition of machine printed, remote Kannada vowels of different font, size and style using fast Discrete Curvelet Transform (DCT).

The coefficient will be obtained from the DCT. Based on the obtained coefficients we can process and apply standard deviation to obtain feature vector. K-NN classifier can be used for the classification. Along with K-NN; delta, reverse nearest neighbor (RNN) and adoptive neighbor [15] can be used for clustering and classification process. On the other side, for feature extraction and graph matching methods for analyzing the Pattern toning and ANN classifier [10], [14] can be considered. The graph matching algorithm uses less number of positional features i.e. character mean, branch points and end points. The results show that the graph matching and the feature extraction are suitable for the typewritten fonts as the fonts are fixed and less in number. It is a universal approach as it recognizes the typed characters and handwritten characters with higher accuracy.

A. Related Issues and Challenges

- Difficulties faced in viewing angles, shadows and unique Fonts.
- More Curves in the Kannada numeral
- Similar shaped numerals.
- Some characters have very similar variation between them and this leads to recognition complexity and reduces the accuracy rate of the recognition system.

III. PROBLEM DEFINITION

Whenever Kannada characters are scanned, usually the scanning system creates picture record or a picture of the whole page. The computer cannot understand the letters on the page, so it cannot search for words or edit it or change the font, as in a word processor. It would use OCR software to convert it into a text or word processor file so that it could do those things. The result is much more flexible and compact than the original page photo of Kannada text. The need for OCR arises in the context of digitizing the documents from the library, which helps in sharing the data through the Internet. The objective is to design and develop the printed Kannada character recognition system, which takes the printed Kannada character as an input image and it is preprocessed and applies neural network classifier to recognizing the printed Kannada character.

A. Proposed Model

The proposed model for this work is shown in fig 3.1. It consists of two phases, training phase and testing phase. It involves some steps as discussed in further section.

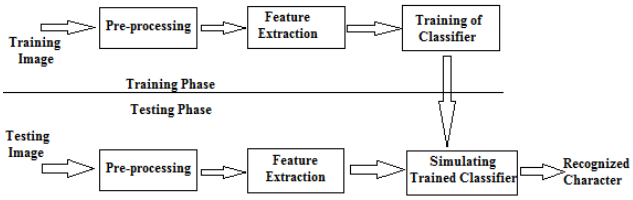


Fig 3.1: Proposed Model

The main steps involved in achieving OCR are as follows:

- Preprocessing
- Feature Extraction
- Classifiers

A. Pre-processing

The input to the system is a digital image containing printed complex Kannada text captured by scanning the input character using a flatbed scanner or digital camera. The input is in RBG format. The preprocessing techniques normally used in any text image processing are shown in fig 3.1.1.

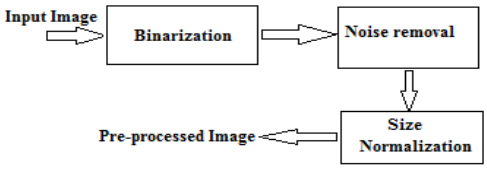


Fig 3.1.1: Image pre-processing

The components of preprocessing are discussed below.

- *Binarization*

Binarization is a method of transforming an image into a black and white image through thresholding. This technique suppress background from the image.

- *Noise removal*

Digital images are having tendency to many types of noises. Noise in a document image is due to poorly photocopied pages. Some noise removal techniques will be used.

- *Normalization*

Normalization is the process of converting random sized images into a standard size. The Bicubic interpolation linear sized normalization techniques could be used for standard sized images.

B. Feature Extraction

Feature extraction is a type of dimensionality reduction technique that efficiently represents interesting parts of an image as a compact feature vector. This approach is useful when image sizes are large and a reduced feature representation is required to quickly complete tasks such as image matching, retrieval and classification applications. In this stage, the features of the characters that are crucial for classifying them at recognition stage are extracted. This is an important stage as its effective functioning improves the

recognition rate and reduces the misclassification. Graph based technique will be used to extract the features in this proposed work.

C. Features Employed

The heart of recognition for any character image is the extracting the features of that image. The main goal of feature extraction technique is to accurately regain the graph based features. The term "Feature Extraction" can be considered to encompass a very wide range of techniques and processes, ranging from simple ordinal / interval measurements derived from individual bands to generate, update and maintaining the discrete feature objects Features extracted from the printed Kannada character image are the global features as follows.

- *Global Features*

The graph based global features extracted in this work are as follows.

Nodes: Number of nodes in a graph of printed Kannada character image.

Lines: Number of lines between the nodes in a graph.

Extreme points: x and y-co-ordinate values of 8 extreme points.

Centroid_x and Centroid_y: calculated the Center location or centroid points (pixel value) of all regions in an image using density equation. Equation shown in Eq(1).

$$xc = 1/m \sum_{i=1}^{m} xi \,.\, mi \qquad \text{Eq.(1)}$$

where M is the sum of intense m_i, m_i is the pixel intense value, x_i and y_i are pixel location on the image, n is the total number of pixels.

Xmax, Xmin: these are minimum and maximum x-co-ordinate values among all the branch points.

Ymax, Ymin: these are minimum and maximum y-co-ordinate values among all the branch points.

D. Classification

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. The classifiers are categorized as statistical methods, such as artificial neural networks, support vector machines and multiple classifier combination. Also some distance based classifiers are available such as Euclidian distance, Manhattan distance city block distance.

IV. GRAPH BASED TECHNIQUE

Graph is a very well-known representation method in the data structure. A graph has main two components Vertex and edges. After formation of graph representation for printed Kannada characters the similarities between the two graphs can be extracted. With the acquired data we can construct mathematical graph of data. In the case of proposed method, undirected weighted graphs will be used. An undirected mathematical graph G is an ordered pair (V, E) in which V represents a set of vertices (nodes) and E represents set of edges. In other words, mathematical graph is a set of vertices that are connected by links called edges. Every edge connects only two vertices, and every two vertices can be called adjacent only if they are connected with an edge. To an every edge E we can assign some non-negative number "w" which is then referred as the weight of the edge E. If all the edges of the graph G have weight assigned to them then the graph G is called weighted graph.

The first feature that we need to extract from the graph is the information whether the graph is connected or not. Graph is connected if we choose one vertex and traveling along the edges of the graph manage to reach all other vertices of the same graph. This process can be achieved by implementing light versions of Depth-First Search or Breadth-First Search algorithms which will tell us if they have searched through all the graph vertices. This will be the main factor for the best match scoring in the identification process.

The connectivity of a given graph will be very important in determining next graph features. We have to identify Eulerian graph or has the graph Euler path (Eulerian trail). Rare mathematical graphs are Eulerian graphs, thus if someone's signature produces that kind of graph it would certainly be of the great influence to the identification process. Graph is Eulerian if and only if all the vertices have an even degree (degree of the vertex is the number of edges connected with the vertex). Eulerian trail is a trail in a graph which visits every edge exactly once. Graph will have Eulerian trail if and only if it has at most two vertices with an odd degree. All graph features described so far can be implemented on both non-weighted and weighted graphs. The graph can be generated for all the printed Kannada character as shown in the figure 4.1.

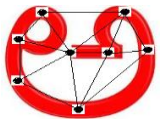


Figure 4.1: Graph for printed Kannada character.

By graph shown above, the features such as number of nodes, edges, Eigen feature etc. can be extracted & given to the classification. By applying the graph based techniques on the printed Kannada character image 23 features have been extracted which are listed in the table 4.1. All the features extracted from the graph based techniques are used as inputs to the BPNN classifier to recognize a given character.

1	E1	10	E10	19	L2
2	E2	11	E11	20	T1
3	E3	12	E12	21	T2
4	E4	13	E13	22	NP
5	E5	14	E14	23	N1
6	E6	15	E15		
7	E7	16	E16		
8	E8	17	C		
9	E9	18	L1		

Table 4.1: List of Extracted Features

Where,
E1-E16: Eigen features of extreme points around character image.
C : Centroid pixel value of the image.
L1, L2: These are minimum and maximum x-co-ordinate values among all the branchpoints.
T1, T2: These are minimum and maximum y-co-ordinate values among all the branchpoints.
Np, Nl : number of points and the number of lines drown on an image.

V. TRAINING & TESTING OF BPNN

A. Training of BPNN

The various printed Kannada characters are trained using BPNN. To complete the proper training, huge number of printed Kannada character images are required. Extracted features of printed Kannada characters are utilized after normalization to train the developed model. Initial runs showed that these settings are sufficient for this study. The fulfillment of classifier after training with back propagation algorithm. The most acceptable mean square set at 1×10^{-3} during the training action. The training courses are carried out with fast back propagation with 500-5000 epochs.

B. Testing of BPNN

Image from the testing file is taken and given them to the training model. The trained BPNN model classifies the given input printed Kannada character image and produces the output of a recognized type as shown in the below table 5.1.

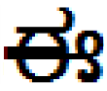
SI. No	Printed Kannada Character	Corresponding Pattern	Recognized Character
1.		0 0 0 0 0 1	A
2.		0 0 0 0 1 0	Aa
3.		0 0 0 0 1 1	I
4.		0 0 0 1 0 0	Ie

Table 5.1: System Level Character Recognition

In this aspect, various printed Kannada character which is not utilized in the training sets, are used to calculate the accuracy of perception. The process of recognition is repeated for various images which admit the trained and untrained images. Classification accuracy is calculated as.

Classification accuracy =

No. of recognized printed Kannada character /
Total no. of testing printed Kannada character

VI. EXPERIMENTATION

In this proposed methodology the graph based feature are used for the printed Kannada character recognition to achieve the greater rate of recognition. The graph based features used in the recognition of the printed Kannada character are intersection points, end points, extreme points, centroid, and minimum and maximum x-co-ordinate values among all the branchpoints. And minimum and maximum y-co-ordinate values among all the branchpoints. During the training stage only 49X8=392 images have been used. In this chapter to access the performance of the proposed method, totally 539 (49X11) printed character images with varying sizes have been tested. Among 539, 510 characters are correctly recognized.

VII. CONCLUSION

The proposed system describes a simple and efficient OCR system for printed Kannada character, a South Indian language. It takes complex Kannada character as input image and it recognizes. The system can be designed to be independent of the font and size of a character. The system uses graph based features. Graph based methods are used for the face recognition, hand recognition etc. But not for the Kannada character. In this proposed the graph based features such as nodes, edges and Eigen features are used for recognizing the printed Kannada character.

REFERENCES

[1] F. W. Mattern and J. Denzler , "Comparison of appearance based methods for generic object recognition," *Pattern Recognition and Image Analysis*, Vol. 14, No. 2, pp. 255–261,2004.
[2] R Sanjeev Kunte and R D Sudhaker Samuel, "A simpleand efficient optical character recognition system for basic symbols in printed Kannada text," Vol. 32, pp. 521–533, 2007.
[3] Thungamani.M, Dr Ramakhanth Kumar P, Keshava Prasanna and Shravani Krishna Rau, "Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments," Vol. 11 No.7, pp. July 2011
[4] Mamatha H.R, Sucharitha S and Srikanta Murthy K, "Multi-font and Multi-size Kannada Character Recognition based on the Curvelets and Standard Deviation," Vol 35– No.11, December 2011.
[5] Vishweshwarayya C. Hallur, Avinash A. Malawade, Seema G. Itagi, "Survey on Kannada Digits Recognition Using OCR Technique," Vol 1, December 2012.
[6] G. G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, "Printed and Handwritten Mixed

Kannada Numerals Recognition Using SVM," Vol. 02, No. pp. 05, 1622-1626, 2010.

[7] Mamatha H R and Srikantamurthy K, "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document," Vol 4– No.5, October 2012.

[8] Niranjan S.K, Vijaya Kumar, Hemantha Kumar G, and Manjunath Aradhya V N, **"**FLD based Unconstrained Handwritten Kannada Character Recognition," *Theory and Application* Vol. 2, No. 3, September 2009.

[9] Anjali Chandavale, Suruchi Dedgaonkar, Dr. Ashok Sapkal, "An Approach for Character Recognition Using Pattern Matching with ANN," Vol 3, ISSUE 10, ISSN 2229-5518, OCTOBER-2012.

[10] Swapnil A. Vaidya, Balaji R. Bombade, "A Comprehensive Survey on Kannada Numerals and Character Recognition**,**" Vol 3, Issue 3, March 2013.

[11] Manjunath A E , Sharath B, "Implementing Kannada Optical Character Recognition on the Android Operating System for Kannada Sign Boards**,**" Vol. 2, Issue 1, January 2013.

[12] Mr.Nithya.E and Dr. Ramesh Babu D R" OCR System for Complex Printed Kannada Characters".

[13] K.S. Prasanna Kumar of the paper, "Optical Character Recognition (OCR) for Kannada numerals using Left Bottom 1/4th segment minimum features extraction".

[14] Umesh R S, Peeta Basa Pati and A G Ramakrishnan, "Set theoretic line segmentation and graph based strategy for bilingual Kannada-English OCR".