International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

OCR for Handwritten Kannada Language Script

Netravati Belagali[1], Shanmukhappa A. Angadi[2]
*[1]Department of Computer Science and Engineering,Visvesvaraya Technological University,
PG Centre, Belagavi, Karnataka, India*
*[2]Department of Computer Science and Engineering,Visvesvaraya Technological University,
PG Centre, Belagavi, Karnataka, India,*

Abstract— The optical character recognition (OCR) is the process of converting textual scanned image into a computer editable format. The proposed OCR system is for complex handwritten Kannada characters. One of the major challenges faced by Kannada OCR system is recognition of handwritten text from an image. The input text image is subjected to preprocessing and then converted into binary image. Segmentation process is carried to extract single character from image. This can be done using connected component labeling. Hu's invariant moments, horizontal and vertical profile features are obtained as features from zoned image. Probabilistic neural network (PNN) classifier is used for character recognition. Finally the recognized output is editable in baraha editor. An accuracy of 94.69% is achieved in character recognition of the domain specific input.

Keywords— Optical Character Recognition (OCR), Feature Extraction, Neural Network, Kannada Scripts

I. INTRODUCTION

An OCR engine was developed between 1985 and 1994 by Hewlett Packard. Optical character recognition is one of the technologies which recognizes the text from the document images and converts the images into a form that the computer can manipulate. OCR is one of the sub fields and research area in pattern recognition. When handwritten text paper is scanned, it produces the document image file. The machine cannot understand the characters on the page, so user cannot edit it or search for words or change fonts as in the word processor. If OCR tool is used to convert the document image into word processor file so that can possible to do all those changes. Scan document with high resolution scanner then OCR tool will produce best result.

Many European languages and some of the Asian languages such as Japanese and Chinese adopted an OCR system with good accuracy level. However less effort has been reported in the OCR system for Indian languages, particularly for south Indian languages like Kannada.

All the OCR tools normally process with five steps, those are preprocessing, segmentation, feature extraction, recognition/ classification and last one is post processing which deals with grammar and spell checking of the documents. Because of the prevalence of internet and multimedia technique OCR has been increasingly developed in recent years. Some of the applications of optical character recognition system (OCR) include digitizing library documents, which can use to share data through the Internet, Form and bank check processing and so on.

Recently Nithya.E [1] developed OCR system for printed complex Kannada characters and correlation method is used for character recognition. B.M. Sagar. [2] used database approach to develop OCR for printed Kannada text and achieved 100%accuracy. Sunanda Dixit [3] proposed method to find error during segmentation of handwritten Kannada text and also discussed weighted bucket algorithm. J.Pradeep.[4] proposed diagonal based feature extraction technique for extracting features of handwritten alphabets . M.S. Patel. [5] Proposed grid based method for recognition of handwritten Kannada text and also explained Euclidean distance method used for classification.

The proposed system extracts 27 features using which the classifier training and testing is done. The text images are scanned using good resolution scanners. Then this text image undergoes into various steps like pre-processing, segmentation, feature extraction and classification. The image containing handwritten Kannada text is taken as input to the system, then image undergoes preprocessing step where it removes background noise and converts it into binary image. Segmentation is carried out to extract lines, words and characters from the input text image. Here connected component labeling is used to extract single character from an image. Zoning method is used to divide the single character image horizontally into three parts. Next some of features like hu's invariant moments, horizontal and vertical profile moments are extracted from each zoned images. Then the neural network classifier is used to classify the segmented characters. Classification is done based on various features which have been extracted from the zoned image. The work in this paper considers 10 types of Kannada handwritten sample images.

The remaining part of the paper is organized into four parts: the overview of handwritten character recognition, this section explains the problems faced during handwritten character recognition process. Proposed system section gives explanation on proposed architecture diagram. Results and discussions section contains the snapshots of proposed system and finally conclusion section deals with conclusion of proposed system along with future work to be carried out. The four parts are introduced as follows.

II. HANDWRITTEN CHARACTER RECOGNITION

Handwritten character recognition is an optimal character recognition problem for handwritten characters. Character recognition helps to improve the interface between man and machine in numerous applications.

One of the major challenges faced by the Kannada OCR system is distinction of similar shaped component in the Kannada script during the character recognition process. Compared to printed Kannada text recognition, handwritten Kannada text recognition become a complex task as different persons have different handwritten styles which diversified font style and font size. One more reason for complexity of handwritten character recognition of Indian language is the presence of vowel modifiers and compound characters in Indian languages.

Some of the character recognition approaches for Indian languages are statistical techniques, structural or syntactic etc. template matching, hybrid or combination approach and neural network approaches. Next section explains the architecture of the proposed system along with methods used to implement it.

III. PROPOSED SYSTEM

In handwritten optical character recognition (OCR) system much work is reported on English and other European languages compared into south Indian languages like Kannada. Kannada handwritten recognition become complex task as it contains diversification in handwriting style and vowel modifier. Different OCR tools have to develop for both handwritten and printed documents. The proposed system helps to reduce storage work of old handwritten documents as it converts text image into electronic records. Nowadays more usage of multimedia technique and internet OCR become one of the raising techniques in an image processing.

The System architecture is given in Figure 1. It shows that the system is subdivided into various modules, like pre-processing, segmentation, feature extraction, Classification/recognition and post-processing. Classification is done using probabilistic neural network. Kannada handwritten text image is given as the input to system; this undergoes pre-processing of image like background elimination, noise removal. Background removal is done using thresholding. Then the image is segmented to extract lines, words and characters. Character segmentation is done using connected components labeling. Then each character is extracted into sub images. Those single character images are divided into three horizontal zones. Feature extraction is done on three zones from each character image. Hu's invariant moments, Horizontal and Vertical profile features are extracted.

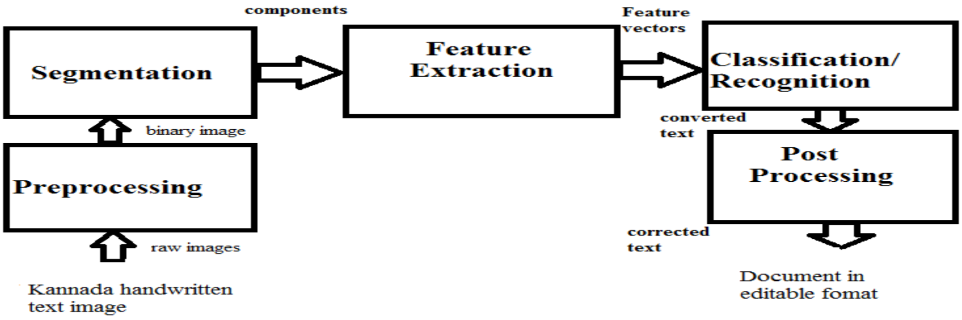


Figure 1: System Architecture

Next classifier is trained. Classification is done on the basis of hybrid features. Probabilistic neural network classifier is used. First the classifier is trained to develop a pattern dictionary, which helps further in classification. Once the pattern dictionary is created the images are tested to check if the given output is correct or not and performance of the system is calculated.

Feature extraction is an important part in handwritten character recognition. The classification depends on the features extracted from the image once the features are extracted. The classifier is trained using these features. The following describes the features extracted and the neural network used.

A. Feature extraction

Zoning

Zoning is one of the feature extraction methods. Here it is used to divide the extracted single character image into three horizontal zones. These zoned images are used during the feature extraction process. Features extraction is applied on individual zones rather than whole image. The figure 2 shows the horizontal zoning on the single character image.

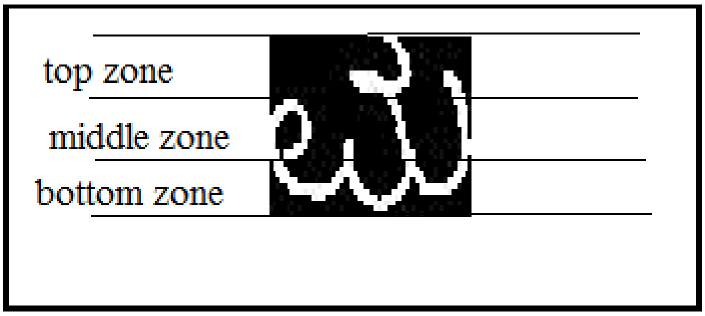
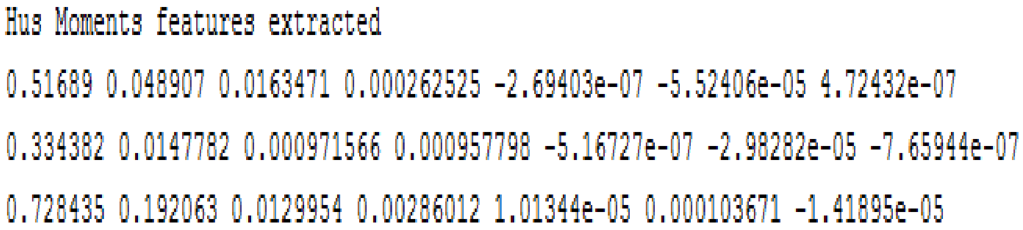


Figure 2: zoned image

Hu's Invariant moments

From each zoned image 7 invariant moments are extracted so totally 21 features have been extracted from single character image.

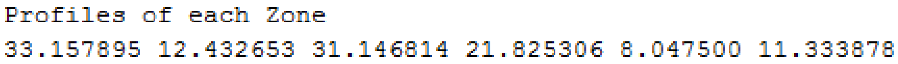


```
Hus Moments features extracted

0.51689 0.048907 0.0163471 0.000262525 -2.69403e-07 -5.52406e-05 4.72432e-07

0.334382 0.0147782 0.000971566 0.000957798 -5.16727e-07 -2.98282e-05 -7.65944e-07

0.728435 0.192063 0.0129954 0.00286012 1.01344e-05 0.000103671 -1.41895e-05
```

Figure 3: Hu's moment features of Kannada letter 'sa'

Horizontal and vertical profile features

Horizontal and vertical profile features are extracted from each character image. Totally 6 features are been extracted from single image.



```
Profiles of each Zone
33.157895 12.432653 31.146814 21.825306 8.047500 11.333878
```

Figure 4: horizontal and vertical features of Kannada letter 'sa'

B. Classifier (Probabilistic neural network)

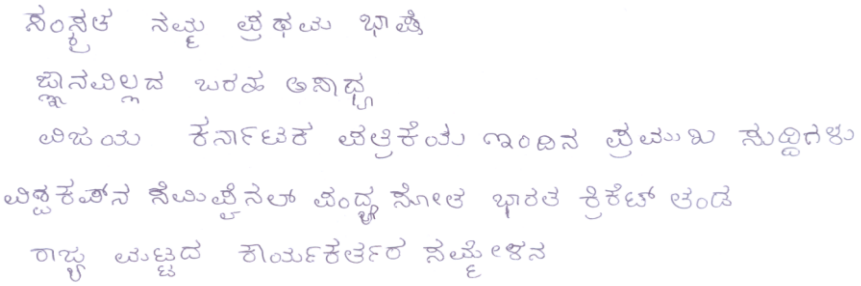
This phase uses probabilistic neural network (PNN) classifier for testing and training of images. PNN is feed forward neural network and it was introduced by D F Specht in the early 1990's. PNN is multi layered network. It consists of four layers, those are input layer, pattern layer, hidden layer and output layer. In a input layer every neuron is used to represent predictor variables. In pattern layer every training data set contains single neuron. Output layer uses the large number of votes to guess the target category.

Once feature extraction is completed and these features are stored in file then next phase is to implement classifier. The file written during feature extraction stage in opened in training process to give input to the neural network. In proposed method 10 sample images of different Kannada handwritten styles are used for an experiment purpose. A total of 980 images are used to train the network using PNN. For all these images, features are extracted for training and knowledge base is created.

IV. RESULTS AND DISCUSSIONS

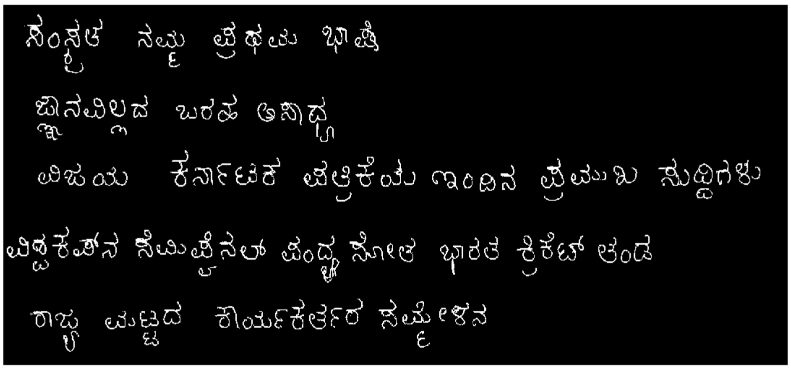
The system is designed and implemented using the Matlab. In order to evaluate the accuracy of the system many images are tested. The below figures shows the snapshots of the proposed system.

Input image



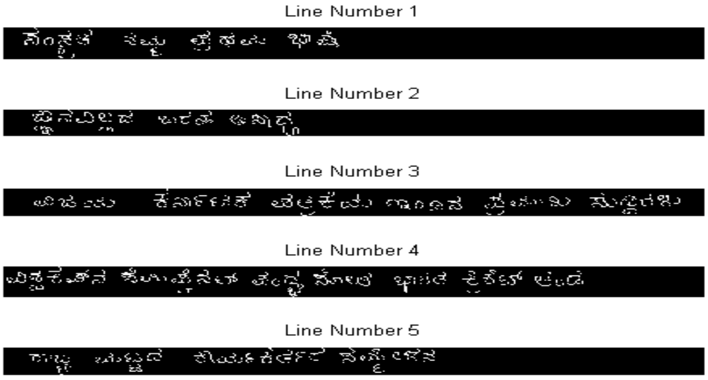
Snapshot 1: input image

Binary image

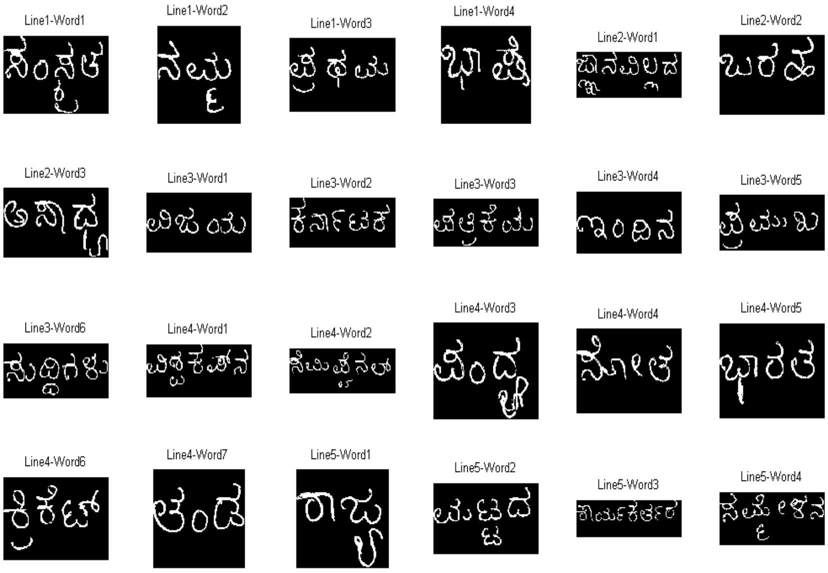


Snapshot 2: binary image

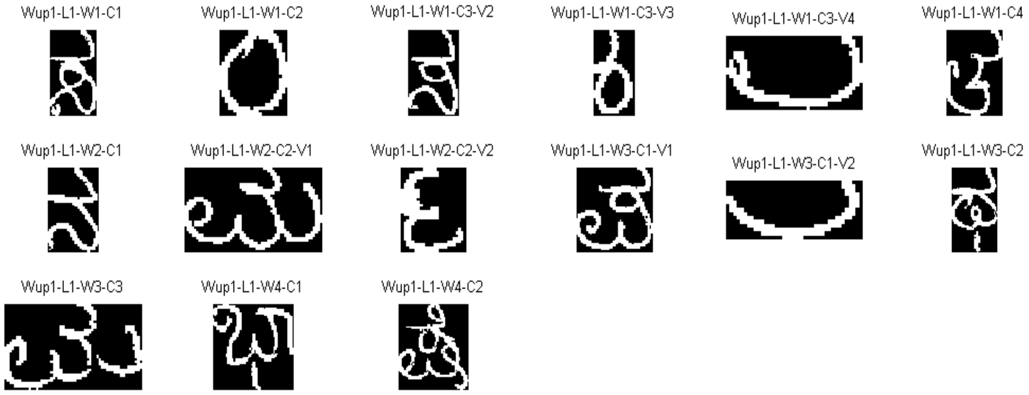
Segmentation



Snapshot 3: line segmentation

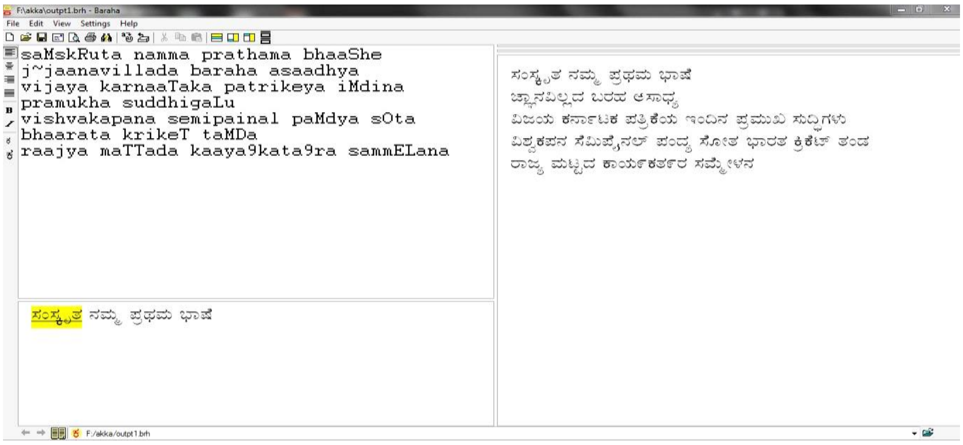


Snapshot 4: word segmentation



Snapshot 6: character segmentation

Output image



```
saMskRuta namma prathama bhaaShe
j~jaanavillada baraha asaadhya
vijaya karnaaTaka patrikeya iMdina
pramukha suddhigaLu
vishvakapana semipainal paMdya sOta
bhaarata krikeT taMDa
raajya maTTada kaaya9kata9ra sammELana
```

ಸಂಸ್ಕೃತ ನಮ್ಮ ಪ್ರಥಮ ಭಾಷೆ
ಜ್ಞಾನವಿಲ್ಲದ ಬರಹ ಅಸಾಧ್ಯ
ವಿಜಯ ಕರ್ನಾಟಕ ಪತ್ರಿಕೆಯ ಇಂದಿನ ಪ್ರಮುಖ ಸುದ್ಧಿಗಳು
ವಿಶ್ವಕಪನ ಸೆಮಿಪೈನಲ್ ಪಂದ್ಯ ಸೋತ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡ
ರಾಜ್ಯ ಮಟ್ಟದ ಕಾರ್ಯಕರ್ತರ ಸಮ್ಮೇಳನ

ಸಂಸ್ಕೃತ ನಮ್ಮ ಪ್ರಥಮ ಭಾಷೆ

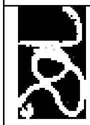
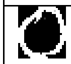
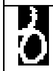
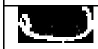
F:/akka/outpt1.brh

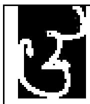
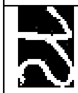
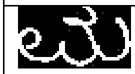
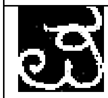
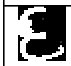
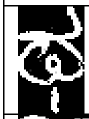
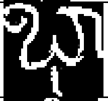
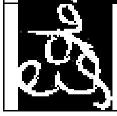
Snapshot 7: output image

Performance Analysis

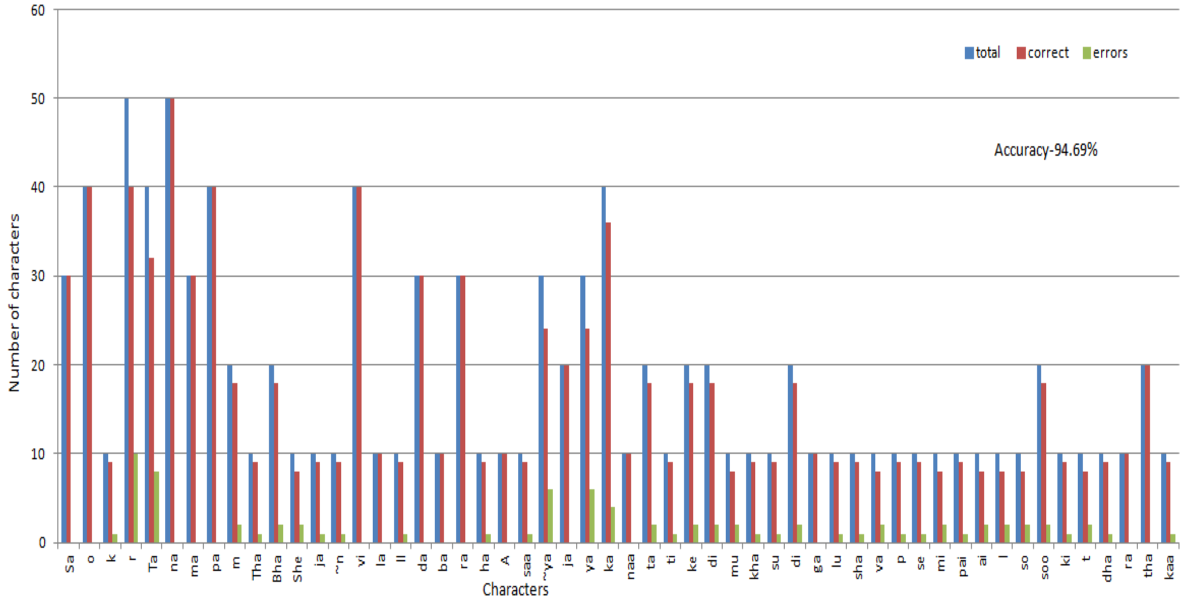
Following table displays performance analysis of some of Kannada characters.

Table 1: Performance analysis table

character image	number of samples tested	number of samples correctly recognized	number of samples miss classified	% of recognition accuracy
	30	30	0	100
	40	40	0	100
	10	9	1	90
	50	40	10	80

	40	32	8	80
	50	50	0	100
	30	30	0	100
	40	40	0	100
	20	18	2	90
	10	9	1	90
	20	18	2	90
	10	8	2	80

For proposed system implementation considered 10 sample images of different handwritten styles. In that every image contains 23 words and 105 characters. So totally 980 characters are tested in that 928 characters recognized correctly. Accuracy level reached 94.69% for this system.
Accuracy= 928/980*100=94.69%



Snapshot 8: Performance analysis graph

V. CONCLUSION

In this proposed system, a zoning based invariant moment feature extraction based PNN classifier is used for recognizing handwritten Kannada characters. Along with invariant moments, horizontal and vertical profile features are extracted from a character image for a better recognition result.

The experimental results and performance analysis table shows the performance of proposed system. Neatly written Kannada characters are considered for the purpose of an experiment.

In future have to work on different styles of Kannada handwritten text and have to improve character recognition accuracy using different methods.

REFERENCES

1. Nithya.E and Ramesh Babu D R, "OCR system for complex printed Kannada characters", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.
2. B.M. Sagar, Shobha G and Ramakantha Kumar P, "OCR for printed Kannada text to machine editable format using database approach", Issue 6, Volume 7, June 2008.
3. Sunanda Dixit, S Ranjitha, H N Suresh, "Segmentation of handwritten kannada text document through computation of standard error and weighted bucket algorithm", International journal of advanced computer technology | volume 3, number 2.
4. 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.
5. M.S. Patel, Sanjay Linga Reddy," An Impact of Grid based Approach in Offline Handwritten Kannada Word Recognition", International Conference on Contemporary Computing and Informatics, 2014.
6. S.A.Angadi, Sharanabasavaraj.H.Angadi, "structural features for recognition of hand written kannada characterbased on svm", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 5,No.2, April 2015.
7. B.V.Dhandra, Shashikala, Gururaj Mukarambi, "Kannada Handwritten Vowels Recognition based on Normalized Chain Code and Wavelet Filters", International Journal of Computer Applications Recent Advances in Information Technology, 2014.
8. Saleem Pasha , M.C. Padma, "Recognition of handwritten kannada characters using hybrid features".
9. B.V.Dhandra, Shashikala, Gururaj Mukarambi, "Kannada Handwritten Vowels Recognition based on Normalized Chain Code and Wavelet Filters", International Journal of Computer Applications Recent Advances in Information Technology, 2014.
10. Samit Kumar Pradhan, Sujoy Sarkar, Suresh Kumar Das, "A Character Recognition Approach using Freeman Chain Code and Approximate String Matching", International Journal of Computer Applications (0975 – 8887) Volume 84, December 2013.