Global Journal of Computer Science and Technology: F
Graphics & Vision
Volume 14 Issue 4 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Zone-Features' based Nearest Neighbor Classification of Images of Kannada Printed and Handwritten Vowel and Consonant Primitives

By Basavaraj. S. Anami, Deepa S Garag & P. Shivakumara

K.L.E. Institute of Technology, India

Abstract- The characters of any languages having scripts are formed by basic units called primitives. It is necessary to practice writing the primitives and their appropriate combinations while writing different characters. In order to automate character generation, primitives" recognition becomes important. In this paper, we propose a zone-features based nearest neighbor classification of Kannada printed and handwritten vowel and consonant primitives. The normalized character image is divided into 49 zones, each of size 4x4 pixels. The classifier based on nearest neighbor using Euclidean distances is deployed. Experiments are performed on images of printed and handwritten primitives of Kannada vowels and consonants. We have considered 9120 images of printed and 3800 images of handwritten 38 primitives. A K-fold cross validation method is used for computation of results. We have observed average recognition accuracies are in the range [90%, 93%] and [93% to 94%] for printed and handwritten primitives respectively. The work is useful in multimedia teaching, animation; Robot based assistance in handwriting, etc.

Keywords: classification, feature extraction, K_Fold cross validation, majority voting, nearest neighbor, printed primitives, handwritten primitives.

GJCST-F Classification : I.4.0

ZONE-FEATURESBASEDNEARESTNEIGHBORCLASSIFICATIONOFIMAGESOFKANNADAPRINTEDANDHANDWRITTENVOWELANDCONSONANTPRIMITIVES

Strictly as per the compliance and regulations of:

Zone-Features' based Nearest Neighbor Classification of Images of Kannada Printed and Handwritten Vowel and Consonant Primitives

Basavaraj. S. Anami [α], Deepa S Garag [σ] & P. Shivakumara [ρ]

Abstract - The characters of any languages having scripts are formed by basic units called primitives. It is necessary to practice writing the primitives and their appropriate combinations while writing different characters. In order to automate character generation, primitives" recognition becomes important. In this paper, we propose a zone-features based nearest neighbor classification of Kannada printed and handwritten vowel and consonant primitives. The normalized character image is divided into 49 zones, each of size 4x4 pixels. The classifier based on nearest neighbor using Euclidean distances is deployed. Experiments are performed on images of printed and handwritten primitives of Kannada vowels and consonants. We have considered 9120 images of printed and 3800 images of handwritten 38 primitives. A K-fold cross validation method is used for computation of results. We have observed average recognition accuracies are in the range [90%, 93%] and [93% to 94%] for printed and handwritten primitives respectively. The work is useful in multimedia teaching, animation; Robot based assistance in handwriting, etc.

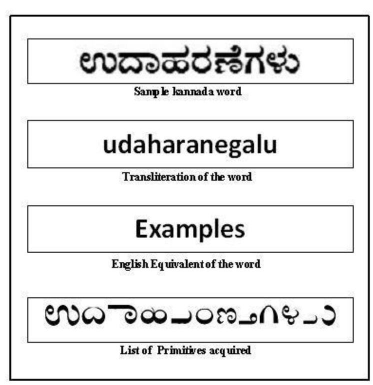
Keywords: classification, feature extraction, K_Fold *cross validation, majority voting, nearest neighbor, printed primitives, handwritten primitives.*

I. Introduction

Natural language processing (NLP) is a field of computer science that deals with understanding and generation of natural languages. Natural language understanding enables computers to understand the natural language and extract meaning from it. Natural language generation involves both spoken and written information. Some applications of both text to speech and speech to text conversion are man-machine interfaces to computers, systems that read and understand printed and hand written text, speed understanding system, text analysis and understanding systems,computer aided instruction systems etc. Ediphones and Dictaphones are examples of speech to text conversion systems.

India is a multi lingual and multi script country with 22 scheduled languages. Every state has its language andcertain dialects. To name a few, Karnataka state has Kannada, Maharashtra has Marathi, Uttarpradesh has Hindi and Gujarath has Gujarathi. These languages are written using only twelve scripts. Devnagiri script used to write Hindi, Marathi, and Sanskrit. All other languages have their own script. Most of the Indian languages are derived from Ancient Brahmi script, Kannada is one amongst them and are called phonetic languages. Therefore, writing these scripts map the sounds of alphabet to specific shapes. The text in all these languages, except Urdu, is written from left to right. Kannada is a scheduled language in India and is spoken by about 60 million people of the south Indian state of Karnataka. The Kannada alphabets are developed from the Kadamba and Chalaukya scripts, descendents of Brahmi script, which are used between the 5th and 7th centuries A.D. The basic structure of Kannada script is distinctly different from the Roman script. Kannada script evolved primarily from stone carving; hence most of the characters are round, curvy and symmetric with straight strokes/wedges. All these characters in a word are isolated.

Kannada script has compound characters and hence different from English. In addition to this, upper and lower case characters are absent in Kannada script. The typical Kannada text looks like as given in Box1. The content of thebox 1 is read as "udaharanegalu" and its English equivalent is "Examples".



Box 1 : Sample Kannada word and its primitives

Author α, σ : K.L.E. Institute of Technology, Hubli-580030, Karnataka .
Author ρ: University of Malaya (UM), Kuala Lumpur-50603, Malaysia .
e-mails: anami_basu@hotmail.com, deepagarag@rediffmail.com, shiva@um.edu.my.

The alphabet of a language is divided into basic characters called vowels and consonants. Two or more basic characters are combined to form compound characters. We recall here the way we have learnt the alphabet and started writing the text in a language, in kindergarten schools, children are made to practice writing the characters and thereby memorize them. The character writing involves combining the primitives. Every character written from an alphabet follows a definite way and depends upon the type of the writer, whether left-hand-writer or right-handwriter. The character generation at once is different from how it is written. Each character in the alphabet has definite way of writing it and is combination of sub parts called primitives. This combination of primitives is a systematic approach in generating or building the characters. The character construction is basic to any medium of learning. While reading, we read the whole character and while writing or constructing a character, we write primitives in an order.

The automation of character construction requires the recognition of primitives from the database of primitives in a language. Many researchers have worked on character recognition, wherein the whole character is considered as one single unit. The focus of the present work is to recognize the images of primitives of Kannada language useful in the construction of characters using syntactic approach. The work is useful for novice learners, multimedia applications, translate-ration and translation etc. The automated script writing and learning by taking technological leverage is considered a new area of research.

In this paper, we have considered the different font types and font sizes of vowels and consonants characters supported by Kannada language software, namely, Nudi and Baraha. We have identified with the help of language experts the primitives of vowels and consonants and manually separated and their images are stored. These images of primitives are preprocessed through binarization, thinning and resizing. The simple zone based features are obtained for these primitives. Nearest neighbor classification is adopted with Euclidean distance measure for recognition of primitives. We have tested for all the combinations of printed primitives with different fonts‟ types and sizes.

The remaining part of the paper is organized into four sections. Section 2 deals with detailed survey on automatic primitive recognition. Section 3 deals with the proposed methodology, wherein different stages of the methodology are discussed. The experimental results and discussion are given in section 4. Conclusion and Future work are given in Section 5.

II. Literature Survey

To know the state-of-the-art in automatic primitive recognition, we carried out the literature survey and following isthe gist of cited papers.

[Leena R Ragha, et. al, 2010] have investigated the moments features on Kannada handwritten basic character set of 49 letters. Four directional images using Gabor wavelets from the dynamically preprocessed original images are found. Then moments features are extracted from them. The comparison of moments features of 4 directional images with original images when tested on Multi Layer Perceptron with Back Propagation Neural Network shows an average improvement of 13% from 72% to 85%. The mean performance of the system with these two features together obtained is 92%.

[Karthik Sheshadri et.al, 2010] have proposed Kannada Character Recognition method based on k-means clustering. A segmentation technique to deco-mpose each character into components from three base classes is used to reduce the magnitude of the problem. The k-Means clustering technique provides a natural degree of font independence and this is used to reduce the size of the training data set to about one-tenth of those used in related works. Accuracy comparisons with related work, shows that the proposed method yields a better peak accuracy. The relative merits of probabilistic and geometric seeding in k-means are also discussed.

[Leena R Ragha, et. al, 2011] have presented the use of moments features on Kannada Kagunita. Four directional images are found using Gabor wavelets from the dynamically preprocessed original image. The Kagunita set is analysed and the regions with vowel and consonant information are identified and cut from the preprocessed original image to form a set of cut images. Moments and statistical features are extracted from original images, directional images and cut images. These features are used for both vowel and consonant recognition using multi-layer perceptron with backpropagation learning.

[Sangame S.K, et. al, 2011] have presented an unconstrained handwritten Kannada basic character recognition using invariant moments and chain code features. Invariant moments feature are extracted from zoned images and chain code. A Euclidean distance based K-NN classifier is used to classify the handwritten Kannada vowels. The method is invariant to size, slant, orientation and translation.

[B.V.Dhandra et. al, 2011] have proposed zone based features for recognition of the mixer of handwritten and printed Kannada digits. The kNN and SVM are used to classify the mixed handwritten and printed Kannadadigits. The reported recognition rates are 97.32% and 98.30% for mixed handwritten and printed Kannada digits using KNN and SVM classifiers respectively.

[G. G. Rajput et.al, 2011] have discussed the implementation of shape based features, namely, Invariant Fourier descriptors and normalized chain codes for OCR of binary images of Kannada hand-written characters. SVM classifier is used for recognition.

The result is computed using five-fold cross validation. The mean performance of the recognition reported for the two shape based features together is 98.45% and 93.92%, for numeral characters and vowels, respectively. The mean recognition rate of 95% is obtained for both vowels and characters taken together.

[K S Prasanna Kumar et. al, 2012] have presented an algorithm to optical character recognition (OCR) for Kannada numerals. The segmentation of a numeral into four equal parts and one of these parts i.e., left bottom segment, is used to extract recognition features. A conflict resolution algorithm is proposed to resolve the conflicting features. A minimum number of features are extracted so as to improve the response time.

[Umapada Pal et. al, 2012] have given a state-of-the-art survey about the techniques available in the area of offline handwriting recognition (OHR) in Indian regional scripts. Various feature extraction and classification techniques associated with the offline handwriting recognition of the regional scripts are discussed in this survey. A separate section is dedicated to the observations made, future scope, and existing difficulties related to handwriting recognition in Indian regional scripts.

[Kauleshwar Prasad, et.al, 2013] have focused on recognition of English alphabet in a given scanned text document with the help of neural networks. The first step involves image acquisition that involves the sub steps,namely, noise filtering, smoothing and normalization of scanned image. The second step, decomposition, involves segmentation of acquired image into sub images. The features improve recognition rate and misclassification. Character extraction and edge detection algorithms are devised. Neural network is used to classify and recognize the handwritten characters.

[Mamatha Hosalli Ramappa, et.al, 2013] have examined a variety of feature extraction approaches and classification methods for different optical character recognition applications. These approaches are des-igned to recognize handwritten numerals of Kannada script. Eight different features are computed from zonal extraction, image fusion, radon transform, etc and ten different classifiers like Euclidean distance, Chebyshev distance, etc are deployed. Artificial Immune system and classifier fusion are considered.

[Om Prakash Sharma et. al, 2013] have presented recent trends and tools used for feature extraction that helps in efficient classification of the handwritten alphabets. The use of Euler Number in addition to zoning increases the speed and the accuracy of the classifier as it reduces the search space by dividing the character set into three groups.

[G. G. Rajput et.al, 2013] have proposed a zone based method for recognition of handwritten characters in Kannada language. The normalized character image is divided into 64 zones and each is of size 8x8 pixels. For each zone,from left to right and from top to bottom, the crack code, representing the line between the object pixel and the background (the crack), is generated by traversing it in anticlockwise direction. A feature vector of size 512 is obtained for each character. A multi-class SVM is used for the classification purpose. The data set has 24500 images with 500 samples of each character. Five-fold cross validation is used and yielded 87.24% recognition accuracy.

[Swapnil A. Vaidya et. al, 2013] have given an overview of the ongoing research in OCR systems for Kannada scripts. They have provided a starting point for the researchers in the field of OCR. The state-of-the-art OCR techniques used in recognition of Kannada scripts, recognition accuracies and the resources available are discussedin fair detail.

[H. Imran Khan et. al, 2013] have proposed a chain code based feature extraction method for developing HCR system. A eight -neighborhood method is implemented, which allows generation of eight different codes for each character. These codes are used as features of the characters" images. These features are used for training and testing the k-Nearest Neighbor (KNN) classifier.

[Nithya E. et.al, 2013] have proposed an OCR system for complex printed Kannada characters. The input to the system is a scanned image of a page of text containing complex Kannada characters and the output is in a machine editable form. The pre-processing step converts the input document into binary form. The lines from the document image are extracted and further segmented into the lines, characters and sub characters. The histogram and connected component methods are used for segmentation and correlation is used for recognition of characters.

[Mamatha.H.R. et, al, 2013] have attempted to measure the performance of the classifier by testing with twodifferent datasets of different sizes. A framework based on the combined concepts of decision fusion and feature fusion for the isolated handwritten Kannada numerals classification is proposed. The combined approach has increased the recognition accuracy by 13.95%.

[Manjunath A. E., et.al, 2013] have proposed a Kannada OCR (Optical Character Recognition) in hand held devices. The work involves Kohonen"s algorithm. The Kohonen network is trained with initial images. The images are thinned Hilditch algorithm. The distortions present in the images are eliminated and images are converted to grey scale images. The grey scale images are segmented and the result is displayed along with a vocal output.

[Sandhya.N et.al, 2014] have proposed a new classification method for Kannada characters, which is used as a preliminary step for recognition. An analysis of Kannada characters is carried out. The syntactic features are identified. At first, the basic features and their exact positions in the characters are identified and recorded. Further, by using a decision tree, the characters are classified. The experimental results show that the syntactic based method using basic features gives good contribution and reliability for Kannada character classification.

From the literature survey, it is observed that researchers have worked on Kannada character recognition. The feature extraction techniques such as template matching, Zernike moments, geometric moment invariants, directional, positional, Fourier transform, etc are used. The classification techniques such as, neural network, support vector machines, nearest neighbor, etc are used. No specific work is observed on recognition of Kannada language primitives in the light of character construction is cited in the literature [1-18]. This is the motivation for the present work on printed and handwritten Kannada language vowels and consonants primitives‟ recognition.

III. Proposed Methodology

The proposed methodology consists of four major steps, namely, identifying the primitives in printed and handwritten Kannada vowels and consonants and obtaining their images, preprocessing of primitives‟ images, primitives‟ recognition and classification of primitives. The steps are shown in the Figure 1.

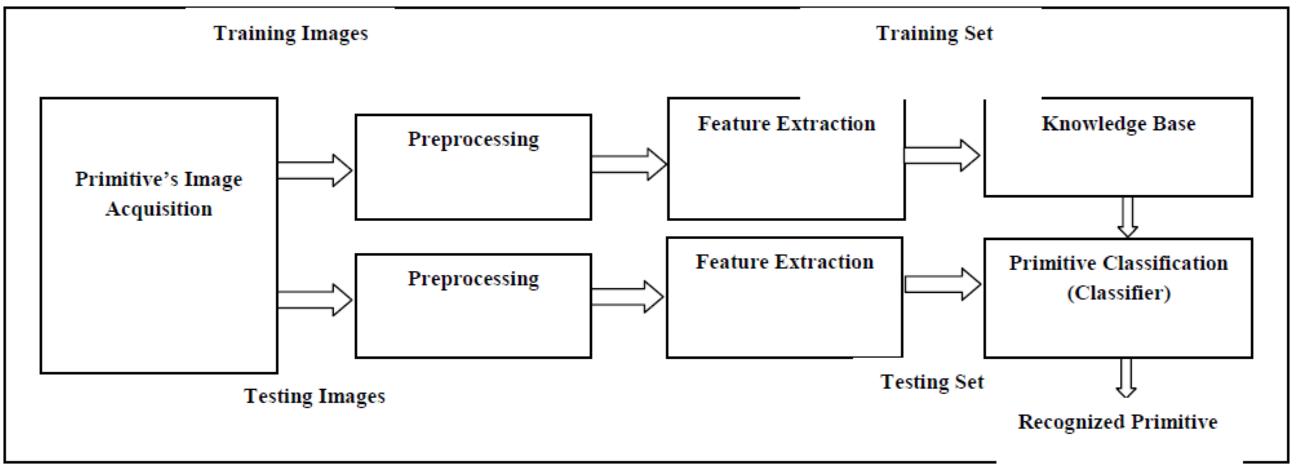


Figure 1 : Phases in Primitive Recognition

a) Identified Primitives and Image Acquisition

This step consists of two tasks, namely, identifying primitives of printed as well as handwritten Kannada vowels and consonants and obtaining their images.

i. *Identified Primitives*

The Kannada language characters are classified into Swaras (vowels), Vyanjanas (consonants), Yogavaahakas (partly vowels and partly consonants), Kagunitha (combination of consonants and vowels) and Wothakshara (conjunct consonants) as given in Table 1. Kannada language script consists of more than 250 basic, modified and compound character shapes giving rise to 18511 distinct characters. We have used the word Kannada and Kannada language interchangeably in this paper. Kannada characters are curve shaped with some regions highly denser than others. Some shapes are wider and some are longer than others, as visible in Table1.

Table 1 : Kannada Vowels, Consonants and Sample Kagunitha

Vowels	Yogavaahaka's		Consonants		Kagunitha
Swaras	Anusvara	Visarga	Structured	Unstructured	
ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ	ಅಂ	ಅಃ	ಕ ಖ ಗ ಘ ಙ ಚ ಛ ಜ ಝ ಞ ತ ಥ ದ ಧ ನ ಟ ಠ ಡ ಢ ಣ ಪ ಫ ಬ ಭ ಮ	ಯ ರ ಲ ವ ಶ ಷ ಸ ಹ ಳ	ಕ ಕಾ ಕಿ ಕೀ ಕು ಕೂ ಕೃ ಕೆ ಕೇ ಕೈ ಕೊ ಕೋ ಕೌ ಕಂ ಕಃ

We have consulted the Kannada language experts and identified the primitives. These 38 primitives are categorized into basic primitives (BP) and character cum primitives (CcP) as shown in the Figure 2. A single primitive, which is also a complete vowel or complete consonant, is defined as Character cum Primitive. One or more basic primitives are joined at appropriate positions to form the given vowels and consonants. It is also observed that symmetry exists in most of the Kannada characters.

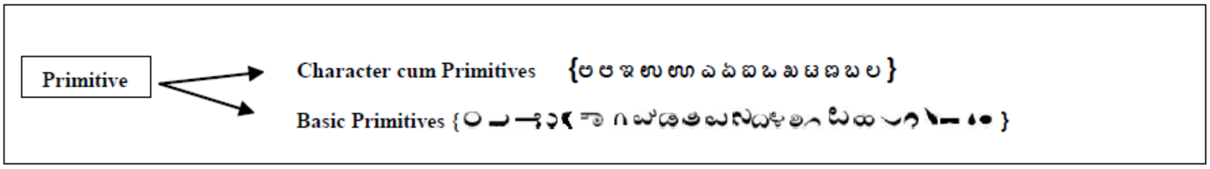
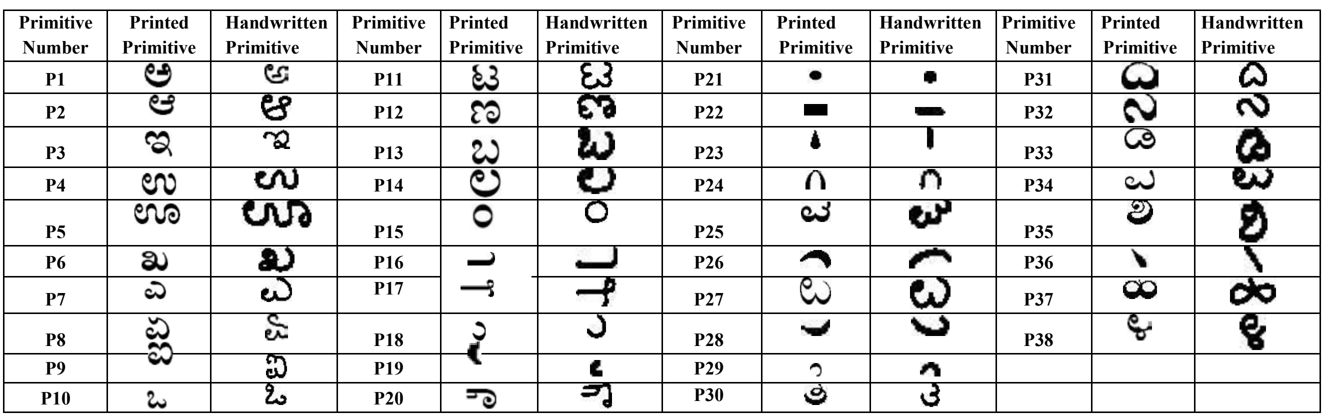


Figure 2 : BP and CcP in Kannada Vowels and Consonants

ii. *Image Acquisition of Primitives*

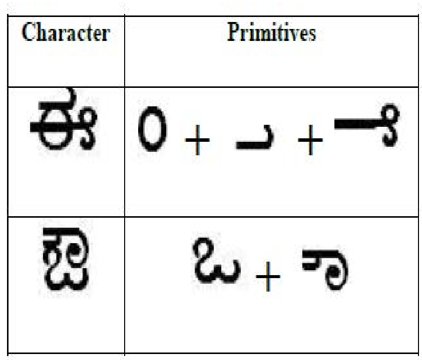
Since there is no standard database available for Kannada primitives, we have created a database of primitives for printed and handwritten Kannada vowels and consonants in consultation with language experts. We have identified 38 primitives to construct all the Kannada vowels and consonants and are given in Table 2. For example, the basic primitives required for constructing the two vowels ಈ (pronounced as „e") and ಔ (pronounced as „aou") are given in Box2. The „+" symbol represents the connection of primitives at appropriate positions for constructing a character. A database of 9120 primitives" images is prepared consisting of different font sizes and font types of printed Kannada vowels and consonants. Handwritten 3800primitives" images are collected from 100 writers who have written all the 38 primitives on plain A4 sheet. These sheets are scanned through HP Laser Jet M 1213nf MFP scanner having resolution of 300 dpi. The primitives are extracted manually from the printed as well as handwritten Kannada vowels and consonants and the images are stored as binary images in .jpg format.

Table 2 : List of Kannada Vowel and Consonant Primitives



Primitive Number	Printed Primitive	Handwritten Primitive	Primitive Number	Printed Primitive	Handwritten Primitive	Primitive Number	Printed Primitive	Handwritten Primitive	Primitive Number	Printed Primitive	Handwritten Primitive
P1			P11			P21			P31		
P2			P12			P22			P32		
P3			P13			P23			P33		
P4			P14			P24			P34		
P5			P15			P25			P35		
P6			P16			P26			P36		
P7			P17			P27			P37		
P8			P18			P28			P38		
P9			P19			P29					
P10			P20			P30					

Box 2 : Sample Kannada Vowels and required Primitives

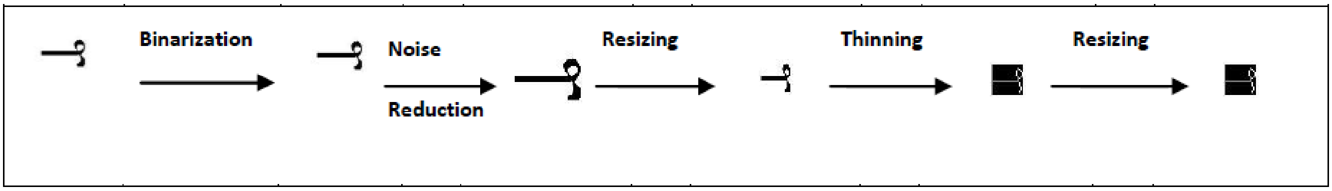


b) *Preprocessing*

We have preprocessed the images of the primitives to make them suitable for feature extraction. Preprocessing involves binarization, noise reduction, size normalization and thinning. The binarization is categorized into two main classes, namely, global and local. We have adopted global approach for converting gray scale image to binary image. Image binarization is performed using Otsu's method. The salt and pepper noise present in the image is removed by applying median filter. The process of thinning involves reducing thickness of each line of the pattern to just a single pixel wide is carried out. Size normalization is required as the size of the primitives vary from one vowel to another. In

Global Journal of Computer Science and Technology (F) Volume XIV Issue IV Version I Year 2014
45

order to bring uniformity among the images of primitives, each image is normalized to the size of 28*28 after finding the bounding box of each image without disturbing the aspect ratio using bilinear standard transformation. The images of primitive obtained after applying all the preprocessing steps to a given sample primitive is shown in Box 3.



Box 3 : Preprocessing steps

c) Feature Extraction and Knowledge Base

We have used zone based feature extraction technique. The number of zones and their sizes are decided based on the classifier accuracy. Figure 3 shows the behavior of the classifier for image sizes ranging from 20*20 to 40*40. Since the primitives, with the image sizes 28*28, have given maximum classification accuracy, the image size of 28*28 is chosen for all the primitives. Figure 4 shows the behavior of the classifier for different zone sizes for the given image size of 28*28. Since the zone size of 4*4 gives maximum accuracy, this size is chosen and the image is divided into 49 zones (7*7) and each is of size 4*4. The information present in each zone is used in defining the features. Figure 5a shows the 7*7 zones of the primitive P7 ( ) and Figure 5b shows the zone feature values obtained for the primitive P7 ( ).

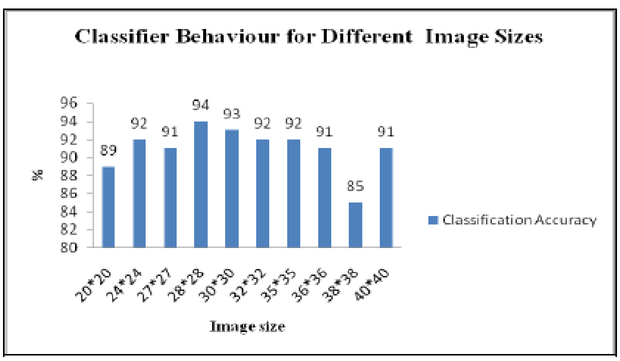


Figure 3 : Behavior of the Classifier for different Image Sizes

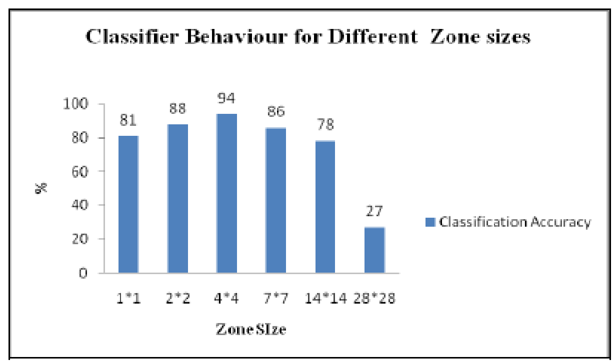


Figure 4 : Behavior of the Classifier for different Zone Sizes

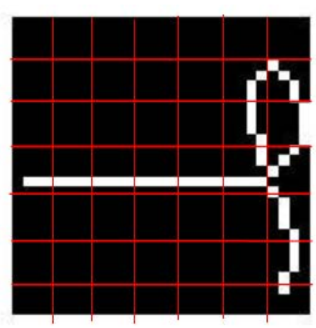


Figure (a) : 7×7 Zones of primitive P9( )

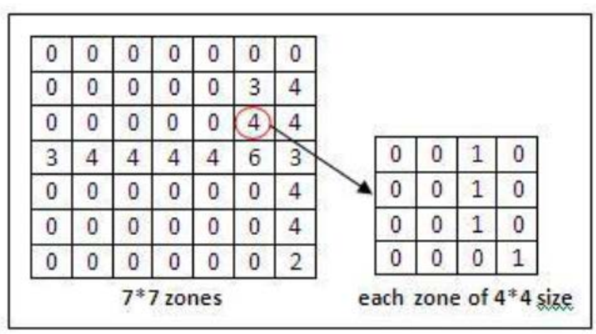


Figure (b) : Sample zone feature values for the primitive P9( )

Figure. 5 : Intermediate steps for feature extraction

A total of 49 features values are extracted from each primitive and this will serve as the feature vector. The feature vector for image (i) denoted by Fi = {z1, z2, z3 … z49}, where zi denotes ith zone value. There could be some zones, which do not contain any part of the primitive at all; therefore the corresponding zone value in the feature vector is zero. The set of feature vectors obtained from the training samples is used as the Knowledge Base (KB). This knowledge base is used to recognize the test samples. We have used nearest neighbor classifier for recognition.

d) Classification

Firstly, a training set consisting of images whose class labels are known is formed. The training set is used to build a classification model which is subsequently applied to the test set, which consists of images with unknown class labels. Evaluation of the performance of a classifier is done based on the counts of test images being correctly classified. The classifier used in the work is the nearest neighbor classifier. It considers each given input pattern and classifies itto a certain class by calculating the distance between the input pattern and the training patterns. The decision is generally based on the class values of the nearest neighbors. In this work, we have computed the distance between features of the test sample and the features of every training sample using the L2 Euclidean distance measures given inBox 4.

$$d(p,q) = \left(\sum_{k=1}^{n} | p_k - q_k |^r\right)^{\frac{1}{r}} \quad \ldots \quad (1)$$

Where r is a parameter varied from 1 to ∞, n is the number of attributes and p_k and q_k are, respectively, the k^{th} attributes (also called components) of the images p and q. $r = 2$ L_2 - Euclidean distance

Box 4 : Distance Measures used

The minimum distance between the test image and training image data in the knowledge base is used to decide the type of the primitive. A nearest neighbor classifier using Euclidean distance measure is used for the classification purpose. The procedure adopted for primitive recognition is given in the form of Algorithm1.

i. *Algorithm 1: Recognition of primitives*

Input : Kannada vowel primitive images.
Output: Recognition of the primitive.
Description: Zone based features and the nearest neighbor classifier are used in the work. Each zone is of 4x4 pixels. Euclidean distance measure is being used.

Start

Step 1. Accept and preprocess the input image to eliminate the noise using median filter and perform thinning.

Step 2. Fit the input image in a bounding box and crop the image to resize to 28*28 pixels.

Step 3. Extract 49 zone values, define feature vector and store.

Step 4 . Repeat steps 1 to 3 until the training images are exhausted.

Step 5. Accept the test sample.

Step 6. Compute Euclidean distance of the test sample with all the trained images.

Step 7. Declare the class of the primitive as the class with minimum distance using Nearest neighbor
classification.

Step 8. Repeat steps 5 and 7 until test images are exhausted.

Step 9. Obtain the accuracy of classification.

Stop.

The accuracy of the classifier is evaluated through k-fold cross-validation method. In this method, each time one of the k subsets of images is used as the test set of images and the other (k-1) subsets are put together to form a training set of images. The advantage of this method is that it matters less how the images are divided. The variance of the resulting estimate is reduced as k is increased. We have computed the average error across all k trials.

IV. Results and Discussion

In experimentation of the methodology, we have considered 60 font styles and 100 font sizes. The different combinations are tried and are as given in Table 3. We have totally five combinations of font styles and font sizes. The bit 0 indicates varying and 1 indicates constant. The experimentation is done on 9120 (38*240) images of printed Kannada vowel primitives. We have considered 240 images with varying font size and font styles for each primitive. The font size and font styles used are given in Table 4. The entire image set is partitioned into training set and test set and classified using K-fold cross validation method.

Table 3 : Combinations of Font Sizes(FS) and Font Styles(FSt)

Combinations	Font Size(S)	Font Styles(F)	Remarks
1.	0	0	Both Font Size and Font Style are Fixed
2.	0	1	Font Size is Fixed and Font Style is Varied
3.	1	0	Font Size is Varied and Font Style is Fixed
4.	1	1	Both Font Size and Font Style are Varied (Nonuniform mix)
5.	1	1	Both Font Size and Font Style are Varied (Uniform mix)

Table 4 : Font Sizes(FS) and Font Styles(FSt) used

Sl No	Font Size	Font Style
1.	12,14,16,……,110 (100 sizes)	Baraha 01
2.	12,14,16,……,110 (100 sizes)	Baraha 02
.....	12,14,16,……,110 (100 sizes)	…..
.....	12,14,16,……,110 (100 sizes)	Baraha 30
.....	12,14,16,……,110 (100 sizes)	Nudi 01
.....	12,14,16,……,110 (100 sizes)	…..
60.	12,14,16,……,110 (100 sizes)	Nudi 30

a) Both Font Size and Font Style Fixed

An experiment is carried out for a total of 39 primitives out of which 14 are character cum primitives and 24 are basic primitives. The image data set has 1520 (38*40) images. For example, we have considered 40 images of font size 60 and font style - Nudi 0.1. The zones based features and nearest neighbor classifier have given 100% recognition accuracy for both the types of primitives. Table 5 gives the results obtained for Character cum Primitives and Table 6 gives the results obtained for Basic primitives.

Table 5 : CcP's recog nition results for SFS and SFSt

Same Size and Same Font	
Image size=28*28 Zone=4*4 No of Samples=38*40=1520	
Primitive	Classification Accuracy (%)
P1	100
P2	100
P3	100
P4	100
P5	100
P6	100
P7	100
P8	100
P9	100
P10	100
P11	100
P12	100
P13	100
P14	100
Average	100

Global Journal of Computer Science and Technology (F) Volume XIV Issue IV Version I Year 2014

Table 6 : BP's recognition results for Same FS(SFS) and Same FSt(SFSt)

Same Size and Same Font				
Image size=28*28 Zone=4*4 No of Samples=38*40=1520				
Primitive	Classification Accuracy(%)		Primitive	Classification Accuracy(%)
P15	100		P27	100
P16	100		P28	100
P17	100		P29	100
P18	100		P30	100
P19	100		P31	100
P20	100		P32	100
P21	100		P33	100
P22	100		P34	100
P23	100		P35	100
P24	100		P36	100
P25	100		P37	100
P26	100		P38	100
Average	100			

b) Font Size Fixed and Font Style Varied

An experiment is carried out for 2280 (38*60) images. For each primitive, 60 images of fixed font size - 60 and 60 different font styles, as given in Table 4, are considered. The classifier is subjected to k-fold cross-validation. We have considered 1520 primitive images in each validation step for training and 760 primitive images for testing. Table 7 and Table 8 give recognition accuracy for this combination of CcP and BP primitives using 3- fold cross validation method. The range of recognition accuracies obtained for both, character cum primitives and basic primitives are given in Table 15.

Table 7 : 3-Fold cross validation of CcP's

3_Fold Cross Validation				
Same Size and Different Font using Euclidean Distance				
Image size=28*28 Zone size=4*4 No of Samples=38*60=2280				
Primitive	1_Fold	2_Fold	3_Fold	Average
P1	70	80	80	76.666667
P2	65	75	50	63.333333
P3	70	100	90	86.666667
P4	100	95	100	98.333333
P5	95	100	100	98.333333
P6	75	85	90	83.333333
P7	60	85	100	81.666667
P8	70	70	85	75
P9	70	75	85	76.666667
P10	75	75	95	81.666667
P11	95	85	65	81.666667
P12	95	85	90	90
P13	70	75	60	68.333333
P14	80	85	85	83.333333
Average	77.857143	83.571429	83.928571	81.785714

Table 8 : 3-Fold cross validation of BP's

3_Fold Cross Validation				
Same Size and Different Font using Euclidean Distance				
Image size=28*28 Zone size=4*4 No of Samples=38*60=2280				
Primitive	1_Fold	2_Fold	3_Fold	Average
P15	75	80	85	80
P16	85	95	80	86.666667
P17	100	100	100	100
P18	85	80	80	81.666667
P19	80	60	60	66.666667

P20	75	85	95	85
P21	25	80	90	65
P22	20	25	25	23.333333
P23	30	25	20	25
P24	75	90	85	83.333333
P25	70	95	100	88.333333
P26	55	55	60	56.666667
P27	70	95	65	76.666667
P28	70	55	60	61.666667
P29	75	80	75	76.666667
P30	80	75	90	81.666667
P31	95	65	80	80
P32	85	95	95	91.666667
P33	65	80	85	76.666667
P34	85	70	80	78.333333
P35	85	90	95	90
P36	95	85	85	88.333333
P37	85	95	90	90
P38	90	95	95	93.333333
Average	73.125	77.083333	78.125	76.111111

c) Font Size Varied and Font Style Fixed

An experiment is carried out on 1520(38*40) images. We have considered 40 images of different font sizes, as given in Table 4, font style of Nudi 01 for each primitive. Table 9 and Table 10 give results of classification for varying font sizes and fixed font style using 4- fold cross validation using Euclidean distance. The range of recognition accuracies obtained for this combination in case of both character cum primitives and basic primitives are given in Table 15.

Table 9 : 4-Fold cross validation of CcP's

2_Fold Cross Validation			
Different Size and Same Font using Euclidean Distance			
Image size=28*28 Zone size=4*4 No of Samples=38*40=1520			
Primitive	1_Fold	2_Fold	Average
P1	100	100	100
P2	95	100	97.5
P3	100	100	100
P4	100	100	100
P5	100	100	100
P6	100	100	100
P7	100	100	100
P8	100	100	100
P9	100	100	100
P10	100	100	100
P11	100	100	100
P12	100	100	100
P13	100	100	100
P14	100	100	100
Average	99.64286	100	99.82143

Table 10 : 4-Fold cross validation of BP'

2_Fold Cross Validation			
Different Size and Same Font using Euclidean Distance			
Image size=28*28 Zone size=4*4 No of Samples=38*40=1520			
Primitive	1_Fold	2_Fold	Average
P15	100	100	100
P16	100	100	100
P17	100	100	100
P18	100	100	100
P19	100	100	100

Global Journal of Computer Science and Technology (F) Volume XIV Issue IV Version I Year 2014

P20	100	100	100
P21	85	85	85
P22	65	70	67.5
P23	70	80	75
P24	100	100	100
P25	100	95	97.5
P26	100	100	100
P27	95	100	97.5
P28	100	95	97.5
P29	100	100	100
P30	100	100	100
P31	100	100	100
P32	100	100	100
P33	100	100	100
P34	100	100	100
P35	100	100	100
P36	100	100	100
P37	100	100	100
P38	100	100	100
Average	96.458333	96.875	96.666667

Table 11 : 5-Fold cross validation of CcP's (Non -uniform mix)

5_Fold Cross Validation						
Different Size and Different Font using Euclidean Distance						
Image size=28*28 Zone size=4*4 No of Samples=38*100=3800						
Primitive	1_Fold	2_Fold	3_Fold	4_Fold	5_Fold	Average
P1	95	95	90	100	100	95
P2	95	85	95	100	100	95
P3	100	100	100	100	90	98
P4	100	100	100	100	100	100
P5	95	100	100	100	100	99
P6	100	100	95	100	100	99
P7	100	95	100	100	100	99
P8	95	95	100	100	90	96
P9	95	100	90	100	100	97
P10	95	95	100	100	100	98
P11	90	100	100	100	100	98
P12	95	90	100	100	100	97
P13	85	95	95	90	100	93
P14	100	100	100	100	100	100
Average	95.714286	96.071429	97.5	99.285714	98.571429	97.428571

d) Both Font Size and Font Style Varied (Non-uniform mix)

An experiment is carried out on 3800 images. In non-uniform mix, we have considered 60 images of different font sizes, as given in the Table 4, of Nudi 01 font style and 40 images of 40 different font sizes for each primitive. Table 11 and Table 12 show the results for both varying font sizes and font styles using 5- fold cross validation using Euclidean distance. The recognition accuracy obtained for basic primitives and for character cum primitives is given in Table 15.

Table 12 : 5-Fold cross validation of BP's (Non -uniform mix)

5_Fold Cross Validation						
Different Size and Different Font using Euclidean Distance						
Image size=28*28 Zone size=4*4 No of Samples=38*100=3800						
Primitive	1_Fold	2_Fold	3_Fold	4_Fold	5_Fold	Average
P15	100	100	100	100	90	98
P16	90	95	100	100	100	97
P17	100	100	100	100	100	100
P18	95	90	100	100	100	97
P19	90	90	95	100	80	91
P20	90	90	100	100	90	94

Global Journal of Computer Science and Technology (F) Volume XIV Issue IV Version I Year 2014

P21	50	60	50	35	50	49
P22	35	20	35	60	40	38
P23	55	75	65	70	85	70
P24	100	95	100	100	100	99
P25	95	100	100	90	90	93
P26	90	90	90	100	100	94
P27	90	90	90	100	100	96
P28	85	95	100	90	100	93
P29	85	95	100	90	90	91
P30	100	100	95	100	100	98
P31	95	95	90	100	100	96
P32	100	100	100	100	90	98
P33	100	95	100	100	100	99
P34	100	100	100	90	100	98
P35	100	95	100	100	100	99
P36	100	100	100	100	90	98
P37	100	95	100	90	100	97
P38	95	100	100	100	100	99
Average	89.166667	89.583333	92.083333	92.291667	91.458333	90.916667

e) Both Font Size and Font Style Varied (Uniform mix)

An experiment is carried out on 3900 (39*100) images. We have considered100 images of 10 font styles and 10 varying sizes, for each font style, as given in Table 4, for each primitive. Table 13 and Table 14 show the results for both varying font sizes and font styles using 5- fold cross validation using Euclidean distance. The recognition accuracy obtained for basic primitives and for character cum primitives is given in Table 15.

Table 13 : 5-Fold cross validation of CcP's (Uniform mix)

5_Fold Cross Validation						
Different Size and Different Font using Euclidean Distance						
Image size=28*28 Zone size=4*4 No of Samples=38*100=3800						
Primitive	1_Fold	2_Fold	3_Fold	4_Fold	5_Fold	Average
1	100	100	100	90	100	98
2	95	100	100	90	100	97
3	100	95	100	90	90	95
4	100	95	100	100	100	99
5	100	100	100	100	100	100
6	100	100	95	100	90	97
7	100	100	100	100	95	99
8	100	90	100	90	85	93
9	90	100	100	90	100	96
10	100	90	95	100	95	96
11	100	100	90	100	85	95
12	100	100	90	85	90	93
13	100	100	90	95	95	96
14	95	100	100	95	95	97
Average	98.571429	97.857143	97.142857	94.642857	94.285714	96.5

Table 14 : 5-Fold cross validation of BP's(Uniform mix)

5_Fold Cross Validation						
Different Size and Different Font using Euclidean Distance						
Image size=28*28 Zone size=4*4 No of Samples=38*100=3800						
Primitive	1_Fold	2_Fold	3_Fold	4_Fold	5_Fold	Average
P15	95	85	70	75	60	77
P16	90	85	95	100	90	92
P17	100	90	95	100	95	96
P18	85	95	95	100	95	94
P19	80	95	100	100	100	95
P20	100	90	90	100	85	93

P21	100	100	100	100	100	100
P22	40	30	50	50	50	44
P23	100	90	100	100	100	98
P24	95	95	100	95	65	90
P25	100	95	95	80	90	92
P26	70	95	55	55	100	75
P27	100	85	85	75	55	80
P28	95	75	85	45	95	79
P29	80	80	85	55	90	78
P30	100	100	100	95	90	97
P31	100	90	95	90	95	94
P32	100	85	95	90	70	88
P33	95	95	90	100	80	92
P34	100	100	95	80	70	89
P35	100	95	95	90	85	93
P36	100	100	90	85	50	85
P37	100	95	100	70	75	88
P38	95	100	100	95	85	95
	92.5	89.375	90	84.375	82.083333	87.666667

Table 15 : Combinations of font sizes and font styles and their recognition accuracy for CcP's and BP's

Combinations	Printed Basic Primitives		Printed Character cum Primitives	
	Euclidean Distance		Euclidean Distance	
	Range	Avg	Range	Avg
Both Font Size and Font Style are Fixed	100%	100%	100%	100%
Font Size is Fixed and Font Style is Varied	25%--93%	76%	63%--98%	82%
Font Size is Varied and Font Style is Fixed	68%-100%	97%	98%-100%	99%
Both Font Size and Font Style are Varied(Non-uniform	38%-100%	90%	93%-100%	97%
Both Font Size and Font Style are Varied(Uniform mix)	44%-100%	88%	93%-100%	87%
Average		**90%**		**93%**
	Handwritten Basic Primitives		Handwritten Character cum Primitives	
	Euclidean Distance		Euclidean Distance	
Average	**93%**		**94%**	

V. Conclusion

We have proposed a zone features based methodology for recognition of both printed and handwritten 38 primitives of 49 Kannada vowels and consonants together. Zones of 4*4 and images of 28*28 are used. The nearest neighbor classifier is used with Euclidean distance measure. The accuracy of the classifier is verified with k-fold cross validation, for k = 2,3 and 5. We have experimented with four combinations of font sizes and font styles for printed primitives and obtained average recognition accuracy in the range [89%, 94%]. Further, we have obtained accuracy in the range [90%, 94%]for handwritten primitives. This work is basic to automation of writing of Kannada kagunitha and wothakshara"s.

References Références Referencias

1. Leena R Ragha and M Sasikumar , (2011) "Feature Analysis for Handwritten Kannada Kagunita Recognition,International Journal of Computer Theory and Engineering, Vol.3, No.11793-8201
2. S. K. Sangame, R.J. Ramteke and Yogesh V. G. (2011) "Recognition of isolated handwritten Kannada characters using invariant moments and chain code" world journal of science and technology, 1(8): 115-120 ISSN: 2231 – 2587
3. K S Prasanna Kumar, (2012) "Optical Character Recognition (OCR) for Kannada numerals using Left Bottom 1/4th segment minimum features extraction" International Journal of Computer Techology & Applications, Vol 3 (1) , 221-225 ISSN:2229-6093
4. Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre,(2013) "Character Recognition Using Matlab"s Neural Network Toolbox", International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1
5. Mamatha Hosalli R amappa and Srikantamurthy Krishnamurthy, (2013) "A Comparative Study of Different Feature Extraction and Classification Methods for Recognition of Handwritten Kannada Numerals" International Journal of Database Theory and Application Vol. 6, No. 4
6. Manjunath A E Sharath B(2013) "Implementing Kan-nada Optical Character Recognition on the Android

Operating System for Kannada Sign Boards"International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1

7. Umapada Pal, Ramachandran Jayadevan and Nabin Sharma (2012), "Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques", ACM Transactions on Asian Language Information Processing, Vol. 11, No. 1, Article 1, Publication date: March 2012.
8. Leena R Ragha and M Sasikumar , (2010) "Using Moments Features from Gabor Directional Images for Kannada Handwriting Character Recognition", International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) – TCET, Mumbai, India
9. Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, Benoy Kumar Thakur (2013) "Recent Trends and Tools for Feature Extraction in OCR Technology", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013
10. Karthik Sheshadri, Pavan Kumar T Ambekar, Deeksha Padma Prasad and Dr. Ramakanth P Kumar(2010) "An OCR system for Printed Kannada using k-means clustering" 978-1-4244-5697-0/10/$25.00 ©2010 IEEE.
11. B.V.Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge (2011), "Zone Based Features for Handwritten and Printed Mixed Kannada Digits Recognition", International Conference on VLSI, Communication & Instrumentation (ICVCI) 2011 Proceedings published by International Journal of Computer Applications®.
12. H. Imran Khan, Smitha U. V, Suresh Kumar D. S (2013), "Isolated Kannada Character Recognition using Chain Code Features", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, Volume2 Issue 8, August 2013
13. Mamatha.H.R , Sucharitha Srirangaprasad, Srikantamurthy K (2013) "Data fusion based framework for the recognition of Isolated Handwritten Kannada Numerals", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 6, 2013
14. Swapnil A. Vaidya, Balaji R. Bombade(2013) "A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction", International Journal of Computer Science and Mobile Computing Vol.2 Issue.6, June- 2013, pg. 179-186
15. Nithya E, Dr. Ramesh Babu D R (2013) "OCR System for Complex Printed Kannada Characters", International Journal of Advanced Research in Computer Science and Software Engineering 3(6), June - 2013, pp. 102-105
16. G. G. Rajput, Rajeswari Horakeri(2011) "Shape Descriptors based Handwritten Character Recognition Engine with Application to Kannada Characters", International Conference on Computer & Communication Technology (ICCCT)-2011.
17. G. G. Rajput, Rajeswari Horakeri(2013) "Zone based Handwritten Kannada Character Recognition Using Crack code and SVM", 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)-2013.
18. Sandhya.N, Krishnan.R, D.R.Ramesh Babu (2014) "Feature Based Kannada Character Classification Method of Kannada Character Recognition" International Journal of Scientific & Engineering Research, Volume 5, Issue 2, February-2014 729 ISSN 2229-5518.