ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(24), DOI: 10.17485/ijst/2015/v8i24/79996, September 2015* ISSN (Online) : 0974-5645

Innovative Feature Sets for Machine Learning based Telugu Character Recognition

J . Jyothi*, K. Manjusha, M. Anand Kumar and K. P. Soman

Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore - 641 112, Tamil Nadu, India; jyothijammigumpula@gmail.com, manjushagecpkd@gmail.com, manandkumar@cb.amrita.edu, kpsoman@amrita.edu

Abstract

In this Information age, all sources of information like historic documents, books, manuscripts are digitized and are available all over the world through internet in the form of scanned copies. These scanned images contain valuable information which are available either in colour or black and white for pleasant viewing. Optical Character Recognition (OCR) technology provides facility to search for keywords in these digital copies. In this paper, new method in which building an OCR system for Telugu language script; mainly focussing on the character recognition module. Features extracted through Discrete Wavelet Transform (DWT), Projection Profile (PP) and Singular Value Decomposition (SVD) is evaluated using k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) classifiers. Most productive results are obtained from the DWT features with SVM classifiers.

Keywords: Discrete Wavelet Transform, K-Nearest Neighbour, Optical Character Recognition, Singular Value Decomposition, Support Vector Machine, Telugu Character Recognition

1. Introduction

Optical Character Recognition (OCR) can be defined as automatic identification of human readable text of any language from the scanned images of typewritten text, humanoid notes, computerized bills, digital prints and many more. OCR is a research area that comes under pattern recognition, Artificial Intelligence and computer vision. The elementary idea underlying the OCR is conversion of scanned copy of text in source language to corresponding machine editable character format. OCR can be integrated with several applications such as TTS (Text To Speech) for assisting visually impaired peoples to read the printed document, banking applications for identifying credit card, accounts number, customer id information etc. The research on OCR started a long time ago. Today good quality commercial OCR systems are available in the market which can identify several different language scripts such as Latin, Arabic etc. Nowadays, even cloud based computing environment are used for document conversion using OCR.

Usually OCR systems follow a feed forward architecture, which includes several sub processes. The pre-processing stage includes several processes such as noise removal, skew correction, thinning, binarization etc. Skew correction helps in aligning correct orientation of character with help of reference line and by using rotation algorithm[1]. Thinning is the process of reducing the width of character thereby increasing the crisp and smoothness of the character[2]. Binarization is the process of converting grey scale image to binary i-age. Histogram-shaped, Clustering, Entropy, Object attribute, Spatial binarization, Locally adaptive algorithms are various binarization techniques[3]. Segmentation is another vital part in OCR which comes after pre-processing. Segmentation the process of splitting the bigger units into the tiniest unit, which can't further be broken down. Feature extraction and classification processes are included in the recognition module of OCR systems which is the main focus in this paper. The architecture of OCR system is shown in Figure

* *Author for correspondence*

1. As shown in the Figure 1, generally OCR systems work in mainly two phases: Training and Testing. During training phase, the feature model for the language built using training image set. The real OCR works in testing stage, where the classification process makes use of the trained feature model for predicting the target class for the segmented character images in the image given.

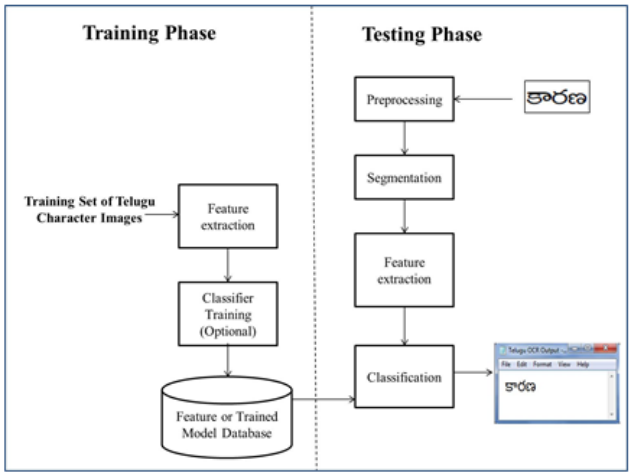

Figure 1. General architecture for OCR process.

2. Telugu Script

In this paper, proposed method focus on developing character recognition methodologies for Telugu OCR systems. Telugu is a Dravidian language and is second most popular speaking language with 75 million native speakers in India. It is the official language of Andhra Pradesh. It's highly derived from Sanskrit and parts of prakrit. Telugu alphabet has strong trace from Brahmi alphabets. Telugu language is syllabic by default. Rinki et al.[4] developed a system in which Telugu scanned documents converted to text using back propagation algorithm. An online recognition system using SVM classifier for Telugu character has been created by Vijay et al[5].

Mani et al.[7] used priority check algorithm for segmenting up paragraph and lines of the Telugu language documents. A character recognition approach using bi linear SVM which can be applied to any document image, irrespective of size and font of script is discussed in[8]. A general model of feature extraction and character recognition can be implemented by finding out base line, mean zones, etc. of the text present in the image[10].

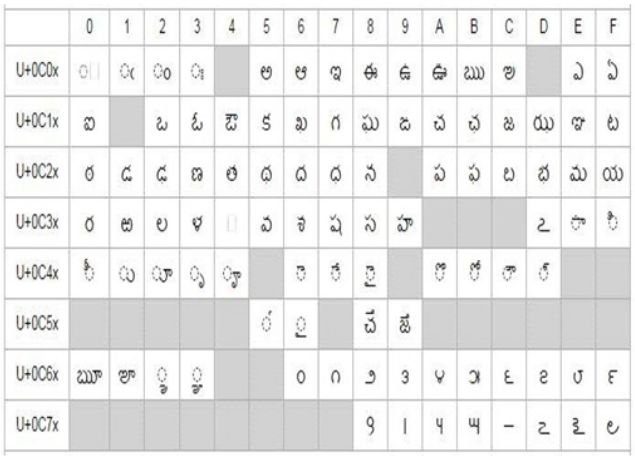

Figure 2. Unicode characters in Telugu language.

Telugu alphabets are known as onnamallu referring elemental alphabets O-Na-Ma-See-Vaa-Yaa-See-Dham-Namaha, originated from inception of a prayer. Telugu words end in vowels. Telugu has 15 vowels (achchu) where 3 got virtually weed out, 36 consonants (hallu), 10 numericals, 8 fractions, 3 dual symbols and 2 diaphones (ai and au). It has both soft (cha, ja) and harsh (chha, jha) pronunciation. Syllables of Telugu consist of consonants and vowels, probably end with vowel. Consonants in Telugu get related to consonants in Sanskrit with an exception of 2. Unicode points for Telugu range from OC00-OC7F. It has full zero, half zero and visarga symbols for producing various nasal sounds like la-La, ra-Ra. Figure 2 shows Unicode Telugu characters.

In this paper, we have experimented Telugu character recognition with SVM and k-NN classifier using Projection Profile, Discrete Wavelet Transform and Singular Value Decomposition features. The Feature extraction methods and classification algorithms used in the experiments which are explained in section 3. Section 4 is about the results we achieved with Telugu character recognition and section 5 is the summary and conclusion of the work towards Telugu character recognition.

3. Character Recognition

Character recognition is the process of converting the segmented character images to its corresponding character code. It includes feature extraction and feature classification processes. In proposed experiments, we have used features based on Singular Value Decomposition (SVD), Projection Profile (PP) and Discrete Wavelet

Transform (DWT). The extracted features are evaluated with k-Nearest Neighbor Classifier (k-NN) and Support Vector Machine (SVM).

3.1 Feature Extraction

The process of squeezing a set of features from high dimensional raw input data using certain algorithms is referred as feature extraction. The character image itself may act as the feature vector but the character image may contain highly redundant data of high dimensional space. The data redundancy and higher dimension of feature vector have to be lowered. These low dimensional features can help much in reducing the processing time taken by classifying algorithm. The feature extraction algorithms used in our experiments is described in following section.

3.1.1 Singular Value Decomposition (SVD)

Singular Value Decomposition is basically a matrix factorization technique which factorizes the given matrix to two unitary matrices and one diagonal matrix. After applying SVD to matrix A of size m n, it is decomposed into U having Eigen vector of AA^T in columns, a diagonal matrix with square root of Eigen values of A in its diagonal and V with Eigen vectors of A^T A. The diagonal values inside are also termed as singular values. These singular values can act as good feature descriptors that can be used for pattern recognition applications. For extracting features from Telugu character images, we have used SVD on different extracted blocks from the image. In proposed feature extraction technique using SVD, first the character image is divided horizontally and vertically to several different strips. All the horizontal strip have equal dimension 8 n, where n is the number of columns in the character image. The vertical strips have the dimension m 8, where m is the number of rows in character image. SVD is applied on the horizontal and vertical image strips and the singular values obtained from the matrix is appended together to create the feature descriptor to represent that character image. From each image strips we can get at most 8 singular values (if m>8 and n>8).

3.1.2 Projection Profile (PP)

Projection Profile based feature extraction is based on the histogram of the character image. If the character image is with white fore-ground pixels on black background then the white pixels (ON pixels) represents the character represented in that image. Horizontal Projection Profile (HPP) is the one dimensional column vector array referring to sum of ON pixels in the image along each row. Vertical Projection Profile (VPP) is a column vector of the image consisting of sum of ON pixels along each column. HPP and VPP can be appended together to represent feature descriptor extracted from the character image. Here PP is used to take features across vertical, horizontal and both for character image of size 32 32. Vertically, 32 features are considered, likewise for horizontally 32 features and for both we append horizontal and vertical values together of size 64.

3.1.3 Discrete Wavelet Transform (DWT)

DWT is the process of sampling the data values discretely in functional and numerical analysis. Its similar to Fourier transform except the fact that transform function is orthogonal. DWT of an image is found by passing the image through a series of Low Pass Filter (LPF) and High Pass Filter (HPF). The output of LPF will result in low frequency component or approximation coefficients and HPF will result into high frequency component or detailed coefficients at every level of decomposition. We are taking only approximation coefficients as feature for the experiments. The output of LPF holds half the original frequency range from first part of the image. Features of the character image are extracted by decomposing the image data matrix into approximation and detailed coefficients. We have experimented with Haar, db2 and db4 wavelets, among them db2 shows good accuracy on the training set.

3.2 Feature Classification

The selected and extracted feature can be further classified using any machine learning algorithms. There are mainly two types of classification algorithms: Supervised (labelled) and Unsupervised (unlabelled). We are using two supervised classifiers for Telugu character recognition k-Nearest Neighbour Classifier (k-NN) and Support Vector Machine (SVM).

3.2.1 k-Nearest Neighbour Classifier (k-NN)

k-NN is simple, powerful, non-parametric algorithm and requires no training process. For the test data features which have to be classified, Euclidean distance is calculated between that data point with all other data

points in training data set. The expression for calculating Euclidean distance between two points is given in[1]. Then these data points are sorted depending on Euclidean distance in ascending order. k parameter is usually fixed manually. To decide the target class of test data features, k nearest data points are considered from the training data set. The maximum labelled class out of k neighbours is taken as the target class for the test data features. For higher values of parameter k there will be less noise[11]. K-NN algorithm is used to classify feature points. We have experimented that the extracted features by changing the number of nearest neighbour (k) parameter.

3.2.2 Support Vector Machine (SVM)

SVM is generalized as kernel machines and are maximum margin methods for classification. SVM is based on finding maximum separability or margin between the different classes with the help of training data features. The main idea during training phase is to find the parameters for the hyper plane which maximally separates the different classes involved in the problem. The general SVM formulation for linear binary classification is given in[1].

$$\min_w; \ \frac{1}{2}\|w\|^2 \text{ subject to } d_k[w^T x_k] \quad (1)$$

where k = 1, 2, 3….m. $w^T x$ is the equation for hyper plane separating the m input samples, x2R. The decision function for new sample x is evaluated using f(x).

$$f(x) = \text{sign}\ (w^T x) \quad (2)$$

4. Experimental Result and Discussion

For this experimentation purpose, we have used two image databases; CMATERdb 3.4.1 and Telugu Character Database. CMATERdb 3.4.1[12] is Telugu handwritten numeral database (contains classes of digit 0 to 9) with 300 images for each class. Each image is of dimension 32 32. 75% of images from each class are taken for training and the retained 25% for testing purpose. The other database is Telugu character database which contains totally 128 character classes including vowels, consonants and most frequently used compound characters in Telugu language. This character database is built with the help of OCROPUS software[13] by considering fonts Pothana and Vemana in different styles and sizes with various degradation levels. The evaluation measure we have used is character recognition accuracy and is estimated by finding the ratio of correctly identified test character images and total number of test character images. All of experiments done are conducted in MATLAB environment. SVD, PP and DWT features are extracted from CMA-TERdb and are experimented with k-NN classifier by varying k from 1 to 20. The accuracy obtained for the recognition process is listed in Table 1. For k equals 4, accuracy reaches 93.20% in case of DWT features. A visual representation of the results obtained is shown in Figure 3. From Figure 3, it is clear that DWT features outperform SVD and PP features. In case DWT features, incrementing k value in k-NN beyond 4 decreases the character recognition accuracy. In case of PP features, for low value of k the accuracy is very low compared to DWT and SVD. On increasing k value the accuracy also increases for PP features and reaches 92.27% when k = 15, afterwards the accuracy decreases. For SVD features, the maximum accuracy obtained is 89.73% in CMATERdb for k = 5,6.

Table 1. Accuracy obtained for SVD, PP and DWT features in k-NN classifier for different k values on CMATERdb 3.4.1

k	SVD	PP	DWT
1	88.53	87.87	92.93
2	88.53	87.87	92.93
3	88.40	88.93	93.07
4	89.07	90.13	93.20
5	89.73	90.40	92.27
6	89.73	90.67	92.27
7	89.47	90.93	92.00
8	89.60	90.67	91.00
9	88.67	90.80	91.20
10	89.33	90.80	91.20
11	89.20	90.80	90.93
12	89.33	91.47	91.33
13	88.80	91.33	91.20
14	89.20	92.00	91.20
15	88.40	92.27	90.40
16	88.13	92.00	90.13
17	88.00	91.73	90.00
18	87.60	91.47	89.60
19	87.87	91.20	89.60
20	87.47	91.33	89.87

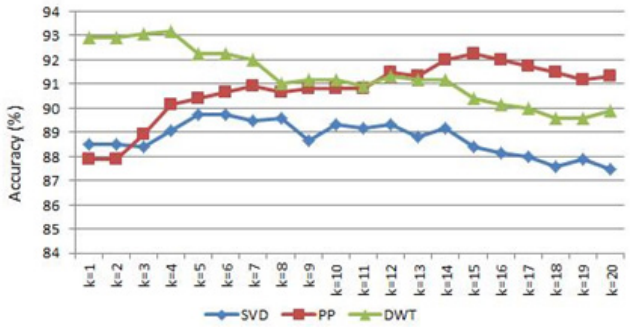

Figure 3. k-NN classification accuracy for SVD, PP and DWT on CMATERdb 3.4.1.

The same SVD, PP and DWT features are evaluated using Linear SVM classifier and the result obtained is shown in Figure 4. Character recognition accuracy is improved in SVM than k-NN for DWT features. Linear SVM classifier could classify 95.47% of the test samples accurately when DWT features are used.

Table 2. Accuracy obtained for SVD, PP and DWT features in k-NN on Telugu character database

k	SVD	PP	DWT
1	89.50	87.70	96.39
2	89.50	87.70	96.39
3	90.24	89.29	96.18
4	90.14	89.29	95.86
5	89.93	89.61	95.33
6	90.14	89.08	95.97
7	89.18	87.27	94.49
8	89.08	87.70	94.80
9	88.65	86.64	93.85

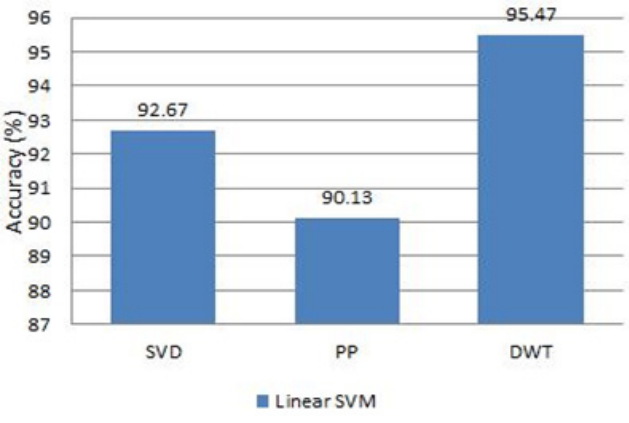

Figure 4. Performance for SVD, PP and DWT features in SVM classifier on CMATERdb 3.4.1

Telugu character image database consisting of 128 different classes is tested by extracting SVD, PP and DWT features for k-NN and SVM classifier. The accuracy obtained for the extracted features in k-NN and SVM classifiers are shown in Figure 5 and Figure 6 respectively. From Figure 5 and Figure 6 it is obvious that DWT features outperform PP and SVD features. DWT features could obtain 97.77% percent accuracy in Linear SVM classifier for the Telugu character image database. The accuracy obtained in k-NN for Telugu character image is listed in Table 2.

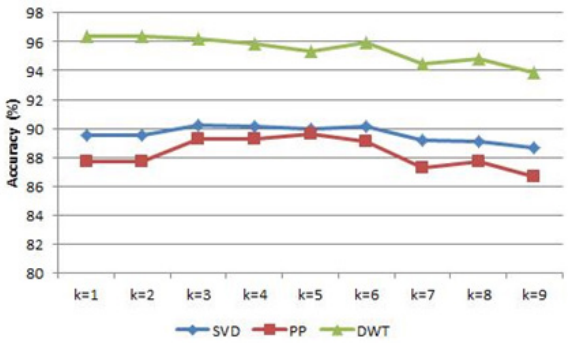

Figure 5. Accuracy for k-NN classification of SVD, PP and DWT features in Telugu character database.

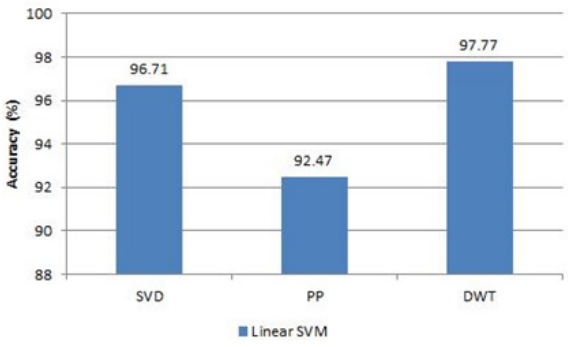

Figure 6. Performance of SVM for SVD, PP and DWT features in Telugu character image database.

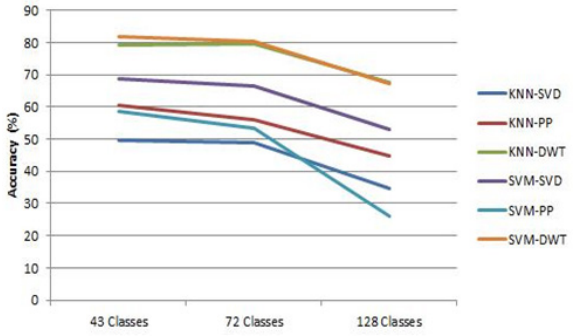

Figure 7. Performance of KNN and SVM classifier using Real time testing for Telugu OCR system.

For extending the experiments to real time document images, we have collected 157 Telugu printed word images and segmented into character level images. The word images are segmented to character by active contour algorithm which is based on level set theory. The segmented character images are then recognized using the trained model of Telugu character image database. Identifying the misclassified characters is done manually by matching with correct Telugu characters. For evaluating the performance of SVD, PP and DWT over the increase in number of character classes, the word images are tested by considering different number of character classes. In our first experiment, only 43 classes with vowels and consonants in Telugu language is considered. The word images are experimented by extracting SVD, PP and DWT features in k-NN and SVM classifier. As k = 4 in k-NN was giving higher accuracy in our experiments with Telugu Character image database, so we set k = 4 in these experiments. Then the above mentioned experiments are repeated for 72 classes which included vowels, consonants and some frequently used compound characters in Telugu language. Again the number of classes is increased by adding some more compound characters and tested with totally 128 characters. The results obtained for the real world Telugu image recognition is listed in Table 3. With 43 classes, SVM classifier with DWT features could achieve 81.77% correct classification rate. For 72 classes also the DWT features in SVM classifier could get the higher accuracy of 80.51%. But when 128 classes are considered the accuracy for DWT is reduced to 67.23% in SVM classifier and 67.48% in k-NN classifier.

Table 3. Accuracy obtained for the real time word images of Telug

No. of Classes	Classifier	SVD	PP	DWT
43 classes (Vowels and Consonants Only)	k-NN	49.57	60.59	79.23
	SVM	68.64	58.47	81.77
72 classes (Vowels, Consonants and Compound characters)	k-NN	48.99	56.16	79.65
	SVM	66.47	53.29	80.51
128 classes (Vowels, Consonants and Compound characters)	k-NN	34.71	44.98	67.48
	SVM	53.05	26.16	67.23

In real time testing the performance of SVM and k-NN classifier is decreased for SVD, PP and DWT features with the increase in number of classes. Graphical representation for Table 3 is shown in Figure 7. It is clear from Figure 7 that DWT features gives a robust performance in real time word image recognition than SVD and PP features. DWT features in both k-NN and SVM gives almost same accuracy, while SVD features gives better performance with SVM classifier than k-NN classifier. The performance of PP features in SVM classifier decreased drastically as the number of characte classes increased.

5. Conclusion

OCR technology has an inevitable role in automatic conversion of the document images to machine understandable text format. In spite of highly efficient OCR systems are available for Latin scripts; activity towards the building of an OCR system for Indian languages is still in its progressing stage. In this paper, we have experimented with SVD, PP and DWT features in k-NN and SVM classifiers in the context of Telugu character recognition. For our experiments, we have used CM-TERdb 3.4.1 numerical hand written Telugu digit database and Telugu Character database containing 128 different classes. On CM-TERdb 3.4.1, DWT features on Linear SVM classifier gives an accuracy of 95.47%. On Telugu Character image database also DWT features are performing better than SVD and PP features. We have performed an analysis of SVD, PP and DWT features on the word images collected from real word document images. DWT features out performs SVD and PP on the real world word images collection. In our experiments, we have considered only frequently used Telugu characters with only limited set of font families. The work can be extended by creating a good training Telugu character database consisting of all different font families and all characters existing for Telugu language script.

6. References

1. Sadri J, Cheriet M. A new approach for skew correction of documents based on particle swarm optimization, International Conference on Document Analysis and Recognition. Barcelona. 2009 July 26–29. p. 1066–70.
2. Gayathri S, Sridhar V An improved fast thinning algorithm for fingerprint image. Int J Engineering Science and Innovative Technology. 2013 Jan; 2(1):264–70.
3. Sauvola J, Pietikinen M. Adaptive document image binarization. Pattern Recognition. 2000 Feb; 33(2):225–36.
4. Singh R, Kaur M. OCR for Telugu script using back-propagation based classifier. International Journal of Information Technology and Knowledge Management. 2010 Jul-Dec; 2(2):639–43.

5. Swethalakshmi H, Jayaraman A, Chakravarthy VS, Sekhar CC. Online handwritten character recognition of Devanagari and Telugu Characters using support vector machines. 10th International Workshop on Frontiers in Handwriting Recognition Suvisoft; 2006 Oct.
6. Lakshmi CV, Patvardhan C. Optical character recognition of basic symbols in printed Telugu text. IE I Journal-CP. 2004 Jul; 7(2):190–204.
7. Bodduluri S, Krishna M, Babu R, Raghunadh MV. A novel way of identifying telugu, tamil and english scripts by priority check using discerning features. IOSR-JECE. 2014 Nov/Dec; 9(6):28–34.
8. Ashwin TV, Sastry PS. A font and size-independent OCR system for printed Kannada documents using support vector machines. Sadhana. 2002 Feb; 27(1):35–58.
9. Aparna KH, Subramanian V, Kasirajan M, Prakash GV, Chakravarthy VS, Madhvanath S. Online handwriting recognition for Tamil. International Workshop on Frontiers in Handwriting Recognition. 2004 Oct; 438–43.
10. Pal U, Chaudhuri BB. Indian script character recognition a survey. Pattern Recognition. 2004.
11. Gongde G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. In on the move to meaningful internet systems. CoopIS, DOA, and ODBASE. Berlin Heidelberg: Springer; 2003; 2888:986–96.
12. Das N, Reddy JM, Sarkar R, Basu S, Kundu M, Nasipuri M, Basu DK. A statistical topological feature combination for recognition of handwritten numerals. Applied Soft Computing. 2012 Aug; 12(8):2486–95.
13. Breuel TM. The OCR opus open source OCR system. International Society for Optics and Photonics in Electronic Imaging. 2008 Jan;.
14. Rao KRM, Ramesh B, Reddy GI. Font and size identification in Telugu printed document. 2013 Apr.; 6(11):92–110.