Hand Written Telugu Character Recognition Using Bayesian Classifier

K.Mohana Lakshmi[1] ,K.Venkatesh[2],G.Sunaina[3], D.Sravani[4], P.Dayakar[5]
ECE Deparment, JNTUH, CMR Technical Campus, Hyderabad, India.
[1] mohana.kesana@gmail.com
[2] kondaboynavenkatesh1995@gmail.com
[3] sunaina.gundu@gmail.com
[4] sravanidabilpura009@gmail.com
[5] pdayakar135@gmail.com

***Abstract*— Identifying handwritten Telugu is difficult way in machine vision systems because of the complex shape of the individual characters and the size of Telugu character set. In this paper, an efficient algorithm is presented to identify individual handwritten Telugu characters based on HOG features and Bayesian classification. The proposed system utilizes features of Telugu scripts for identifying handwritten Telugu text efficiently. The recognition rate for Telugu script is 87.5%.**

Keyword - Handwritten character recognition, Pre-processing, FeatureExtraction, Bayesian Classifier.

I. INTRODUCTION

Machine perception and recognition of handwritten character any language is a difficult task. It has many applications in banking, mail search, forms processing in administration and insurance. Telugu script presents additional challenges for handwriting recognition systems due to its highly connected components, curve nature forms of each letter and regional differences in writing style. Off-line Telugu Handwriting Recognition is a difficult task due to the high variability and uncertainty of human writing (infinite variations of shapes resulting from the writing style, scanning methods etc. The field of handwriting recognition can be divided into two approaches; on line recognition [1] and offline recognition (Miled H. 1998, Mantas J. 1991). Telugu is the primary language in Andhra Pradesh, Telangana, and Pondicherry. There are more than 10,000 old handwritten documents in Telugu. To identify old literature as new one, we have proposed an efficient mechanism to identify Telugu character set. Optical recognition is done off-line after the characters have been totally written or printed. Handwritten characters can be identified, but the characteristics of such systems can be largely dependent upon attribute of the input documents as well as the writing style of the writer. But, when it comes to completely unsupervised handwritten character recognition, machines are efficient as humans. However, the computer reads fast and technical advances are continually bringing the technology closer to its ideal.

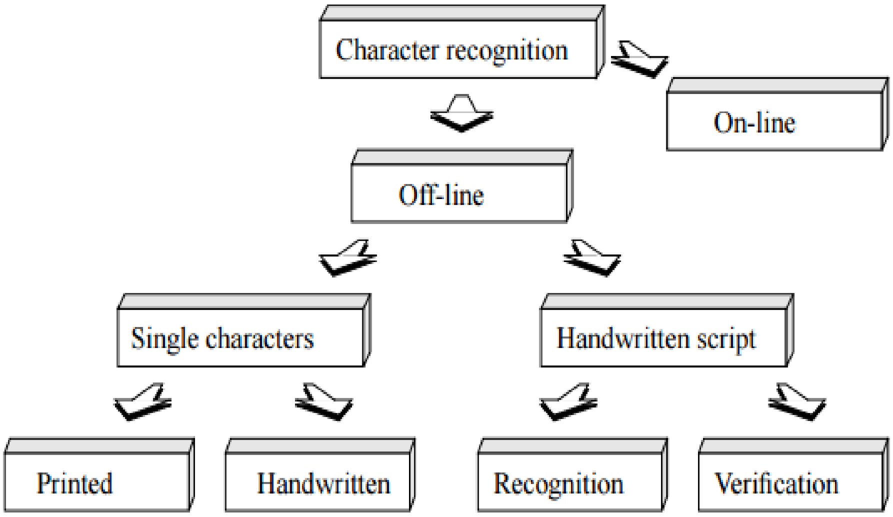


Figure 1. The different areas of character

U. Pal and B.B. Chaudhuri[2] suggested an enhanced quadratic classifier for identification of handwritten numerals in Devnagri, Bangla, Telugu, Oriya, Kannada and Tamil scripts. The primary idea behind automatic character recognition is to first train the machine classifier with instances of patterns that may occur and their appearance. In OCR these patterns are alphabets, numerals and some special symbols like commas, question marks etc., Using these samples the machine constructs a prototype or a description of each class of characters.

Now, for recognition, the new sample of characters are compared to the previously obtained descriptions and assigned the class that gives the best match. In most commercial systems for character recognition, the training process is performed.A typical OCR system consists of several components. In figure 2 a common setup is illustrated. The first step in the process is to digitize the analog document using an optical scanner. First inside the document, we need to locate the regions containing text and then extract them through a segmentation process.

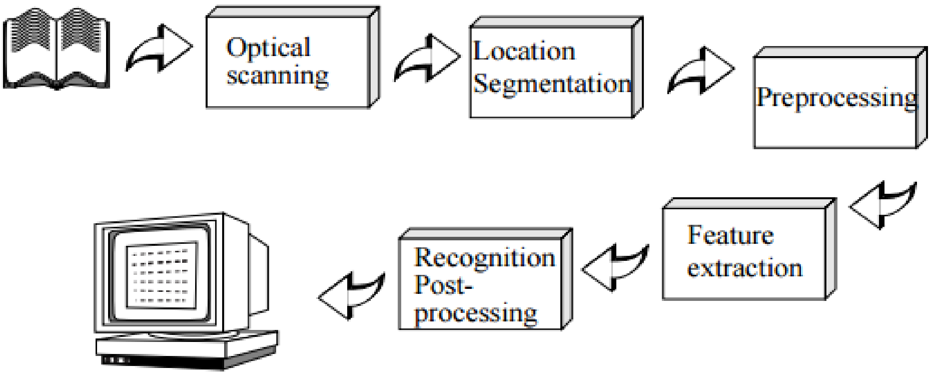


Figure 2. Components of an OCR system

Each symbol can be identified by comparing the extracted features with descriptions of the symbol classes created previously in the training phase. Finally, textual information is used to construct the words from the original text. The next section describes these steps and methodologies in detail.During the scanning process may contain a certain amount of noise. The characters may get broken or smeared based on the poor resolution of the scanner or the methodology applied for thresholding. The most widely accepted methodology for smoothing, moves a window over the binary image of the character, using specific rules to the contents of the window for smoothing.

The aim of feature extraction is to find the basic properties of the character set, and it is largely acknowledged that this is the most challenging issue in pattern recognition. The most obvious way of describing a character is by its original raster image. Another method is to calculate specific features that still describe these characters, but eliminate out the insignificant attributes. The methods for calculation of such features are frequently partitioned into three main classes, where the features are calculated from: (1) The distribution of points. (2) Transformations and series expansions. (3) Structural analysis.

The classification is the process towards distinguishing each character and allocating it to the suitable character type. There are many advantages of OCR, however to be specific it increases the proficiency and effectiveness of office work. The capability to search through a document instantaneously is very useful, more so in an office premise where one has to deal with high volume scanning or high document inflow. This paper describes the method of recognizing characters in a document image. Section II describes Telugu character set and existing methods for handwritten. Section III describes the experimental date and classifier used. Section IV describes the step by step algorithm. Section V reports the final conclusions.

II. TELUGU SCRIPT

There are 18 characters vowels and 36 characters consonants in Telugu language. Telugu characters and their pronunciations are given by

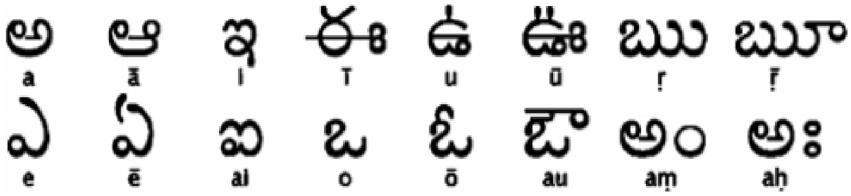
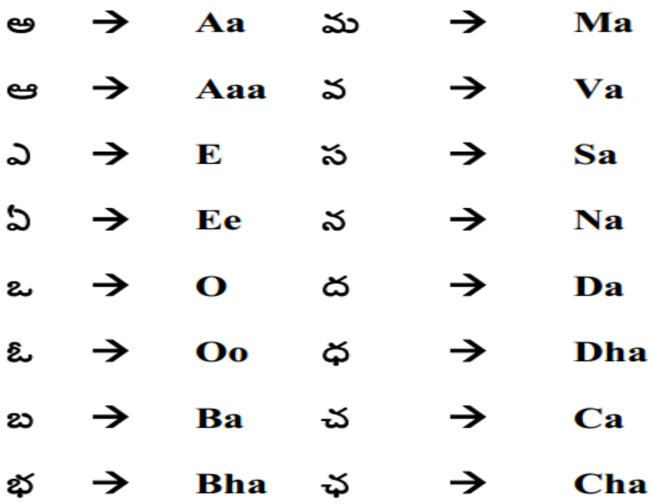


Figure 3. Some of Telugu characters



అ	→	Aa	మ	→	Ma
ఆ	→	Aaa	వ	→	Va
ఎ	→	E	స	→	Sa
ఏ	→	Ee	న	→	Na
ఒ	→	O	డ	→	Da
ఓ	→	Oo	ధ	→	Dha
బ	→	Ba	చ	→	Ca
భ	→	Bha	ఛ	→	Cha

Figure 4. Some of Telugu character set pronunciation.

III. EXPERIMENTAL DATA

The flow chart is mainly containing if two parts:

- Training of image.
- Testing of image.

Training and Testing of image:

STEP 1*-Read input image:*

The basic step that is required to start the procedure is to select or consider an image for the classification The image should be first checked of its format and made sure that it is saved in the library of the files of the software we are using for its execution along with function format. Image file saved in the software library: ‘imagename.jpg’

Image file read in the coding: **imread(['train/',d(i).name]);**

The image is named and saved for its usage as an input for simulation

STEP 2*-Resize Image:*

After selecting the input image, then resize the image i.e [60 64].

Image resize in the coding: **imresize(A,[60 64]);**

STEP 3*-Converting image into gray***:**

Image in jpeg form, is transformed into gray form for easy analysis. The extraction of the histogram values and zoning is done accurately in a gray image.

Converting to gray in the coding: **rgb2gray(A);**

STEP 4*-Converting image into binary:*

Converting image into binary: **im2bw(A,graythresh(A));**

STEP 5*-Creating Filter:*

Creating average filter**: fspecial('average')**

STEP 6*-HOG Feature Extraction:*

HOG features extraction in coding: **hogf(j,:)=extractHOGFeatures(bw);**

STEP 7*-Training and Testing of Bayesian Classifier:*

Creating bayesian classifier: **Bayesianobj.predict(feat);**

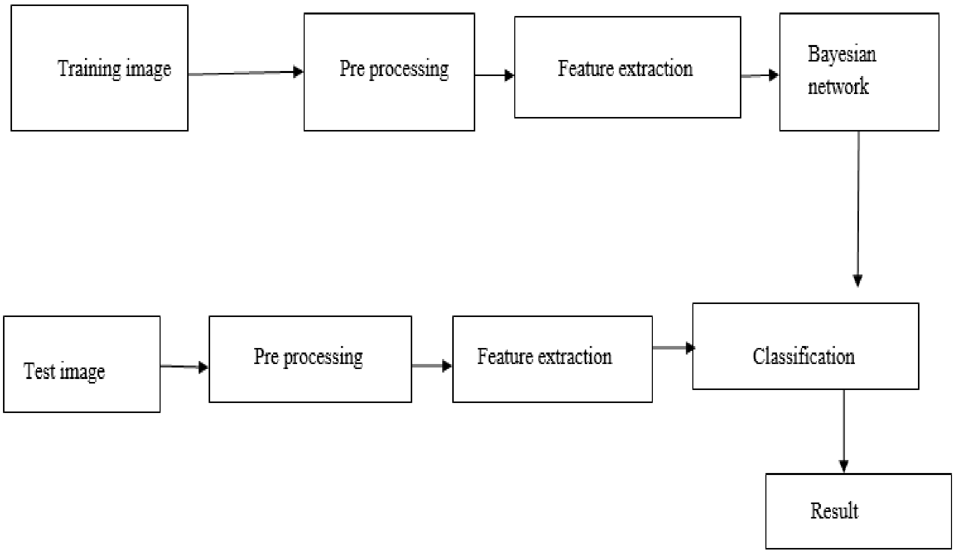


Figure 5. Block Diagram

A. *Pre-processing:*

The fundamental standard of recognition-based character division is to utilize a portable window is used to give the speculative divisions which are affirmed (or not) by the grouping. Picture obtaining and pre-processing are the two moderately basic stages, which are introduced first. Picture securing is at the picture representation level of pattern recognition(PR). It is the methodology of securing a machine representation of an archive to be perceived. A digital scanner is utilized at this stage to secure 200 dpi, 8-bits light black-level pictures. Pre-processing is at the picture-to-picture conversion level and repaying a low quality-unique and/or low quality-examining. There are two techniques to improve the obtained picture in the proposed framework, which are binarization and smoothing. Character division could be taken as recognition-based procedure. Dissection implies the decay of the image into a grouping of sub-images utilizing general characteristics. Each one sub-picture is dealt with as a character for recognition. It is worth saying that arrangement of characters is completed at a later stage. Projection analysis, connected component preparing, and white space and pitch discovering are a percentage of the regular dissection strategies utilized by OCR frameworks. These procedures are suitable for scripts which have large character spaces in between then. The dissection system is utilized for cursive scripts, a more `intelligent and particular dissection strategy for the specific script is required and no surety that high segmentation exactness might be accomplished.

The essential guideline of recognition-based character segmentation is to utilize a portable window of variable width gives experimental divisions which are affirmed (or not) by the order. Characters are by- results of the character recognition for frameworks utilizing such a guideline to perform character segmentation. The principle point of interest of this strategy is that it sidesteps genuine character segmentation issues. On a basic level, no particular segmentation algorithms for the particular script is required and recognition failures are principally because of disappointments throughout the classification stage. Consequently, more cursive script OCR frameworks utilize this procedure for enhancing the recognition correctness. This methodology is otherwise called without division recognition because of the virtual nonattendance of the character partition stage.

Binarization is an extraordinary instance of thresholding, of which there are just two states of yields in the ensuing picture, either dark or white. It diminishes the computational necessities of the framework and may empower evacuation of some noise. A document could be binarized comprehensively or adaptively. Unless the archive is considered as uneven shaded paper, worldwide thresholding is sufficient to do the binarization. Two worldwide thresholding calculations were concentrated on and executed.

B. Feature Extraction:

Histogram of Oriented Gradients

HOG is a feature [13] selection method gives the parameters as shown in Table I. The parameters of these features are flexible and are represented without any bias or variance.HOG starts by detecting the boundary or shape of an character and normalizing them into gray-scale. The gradient is calculated for the intensity of the detected characters. Weighted voting is applied to spatial and orientation cells. The normalization was performed on overall areas are made using L2 norm. The extracted features are rescaled in therange[0,1] by using the formulae:

$$Data = \frac{D - min}{max - min}$$

where D is a row vector containing the features, *min* is the minimum value of D and *max* is the maximum value in D.

BAYESIAN CLASSIFIER:

A. Bayesian Network:

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem:

There are two types of probabilities −

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

P(H/X)= P(X/H)P(H) / P(X)

Bayesian networks represent a set of variables in the form of nodes on a directed acyclic graph (DAG). It maps the conditional independencies of these variables (D. Heckerman, 1995). Bayesian networks bring us four advantages as a data modeling tool Firstly, Bayesian networks are able to handle incomplete or noisy data which is very frequently in image analysis (D. Bellot 2002). Secondly, Bayesian networks are able to ascertain causal relationships through conditional independencies, allowing the modeling of relationships between variables. The last advantage is that Bayesian networks are able to incorporate existing knowledge, or pre-known data into its learning, allowing more accurate results.

$$P(C = T|A = T) = \frac{P(C = T, A = T)}{P(A = T)}$$

where

$$P(C = T, A = T) = \sum_{S,W,B\in\{T,F\}} P(C = T)P(S) \times P(W|C = T)P(B|S, C = T)P(A = T|B)$$

$$P(A = T) = \sum_{S,W,B,C\in\{T,F\}} P(C)P(S)P(W|C)P(B|S,C) \times P(A = T|B)$$

V. EXPERIMENTAL RESULTS

The below given figures gives the procedure of the program and outputs of the program at certain levels, and explained clearly.

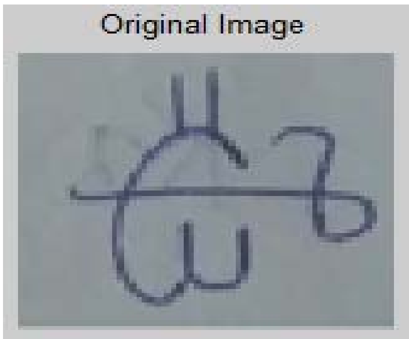


Figure 6. Input Image

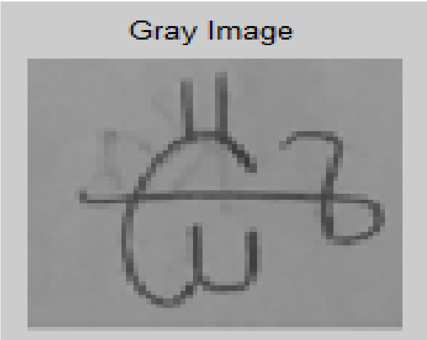


Figure 7. Input Image To Gray Image

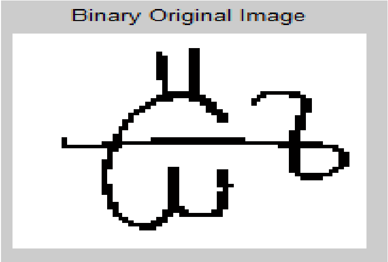


Figure 9. Gray image to Binary Image

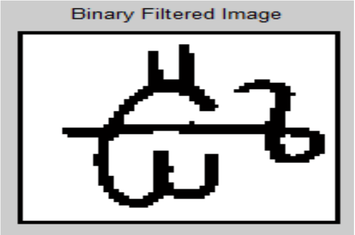


Figure 10. Binary Filtered Image

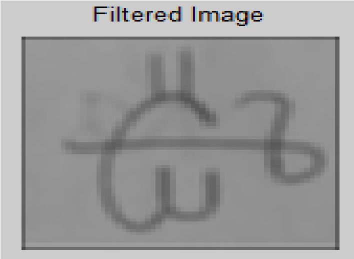


Figure 11. Filtered Image

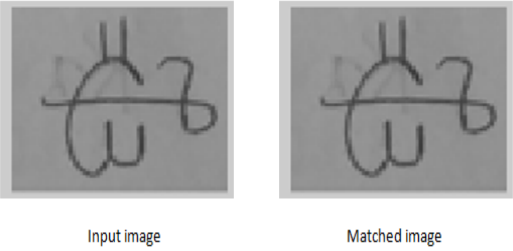


Figure 12. Input and Matched Images

V. CONCLUSION

This paper suggested a method for application of HOG and Bayesian Network classification for Telugu characters. The experimental results shows, the performance characteristics of the Bayesian network. Furthermore, Bayesian Network can produce higher recognition rates using Nearest Neighbourhood method. The efficiency of Bayesian network based model has its average of standard deviations of testing recognition rate less than the standard deviations of NN-based classifier. The recognition accuracy was calculate based on the number of positives and negatives in the training and testing dataset.

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to Dr.A.Raji Reddy and G.Srikanth providing us MATLAB software for simulation.We thank K.Mohana Lakshmi, Department of ECE for her consistent guidance for result oriented implementation. We would also like to thank all those who contributed towards the generation of our data set.

REFERENCES

[1] R. Plamondan, S.N. Srihari, “Online and offline handwriting recognition: A comprehensive survey”, IEEE Trans. On PAMI, Vol 22(1) pp 63 – 84, 2000.

[2] U. Pal and B.B. Chaudhuri, “Indian script character recognition: A survey”, Pattern Recognition, Elsevier ,Vol. 37, pp. 1887-1899, 2004.

[3] Technology Development for Indian Languages http://tdil.mit.gov.in/

[4] Lim, Jae S.,” Two-Dimensional Signal and Image Processing”, Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.

[5] P. K. Sahoo, S. Soltani, A.K.C Wong and Y C Chen, “A survey of Thresholding Techniques”, Computer Vision, Graphics and Image processing, vol 41, pp 233-260, 1988

[6] Otsu.N, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cybernetics, vol.9, pp.62-66, 1979

[7] L. Lam, S.W. Lee and C.Y.Suen, “Thinning Methodologies: A Comprehensive Survey”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.14.pp 869-885,1992

[8] Richard G. Casey And Eric Lecolinet “ A Survey Of Methods And Strategies In Character Segmentation”, IEEE Trans. On Pattern Analysis And Machine Intelligence, Vol 18, Pp 690-706,1996

[9] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 1996.

[10] Lajish V. L., “Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks”, Proc. 4th Int. National conf. on Innovations in IT, 2007, 188 – 192.

[11] Lajish V. L., “Handwritten character recognition using gray scale based state space parameters and class modular NN”,Proc. 4th Int. National conf. on Innovations in IT, 2007, 374 – 379.

[12] G. Raju, “Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients”, Proc. of 14th International conference on Advanced Computing and Communications, 2006, pp 217 – 221

[13] N. Dalal, B. Triggs, “Histogram of oriented gradients for human detection,” Conference on Computer Vision and Pattern Recognition, 2005, Vol. 1, pp. 886-893