# RECOGNITION AND LEXICON BUILDING OF TELUGU HANDWRITTEN CHARACTERS USING HOG FEATURES

Neerugatti Varipally Vishwanath[1], K.Manjunathachari[2], K. Satya Prasad[3],
[1]Research Scholar, Department Of ECE, JNTUK, Kakinada, Andhra Pradesh,India
[2]Professor, Department of ECE, GITAM University, Hyderabad Campus,India
[3]Professor, Department of ECE, JNTUK, Kakinada, Andhra Pradesh,India

**Abstract**

**This paper manages the status of Telugu characters on Telugu abuse connected arithmetic choices. Composed character recognition has various applications in post workplaces, perusing helps for visually impaired, library computerization and mixed media framework plan. Telugu original copies contain otherworldly writings and a large group of subjects like craftsmanship, medication, music, crystal gazing, law and physical science. There's partner degree innate 3D highlight for characters on Telugu known as profundity. This profundity is corresponding to the essayist's stylus weight connected at each pel reason. This 3Dfeature of every pixel in a photo is utilized to recognize the Telugu characters inside the blessing work. The picture is part of zones and along these lines, the promotion of the pel forces in each zone is utilized as a component vector to recognize the Telugu characters. According to the writing overview, the fame exactness for composed characters is a littler {amount} than an hour and moreover, appallingly less measure of work is accomplished for Telugu character recognition. Abuse the arranged strategy the best recognition precision got for Telugu characters is ninety-six.**

**Keywords: HOG futures, OCR, feature extraction, Bayesian classifier.**

## I. INTRODUCTION

The most serious issue with doing research on Handwritten Character Recognition (HCR) is the absence of a standard existing database for Indian scripts. For Latin numerals, there are a few standard databases, for example, NIST, MNIST, CEDAR, and CENPARMI. Little databases gathered in a research center environment are utilized as a part of the present work. By measuring the pixel organizes (X, Y, Z) of Telugu characters a database is produced. In 1870 the historical backdrop of character recognition began with the imagine of retina scanner. Later, with the innovation of the successive scanner in the year 1890, noteworthy changes occurred in the character recognition frameworks. In the mid-1900s distinguishing of characters was at first regarded as a guide for the visually impaired and was endeavored effectively by the Russian researcher Tyurin. The greater part of the exploration was contributed for printed content utilizing layout coordinating methodologies as a part of the early work. Less research need was given for manually written content as a result of the recognition challenges lying in them. For the most part, factual methodologies were utilized for perceiving manually written content. In the mid-1950s, with the assistance of electronic tablets, the information is caught utilizing x-y facilitates. Utilizing such developed information obtaining gadgets, analysts were inspired to chip away at web-based penmanship recognition frameworks. Later in the 1990s, a blend of picture handling and example recognition strategies with computerized reasoning techniques was abused. Nowadays, capable PCs with precise electronic devices like electronic tablets, scanners, cameras, there are effective strategies like neural systems, fluffy sets, and concealed Markov model being utilized for character recognition. For more than two thousand years palm leaves have been a well known written work medium in South and South East Asia. A gathering of palm leaves is accessible in traditional Indian dialects

like Tamil, Pali, and Sanskrit and in addition in Telugu [1], [2]. The original copies contain vital data and it is excessively troublesome, making it impossible to store them for a considerable length of time to come. There are a few variables for crumbling of palm leaves, for example, regular elements (temperature, dampness) and bug nibbles. There is no innovation to build the life of palm takes off. Along these lines, changing the information on Telugu into machine clear frame is the main method for saving them. Perceiving Telugu characters is a more perplexing errand than perceiving transcribed characters because of maturing and crumbling of Telugu [3]. In all the Indian provincial scripts there are compound characters separated from numerals, consonants, and vowels. Every compound character is the mix of two or more essential characters. There are a few issues in perceiving the written by hand characters, as a result of the distinction in composing style, size and state of the characters which differ from individual to individual. For each fundamental character on the Telugu, the pixel purposes of the character are recognized. For each pixel point the X, Y, and Z directions are measured, where Z directions are the profundity of the space relative to the weight connected by the author for every purpose of the Telugu character. Better recognition precision is gotten utilizing this 3D highlight on palm takes off. For Telugu characters, the most extreme variety of characters is found in the Y-bearing. For all the three planes of projection i.e. XY, YZ and XZ the proposed technique is connected and promote Nearest Neighborhood Classifier is utilized for order. The most noteworthy recognition precision was observed to be in YZ plane of projection [4],[5],[6].

## II. RELATED WORK

The pal et al. review report that recognition of characters should be possible utilizing two methodologies to be template format based and feature based methodologies [4]. Feature-based methodologies are recognizable for manually written character recognition contrasted with format based methodologies. As the qualities of the character shift from individual to individual, as far as composing style, size, and shape, layout based methodologies are not recognizable for manually written character recognition. At first, for essential characters feature based methodologies were utilized and for compound characters' format based methodologies were utilized as character recognition frameworks. In the present work, the measurable elements of the characters are considered for Telugu transcribed character recognition. Panyam Narahari Sastry et al [7], [8]. chipped away at palm leaves utilizing 2D relationship. The characters were ordered in view of the measure of closeness amongst them and the recognition exactness got is 90% in YZ plane of projection. Utilizing Radon change the recognition exactness for Telugu characters acquired is 89% in YZ plane of projection. Arica et al. concentrated on preprocessing methods, line, and word divisions. The historical backdrop of character recognition, the issues connected with on the web and disconnected manually written character recognition and advancements of character recognition frameworks were investigated in this paper. For online character recognition, the x-y directions were initially presented in 1950. Basic methodologies were initially presented for character recognition frameworks. Later shape recognition procedures were presented with no semantic data. In the mid-1990s, these procedures were consolidated with computerized reasoning techniques. Late works have been done on capable PCs, with effective hardware like scanners, cameras and so forth., and proficient calculations were actualized. Pawar Vijaya Rahul et al. contributed their work on Handwritten Devanagari scripts utilizing Learning Vector Quantization took after without anyone else Organizing Map (SOM). Two completely associated layers are utilized as a part of SOMs. The primary layer takes the qualities from the information design and the second layer totals the contributions to register a solitary winning unit. For preparing 2000 examples and for testing 200 specimens were utilized as a part of their work. The recognition precision was observed to be 85% to 95%. Keerthi Prasad et al. taken a shot at Kannada characters utilizing PCA (Principal Component Analysis) and DTW (Dynamic Time Wrapping) approaches. Utilizing PCA the real data is removed from information sets by ascertaining the Eigenvectors of the Covariance lattice. DTW gauges the wrapping separation for finding the comparability between any two-time arrangement. The base separation is utilized for the arrangement. They reported the recognition precision to be 87.5% utilizing PCA and 63.7% utilizing DTW approaches. Rituraj Kumar et al. contributed take a shot at online manually

written Kannada characters [9]. The measurable Dynamic Time Wrapping is utilized as a classifier that utilizations X, Y Coordinates, and their first subsidiary as components. The work is done for 295 classes of Kannada characters, numerals, accentuations and unique images [10], [11], [12]. The database was worked by gathering information from 69 unique clients and was caught utilizing the Tablet PC. The best recognition precision was observed to be 87.9%. Suresh Sundaram et al. have proved Online Tamil characters utilizing 2D PCA. To the routine components of PCA polynomial and quartile, elements are consolidated to frame a novel set for the examination. Promote, Eigenvectors are gotten from getting the diminished list of capabilities and for arrangement, Mahalanob is separation is utilized. This demonstrates a change in RA by 3% over the customary PCA. Venkatesh N et al. investigated the selection of classifiers for online transcribed Kannada and Tamil characters in progressive recognition [10]. In the primary stage PCA with NNC took after by Dynamic Time Wrapping is utilized to determine the mistaking characters for the second stage. To keep away from uncertainty among the mistaking characters for the optional classifiers, fitting weights are doled out to them. The most astounding weight among them is thought to be the last name. Utilizing single stage classifier they reported the recognition precision to be 76.5% and utilizing able classifier they reported it to be 92.2%.

## III. PROPOSED SYSTEM

TELUGU is one among the language in India and Asia, with a rise of quite 846 million speakers a square measure there. The sweetening of OCR particularly in Asian and Indian scripts is admittedly at a moderately promising stage, whereas it very is seen that OCR technology is in an exceedingly grown-up stage of growth for English and different Roman / Latin scripts. Among all the explanations square measure the sophistication of the writing, significantly in Telugu. Whereas probably ten thousand syllables square measure of times used at intervals the language, the writing units' square measure composed of mixtures of sixteen vowels and thirty-six consonants. A sensible OCR system for Telugu script was developed and planned by Negotal, wherever truly the quality of Telugu script and strategy for its reduction were planned. Their approach includes recognition and identification of associated parts. At intervals, this paper we have a tendency to propose a more robust and a strong recognition approach that initial uses the component distributions of the script and once exploits the structural data of Telugu writing.

Telugu is one among the prehistoric languages of India. The essential alphabet set of Telugu consists of sixteen vowels and thirty-six consonants. The Telugu language consists of straightforward and compound character shaped from basic alphabet set. Some characters in Telugu square measure created quite one connected symbol. Compound characters' square measure shaped by associate modifiers with consonants, of import in an exceedingly vast variety of attainable mixtures square measure there in an exceedingly Telugu script. Telugu includes a varied writing with a sizable amount of distinct character shapes composed of straightforward and compound characters evolved alphabet set. Telugu could be a phonetic language and written from left to right like West Germanic and additionally, in the Telugu language, every and each character represents a linguistic unit. Not a lot of work has been according to on the event of OCR systems for Telugu. Therefore, development of associated OCR system for Telugu is a very important space of current analysis. The identification method is extremely tough for Telugu as a result of it consists massive and numerous teams.

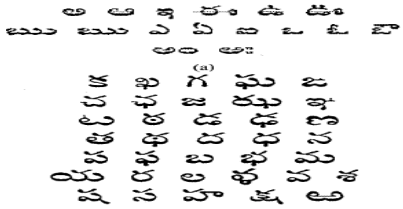


Fig 1. Telugu characters

The major steps concerned in recognition of characters embody, pre-process, segmentation, feature extraction, and classification

**Input Page Description**

i) Forms square measure created with applicable letters on the pages. It's created so mechanical extraction is probable.

ii) Every row contains ten characters excluding the last page every page confined ninety characters. Wholly 983 Telugu characters'

square measure to be overflowing with single handwriting.

iii) Every page envelopment a circle at the highest of the page that denoting the fact variety of the one set of handwriting.

iv) Every page containing the parallel and vertical block in spite of appearance right corner for in situ of the proper orientation of the check image.

v) Each vertical and horizontal line extends to every feature or row for characteristic the initial column or row positions within the look of the pel.

vi) Every block contains the breadth of fifty pixels and a length of sixty pixels.

**Preprocessing**

Pre-process can progress the image lucidity by cleaning- up the image and increasing the entrance worth. The preprocessed image can offer input to the next part. PRE-process the sequences of pre-processing steps square amount as trails

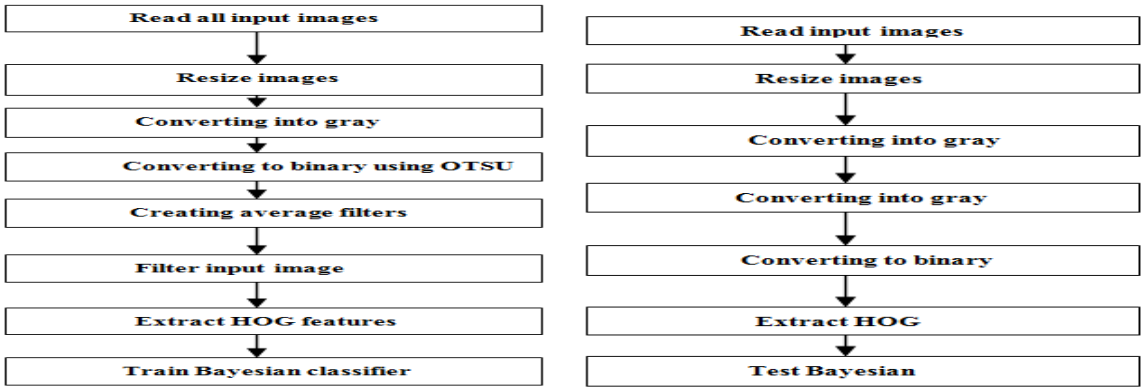



Fig 2: flowchart

**Noise Removal**

Noise is outlined as any squalor within the image because of external disturbance. The quality of written documents be contingent on varied factors as well as the excellence of paper, aging of documents, quality of pen, the color of ink etc. Some models of noise square measure salt and pepper noise, Gaussian noise. These noises will be removed to bound extent exploitation filtering technique.

**Thresholding:**

The job of thresholding is to extract the forefront (ink) from the background. Given a threshold, T between zero and 255, replace all the pixels with a gray level underneath or up to T with black (0), the remainder with white (1). If the brink is simply too low, it should cut back a number of objects and a few objects might not be perceptible. If it's too high, we tend to could embrace unwanted background info. The acceptable threshold worth chosen will be applied globally or natively. Otsu's algorithmic program is that the usually used international thresholding algorithmic program.

**SEGMENTATION**

Segmentation step contains line segmentation, word segmentation, and character segmentation. Traditions for character segmentations square measure supported

i) White area and pitch

ii) Projection analysis and

iii) Connected part labeling.

Standardization

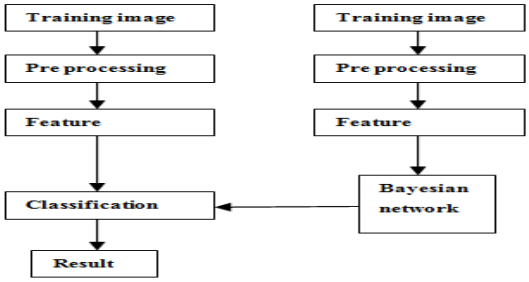
It's the scheme of changing the random sized image into a customary sized image. This size standardization avoids response category variation among characters. Bilinear, Bicubic interpolation techniques square measure some ways for size standardization.

**FEATURE EXTRACTION**

Options square measure a collection of numbers that capture the salient characteristics of the divided image. There are totally different feature extraction ways projected for character recognition.

## CLASSIFICATION

The feature vector obtained from the previous part is appointed a category label and recognized exploitation supervised and unsupervised technique. The information set is split into coaching set and check set for every character. Character classifier will be mathematician classifier, Nearest neighbor classifier, Radial basis performs, Support vector machine, Linear discriminant functions and Neural networks with or while not back propagation.




**III BLOCK DIAGRAM**

**Word-Level Recognition:**

The purpose of this phase is to confirm the recognition in word level by means of a graph representing the whole word replacements. This goal is achieved by computing the most probable word which exists in the lexicon. The system can also extend the lexicon if the probability of an unknown word exceeds a certain confidence level. In other words, their cognition is not restricted to the words in the lexicon. Given a segmented word, let G(V,A)be a multistage word graph, where the vertices represent the candidate segmentation boundaries between the characters and the edges represent the candidate character labels with the corresponding HMM probability measures. The word graph has the set of vertices $v_k \in V, k = 1, \ldots, K+1,$ where K is the total number of segments obtained from the segmentation algorithm. The source $v_1$ and the sink $v_{K+1}$ represent the left and right word boundaries, respectively. The edges in G are of the form $< v_k, v_{k+n} >_j$ and their associated costs are denoted as

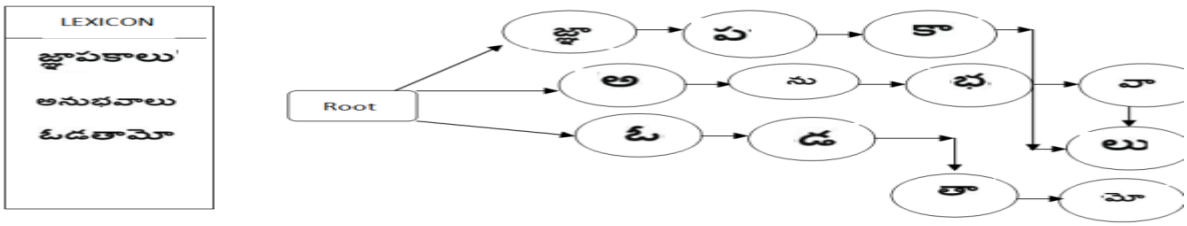
$$a_{k,n}(j) = -\log P(O_{k,n}|\lambda_i), \quad 1 \le j \le \Gamma_{k,n}, \quad 1 \le n \le N,$$

And the character labels are denoted as

$$\arg(a_{k,n}(j)) = i, \quad 1 \le i \le C,$$

Where $\Gamma_{k,n}$ is the total number of edges between vertices $v_k$ and $v_{k+n}$ which is calculated in the HMM ranking stage and C is the number of characters in the alphabet. Fig. 11b indicates the word graph of Fig. 11a, with the most probable candidates, taken from the list $(0 \le \Gamma_{k,n} \le 2, \ for\ 1 \le k \le K)$ for N=3 and K =10. The shortest path from source to the sink can be obtained by minimizing the cumulative $-logP(O_{k,n}|\lambda_i)$ measure

.




However, the sequence of characters lying on the edge of the shortest path may not form a valid word in the lexicon. At this point, lexicon information is necessary for achieving further improvements in word-level recognition rates. For this purpose, most of the available methods match the unknown word graph against the words in the lexicon using dynamic programming technique and ranking every word in the lexicon. Then, the word with the highest rank is chosen as the recognition result. Recognition rates are quite satisfactory for small size vocabularies. However, as the size gets larger, the highest rate may not always correspond to the correct word. The time complexity of these algorithms is $O(K \times £)$ to match K segments to a lexicon of size £. In some studies, the lexicon is stored in a hash table [6] or in a tree structure . If it is represented as true, the complexity is of order $O(K \times \Lambda)$, where K is the number of segments and _ is the number of

nodes in the tree. In this study, the lexicon is also stored in a trie data structure.

## IV. CONCLUSION

Telugu character recognition is a generally less inquired about range according to the writing. Because of modifiers and number of Telugu characters, it is hard to have high recognition precision. Characters like Va, Ma, Pa, Ya are fundamentally the same as and confound the PC in the recognition. Because of less number of clients and no standard database as reported in the writing, 4620 pictures for PLCR is produced in XY, YZ and XZ planes of projection. Out of these pictures, 924 are utilized for testing and alternate pictures are utilized for preparing. The proposed Telugu character recognition technique yielded 96% recognition exactness for essential Telugu characters in YZ plane of projection. This calculation can be enhanced and utilized for multilingual scripts as a major aspect of future work. This work can be stretched out to mechanize the Telugu character recognition framework.

## REFERENCES

[1] P.N. Sastry and R. Krishnan, "Isolated Telugu character recognition using radon transform, a novel approach," in *World Congress on Information and Communication Technologies(WICT)*, 2012, pp. 795–802.

[2] P. N. Sastry, R. Krishnan, and B. V. S. Ram, "Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach," *J. Applied Engine. Sci*, vol. 5, no. 3, pp. 22–32, 2010.

[3] U. Pal and B. Chaudhuri, "Indian script character recognition: a survey," *Pattern Recognition*, vol. 37, no. 9, pp. 1887–1899, 2004.

[4] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, no. 2, pp. 216–233, 2001.

[5] P. N. Sastry, R. Krishnan, and B. V. S. Ram, "Telugu character recognition on palm leaves- a three-dimensional approach,"*Technology Spectrum*, vol. 2, no. 3, pp. 19–26, 2008.

[6] U. Pal, R. Jayadevan, and N. Sharma, "Handwriting recognition in Indian regional scripts: A survey of offline techniques,"*ACM Transactions on Asian Language Information Processing(TALIP)*, vol. 11, no. 1, pp. 1:1–1:35, 2012.

[7] P. N. Sastry and R. Krishnan, "A data acquisition and analysis system for Telugu documents in Telugu,"in *Proceeding of the Workshop on Document Analysis And Recognition*, ser. DAR '12. New York, NY, USA: ACM, 2012, pp. 139–146. [Online].Available:http://doi.acm.org/10.1145/2432553.2432578

[8] G. Keerthi Prasad, I. Khan, N. Chanukotimath, and F. Khan, "On-line handwritten character recognition system for Kannada using principal component analysis approach: For handheld devices," in *World Congress on Information and Communication Technologies (WICT)*, Oct 2012, pp. 675–678.

[9] R. Kunwar, K. Shashikiran, and A. Ramakrishnan, "Online handwritten Kannada word recognizer with unrestricted vocabulary,"in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2010, pp. 611–616.

[10] S. Sundaram and A. Ramakrishnan, "Two-dimensional principal component analysis for online Tamil character recognition, "*The Proceedings of 11th ICFHR*, pp. 88–94, 2008.

[11] V. N. Murthy and A. G. Ramakrishnan, "Choice of classifiers in hierarchical recognition of online handwritten Kannada and Tamil akharas," *Journal of Universal Computer Science*,vol. 17, no. 1, pp. 94–106, Jan 2011.

[12] A. Senior and A. Robinson, "An off-line cursive handwriting recognition system," *Pattern Analysis and Machine Intelligence,IEEE Transactions on*, vol. 20, no. 3, pp. 309–321,Mar 1998.34