International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 4th Oct 2021. Link
:http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=ISSUE-11

DOI: 10.48047/IJIEMR/V10/I11/50

Title **A comparative study on Telugu Optical Character Recognition for Printed Text Documents**

Volume 10, Issue 11, Pages: 309-315

Paper Authors
Srinivasa Rao Dhanikonda, Subhash Chandra N

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

A comparative study on Telugu Optical Character Recognition for Printed Text Documents

[1]Srinivasa Rao Dhanikonda, [2]Subhash Chandra N
[1]Research Scholar, CSE, JNTU Hyderabad.
[2]Professor, CSE Department; CVRCOE,JNTU Hyderabad.
[1]srinivasarao.dhanikonda@gmail.com ,[2]subhashchandra.n.cse@gmail.com

Abstract:

Character recognition is a major area of research. This is the conversion of the scanned image into a machine/human readable format. This term refers to scanned images or documents. An OCR (Optical Character Recognition) can convert any image text to machine-readable text. OCR can be used for many purposes. It is useful in understanding scripts in any language. The text is also converted with high accuracy. MS Office is used to edit any image text that has been through OCR processing. OCR has made it unnecessary to digitize image text. Deep earning is another advanced technology used extensively in OCR. You can choose to have the documents printed either as text or handwritten. Handwriting can be either constrained or free. The image file will not contain text that is in a computer format. OCR for English vocabulary is well-constructed. OCR is required for Indian languages. OCR is difficult for Telugu languages as vowels and consonants play an important role in the formation of words, along with vattus or gunithas. The compound character could be made up of a combination of vowels or consonants. This paper examines the research that has been done to date on the OCR methods for the Telugu Language.

Key Words: OCR (Optical Character Recognition), handwritten, scripts, Indian languages

I. Introduction

Sometimes scanning documents is required. These scanned files must be edited as necessary. This task is handled by an optical character recognition (OCR). OCR converts text-rich images into a format that is easily editable with MS Word. From the scan, it creates a digital file. It can also be used for exporting data. It makes it easier to search for files or edit text. It is smaller and easier to store. It is computationally faster and more efficient [1]. To save scanned documents, a bitmap file can be used. Although the scanned document is easily read by humans, a computer interprets it as a series of white and black dots. The computer system attempts determine if the sequences of dots correspond with any letter or alphabet. OCR allows you to easily edit and use the document. After OCR processing, a word processor can be used for editing the document. It is also accessible to visually impaired persons. Screen readers can read machine-readable text even for the blind. OCR is extremely useful in office work because most office tasks can easily be completed with the help of scanned documents [2]. OCR accuracy can reach 93% which is very impressive. OCR has not improved the quality language scripts. Nearly 18 languages are recognized in India. This paper is a major step towards the development of an OCR system for Telugu. There are 52 alphabets and 36 consonants in Telugu. 16 vowels. The Telugu alphabets can be written in a different way than the English alphabets. Telugu script can be broken down into simple and complicated character sets. A character can be considered simple if it has one cipher. If a character has only one cipher, it can be considered simple. Each character can be broken down into individual ciphers in Telugu, which can then be analyzed once recognized [3]. Telugu has 104 ciphers. Practically, there are only 100 ciphers. This manual approach is time-consuming and tedious. This method is problematic because you can use ciphers in multiple forms. Another Telugu OCR system can segment and recognize the scanned image using glyphs. It was able to test around 5000 pages from 30 books. It was more efficient than the Telugu OCR Drishti method.

Characteristics of Telugu Script

India is home to many languages, and approximately 150 of these languages have been documented. The Constitution of India lists 22 languages that are commonly used in the Eighth Schedule. Each language has its own script. People speak many other languages in daily life. Many languages share a common script, while some languages don't have any script at all. The Brahmi script is the basis of most Indian scripts. The major language families are Dravidian (Indo-Aryan), Astro-

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Asiatic, Tibeto-Burman, and Astro-Asiatic. Telugu and Tamil are part of the Dravidian language family. Telugu is spoken mainly in South India's Andhra Pradesh and Telangana. It also appears in many parts of Karnataka and Tamil Nadu, where there are approximately 110 million people. A syllable is an organized sequence of sounds. Akshara, the spoken unit of the language, has a direct correspondence with the written word. The script does not require spelling rules because there is an association between spoken and written syllables. There are 38 consonants and 18 vowels in the script (Fig. 1). Two of the 38 consonants and vowels (lu, luu and dza), are no longer in use. The character set also includes a 'halant', but they are not independent.

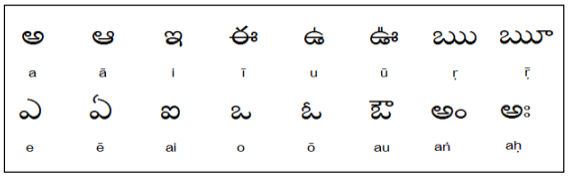

Figure 1(a): Telugu vowels

Figure 1(b): Telugu consonants

To form consonants, the base consonants that end in halant are combined together with the first vowel (/a/). Base consonants do not have a vowel sound. They are not used often in modern languages, but they do appear in a few words ends and when words from other languages are written in Telugu.

Combining consonants with vowels creates consonant vowel forms. (Fig. 2). A vowel modified consonant is one that has a base consonant and a vowel sound. They come in distinct shapes. In most cases, the vowel modifiers (gunintalu) replace the inclined top portion of consonants. Vowel modifiers, or matras, are attached in different places. They can be written at the bottom or right and form separate shapes. Except for the vowel /a/, each vowel contains the appropriate vowel modifier. The consonants are formed by combining the vowels. The matras are not independent entities.

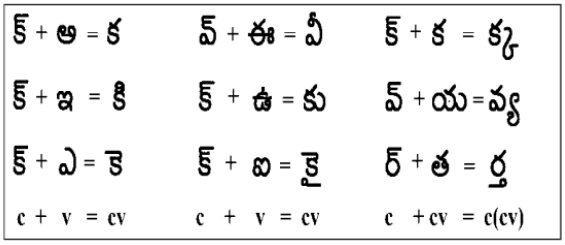

Figure 2: Formation of consonants, vowel modified consonants and conjunct consonants

Each OCR implementation involves a series of pre-processing and then the actual recognition. There are many factors that influence the number and type of pre-processing algorithms used on scanned images. These include the date of the document, paper quality, resolution, resolution and the format and layout of images and text. These are typical stages of pre-processing.

Binarization: Binarization refers to converting grayscale images into binary images. It is essential that you identify the objects of particular interest in any image analysis or enhancement problem. Binarization allows you to distinguish the background (text), from the foreground. You can binarize an image by first defining a threshold value that will determine the intensity. Then, convert all intensity values above this threshold to one intensity and all intensity levels below it to the selected intensity. Most reports indicate that binarization can be performed locally or globally. A single intensity value is applied to the whole picture using global techniques. You may apply various intensity values across portions of a picture using adaptive or local thresholding. The proximity of the pixels to which the thresholding is applied determines the threshold values.

Noise Removal: Noise can often be found in scanned documents due to the printer, scanner quality, age, and print quality. It is important to remove this noise from the image before processing it. Low-pass filtering the image is a common approach that can be used for processing later. A

filter is required to remove as much noise from the signal as possible in order to reduce it.

Thinning: A process called skeletonization or thinning is where an object's representation is created that is one pixel wide. It preserves an object's connections and ends points (Gonzalez & Woods, 2002). To make an image easier to recognize and analyze, it reduces its information. This makes it easier to identify relevant features. Figure 3 shows an example image of a photo before and after thinning. There are many different existing thinning methods that have been created Hilditch is the most often used algorithm. There are many variations.

Skew Detection and Correction: A few degrees of tilt (skew) can be expected when a human or machine operator is feeding a document into the scanner. Text lines in a digital picture connect with horizontal directions at an angle known as the skew angle. A variety of skew estimating techniques exist. There are two types of skew estimation methods. The projection profile of a document is the first. Another one is based on the clustering of related neighbor components. Skew estimation can also be done using techniques based on the Fourier transform and Hough transform. Chaudhuri & Pal (1997) provide a comprehensive overview of the various skew correction methods. The projection profile is a popular method of skew detection. Horizontal projection profiles are A one-dimensional array in which each element represents the number black pixels per row. The horizontal projection profile is useful for documents that have horizontal text lines. It features a peaked that is equal in width to the character height, and valleys that are equal in width to the spacing between lines. Since scan lines align with text lines, the projection profile is at the correct skew angle. It has a maximum height peak for line spacing valleys and text.

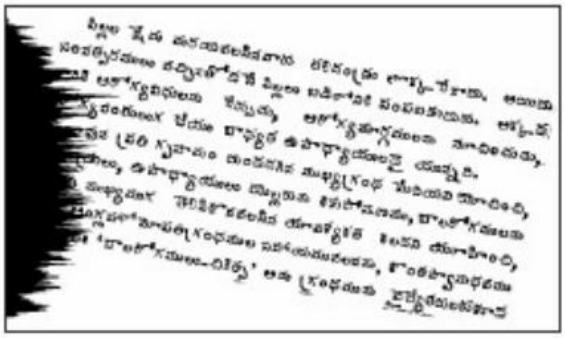

Figure 3(a): Text image skewed

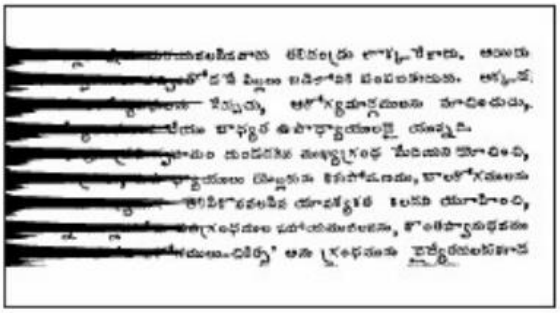

Figure 3(b): text image straightened

Line, Word, and Character Segmentation: After correcting the tilt, the text must be divided into lines. Each line must then be broken into words. Finally, each word must be broken into its constituent characters. To extract lines from a document, horizontal projection is the most common method. Horizontal projection of a document image will produce lines that are separated and not tilted. Figure 3(b) shows the separation of peaks and valleys. These serve as separators for the text lines. These valleys can be easily identified and used to locate the boundaries between lines. Figure 5 shows an image with 3 text lines (left) and 3 segments (right), which were created using horizontal projection profiles. Vertical projection profiles also give column sums. By looking at the horizontal projection profile, you can separate lines by looking for minima. To separate words, you can also look at the minima on a vertical projection profile. Vertical projections of line images are useful for extracting words and individual characters. Vertical projection profiles cannot be used to segment adjacent or overlapping characters in a word ("kerned characters") by using zero-valued valleys. This problem requires special techniques.

Feature Extraction and Selection: Feature extraction is the process of determining the form of an underlying character using a set (or feature) of parameters. It is important to select features that can help distinguish between characters. Thinned data can be used to identify features like straight lines, curves, or significant points along curves.

Classification: Based on the extracted features and their relationships, the classification stage of an OCR process assigns labels for character images. This is where OCR finally recognizes characters and outputs them as machine editable text. The most widely used classification method is template matching. Template matching is based on individual pixels in an image. A set of prototypes or images is used to classify an input image.

Recognition is provided by the template which matches closest to the unknown. Following feature extraction, classification strategies are based mainly on the identification of the nearest neighbor pixel. The similarity between images is determined by the distance between vectors. The two most popular classifiers are binary-tree and nearest-neighbor. The Euclidean distance measurement is the most commonly used distance calculation.

II. Literature survey

Many authors proposed many techniques for Telugu OCR, in this section we are going to discuss some of those proposals and their achievements

Telugu handwritten characters recognition was worked on by P. N. Sastry and his colleagues in 2010. They employed zonal-based feature extraction from scanned handwritten pictures to accomplish their aim. They concentrated on sorting and categorizing the photographs of the characters into certain areas. The statistical characteristic was then determined for each of the zones that had been selected. Their strategy had a 78 percent success rate.

Online character recognition has been stressed by Swethalakshmi et al. (2006). They are mainly concerned with the extraction and classification of elements that may be used to identify handwritten characters. After that, they put them on a list. They employed support vector machines to develop the stroke recognition engine. Accuracy ranged from 74.62% to 99.91%, depending on the character groupings that were being analyzed.

The 43 Telugu characters were studied by Pradhan et al. (2012) in trie format. Syntactic PR is now being used for character recognition. They were interested in how discrete patterns may be combined to generate a sequence triangle. The greatest results were achieved by using an approximate match rather than an exact match.

Driven by the Telugu Language Handwritten Numbers 0-9 by Jagan Mohan Reddy et al. (2019). Throughout the ages, they have enjoyed remarkable success due to their concentration on numbers. More than 95% of characters were correctly transcribed. For the number 3, they had an accuracy rate of 98%. The accuracy of the numerals 7, 9, and 4 was lower than that of other numbers because of their ambiguity.

Using optical character recognition, Prameela et al. (2017) has been working on handwritten Telugu characters. During the pre-processing stage, they concentrated on median filtering, which removes the border edge pixels points. Three three-by-three squares make up each character. In order to determine the euclidean distances, the closest pixel in that zone was used. To categorize, they turned to methods like quadratic discriminate and support vector machines.

Prasad et al. (2016) is currently working on handwritten Telugu characters using both static and adaptive zone methods. To extract the geometrical features, a genetic algorithm was used. To extract density-based features, the distance was used. The first 11 characters were accurately identified with 100% accuracy. They also achieved 82.4% accuracy. The accuracy of the second approach was 100%, while the accuracy for the 50 remaining characters was 88.8%.

Andrew et al. Andrew et al. (2017) suggested a method to identify handwritten Telugu scriptwriters. They have compiled the handwritten data from more than 150 people to create a database. They gave it the name “IIITSTHDD”. They

compared features from their data set to determine if noise effects were present.

Using CNN with the "kha" dataset, N. B. Muppalaneni (2010) developed the Telugu "guninthalu" a prediction system for handwritten Telugu. Vowels and consonants combine to form the Telugu composite letter. It follows the same rules. Handwritten images are a difficult task. One of the most challenging tasks is to identify compound characters. The author completed this task with an accuracy of 79.61% in test and 96.13% in training. This feat was achieved by the author's machine-learning model, which uses CNN architecture.

III. Research study in identification of character regions

Word region segmentation: English has a well-developed segmentation system for documents. Maximally stable extremal regions are used to separate characters in English in many papers. MSER cannot directly be applied to Telugu because most of the dheergams are separate. Minor changes were made to MSER in order to account for vatus and dheergas.

Character level segmentation: The Connected Components algorithm in Image Processing allows us to segment every character of the word. After the image has been binarized, the algorithm is used to separate the letters and vattus. The components are groups of binary pixels that contain the letters and vattus. The components are also stripped of little blobs. Some vattus in Telugu do not connect to the base letter. We measured the overlap distance in horizontal and vertical directions to connect the base letter and its vattu and then grouped them.

Author	Methodology implemented	Results
P.N. Sastry et al. [7]	Using the Zonal Features of a character in an image, Telugu Character Recognition have performed.	The character in the picture has been identified 78 percent of the time.
Swethalakshmi et al. [8]	Handwritten characters recognised with the assistance of a series of strokes that represents them	They were able to get a range of 74 to 99 percent accuracy for different characters.
Pradhan et al. [9]	The trie data structure was utilised to work with the Telugu language's 43 characters.	They obtained good results and accuracy with their architecture
Jagan Mohan Reddy et al. [10]	Worked on the 0 to 9 handwritten numbers written in the Telugu language	Obtained an accuracy of more than 95% for every digit
Prameela et al. [11]	Offline work was done on it. Optical character recognition is being used to read handwritten Telugu characters.	They had favourable outcomes with the strategy they presented.
Prasad et al. [12]	Adaptive and static zoning approaches were used to create the handwritten Telugu characters.	For the first 11 characters, they achieved an accuracy of 100% and 82.4 percent, while the remaining 50 characters with the second technique achieved an accuracy of 100% and 88.8 percent.

Muppalaneni [15]	Worked on Telugu Handwritten Guninthalu	For compound characters, achieved a 79% test accuracy and a 96% training accuracy.

Table 1: Comparative study of different methodologies

IV. Conclusion and Future Direction

Because of the wide range of applications, the Telugu language's optical character recognition (OCR) is a current research topic. OCR for printed text, which is still in the early stages of development, necessitates improvisation at the time of processing. It also has to be capable of dealing with broken characters and segmentation problems. None of the approaches listed above can guarantee an accuracy of 99 percent. Because of the segmentation and categorization of Telugu characters, it isn't easy to recognize them. We looked at a variety of classification and segmentation frameworks for this project.

On the other hand, more precision is required to enhance segmentation algorithms. It is necessary to design a hybrid model to address the issues indicated in the present study. It is necessary to enhance segmentation algorithms so that each character may be segmented according to its "vatu" or "gunintham". The accuracy of the network may be further enhanced to increase the accuracy of the classifier. The same optical character recognition study will be undertaken again, but this time by offering a dynamic line segmentation approach to extract Telugu characters from document photos to address issues with current systems. This work is a continuation of the Binarization, and Skewness correction work done before. Projection Profiles are a kind of projection profile (Horizontal & Vertical). Calculate the distance between pixels in a projection profile, the distance between words in word segmentation, and the distance between characters.

References:

[1] T.V. Ashwin and P.S. Sastry. "Font and size independent OCR for printed Kannad,a documents using SVM classifier". In Proc. of Open Research Forum, Fifth ICDAR, pages 14-18. 1999.

[2] R.L. Brown. "The fringe distance measure: an easily calculated image distance measure with recognition results comparable to Gaussian blurring". IEEE Trans.System Man and Cybernetics, 24(1) :111-116, 1994.

[3] R.C. Gonzalez and R.E. Woods. "Digital Image Processing", Addison-Wesley, 1993.

[4] A. Negi, C. Bhagvati, and B. Krishna, "An OCR System for Telugu," in ICDAR, 2001.

[5] P. P. Kumar, C. Bhagvati, A. Negi, A. Agarwal, and B. L. Deekshatulu, "Towards improving the accuracy of telugu ocr systems," in 2011 International Conference on Document Analysis and Recognition, pp. 910– 914, Sep. 2011.

[6] C. V. Lakshmi and C. Patvardhan, "A high accuracy ocr system for printed telugu text," in TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, vol. 2, pp. 725–729 Vol.2, Oct 2003.

[7] Sastry, P. N., Krishnan, R., and Ram, B. V. S. 2010. "Classification and identification of Telugu handwritten characters extracted from palm leaves using decision tree approach". J. Applied Engn. Sci. 5, 3, 22--32.

[8] H. Swethalakshmi, Anitha Jayaraman, V. Srinivasa Chakravarthy, C. Chandra Sekhar. "Online Handwritten Character Recognition of Devanagari and Telugu Characters using Support Vector Machines". Tenth International Workshop on Frontiers in Handwriting Recognition, Université de Rennes 1, Oct 2006, La Baule (France).

[9] Pradhan, S. K., & Negi, A. (2012). "A syntactic PR approach to Telugu handwritten character recognition". Proceeding of the Workshop on Document Analysis and Recognition - DAR '12. doi:10.1145/2432553.2432579

[10] D. Jagan, Mohan Reddy and A. Vishnuvardhan Reddy, "Recognition of handwritten characters using deep convolutional neural network", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 6S4, pp. 314-317, 2019.

[11] Prameela, N., Anjusha, P., & Karthik, R. (2017). "Off-line Telugu handwritten characters recognition using optical character recognition". 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA). doi:10.1109/iceca.2017.8212801

[12] Prasad, S. D., & Kanduri, Y. (2016). "Telugu handwritten character recognition using adaptive and static zoning methods". 2016 IEEE Students' Technology Symposium (TechSym). doi:10.1109/techsym.2016.7872700

[13] C. Andrew, S. Reddy, V. Pulabaigari and U. Pal, "Text Independent Writer Identification for Telugu Script Using Directional Filter Based Features," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 65-70, doi: 10.1109/ICDAR.2017.330.

[14] Convolutional neural network, https://en.wikipedia.org/wiki/Convolutional_neural_ network, Accessed on 25 Januray 2021.

[15] N. B. Muppalaneni, "Handwritten Telugu Compound Character Prediction using Convolutional Neural Network," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-4, doi: 10.1109/icETITE47903.2020.349.

[16] A. Negi, C. Bhagvati, and B. Krishna, "An OCR System for Telugu," in ICDAR, 2001.

[17] P. P. Kumar, C. Bhagvati, A. Negi, A. Agarwal, and B. L. Deekshatulu, "Towards improving the accuracy of telugu ocr systems," in 2011 International Conference on Document Analysis and Recognition, pp. 910– 914, Sep. 2011.

[18] C. V. Lakshmi and C. Patvardhan, "A high accuracy ocr system for printed telugu text," in TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, vol. 2, pp. 725–729 Vol.2, Oct 2003.

[19] D. K. Rao and A. Negi, "Orthographic Properties Based Telugu Text Recognition Using Hidden Markov Models," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 05, pp. 45–50, Nov 2017.

[20] G. Chen and D. Needell, "Compressed sensing and dictionary learning," Finite Frame Theory: A Complete Introduction to Overcompleteness, vol. 73, p. 201, 2016.

[21] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," Journal of Machine Learning Research, vol. 12, no. Nov, pp. 3259–3281, 2011.

[22] C. N. Duong, K. G. Quach, and T. D. Bui, "Are sparse representation and dictionary learning good for handwritten character recognition," in 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 575–580, Sept 2014.

[23] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," Network: Computation in Neural Systems, vol. 7, no. 2, pp. 333–339, 1996. PMID: 16754394.

[24] Sukumar Burra1, Amit Patel2, Chakravarthy Bhagvati1 and Atul Negi1, "Improved Symbol Segmentation for TELUGU Optical Character Recognition", ISDA 2017.