International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Printed Gujarati Character Recognition: A Review

Nidhi Desai[1], Prof. Mosin Hasan[2], Prof. Prashant Swadas[3]

[1]Student (M.Tech), Dept. of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India
[2]Professor, Dept. of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India
[3]Professor, Dept. of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India

---

Abstract - *For years, people have been fascinated by optical character recognition (OCR).Optical Character Recognition (OCR) is a method of transforming a handwritten or printed text image, photo, or scanned document into machine-encoded text using mechanical or technological means. Western scripts can be read with a variety of commercial OCR systems. However, Indian scripts, such as Gujarati script, do not have adequate work. On the other hand, except for the Gujarati script, there are few OCRs available for several Indian scripts. In today's world, OCR and HCR are widely utilized for data input from printed or handwritten records. A survey of text recognition strategies for Gujarati script is presented in this study. The Gujarati script is used to classify this survey. The goal of this study is to summarize the existing research on the topic of OCR. It gives an overview of many aspects of OCR and examines related solutions for overcoming OCR problems.*

Keywords: **OCR, character recognition, Gujarati characters, Gujarati script, Online Recognition, Offline Recognition.**

1. INTRODUCTION

India is a country that speaks several different languages and uses several different scripts. In India, there are 22 official languages written in 12 scripts. Kashmiri, Devanagari, Gujarati, Bengali, Oriya, Kannada, Telugu, and Malayalam are only a few of the Indian scripts that are evolved from the Brahmi alphabet [1]. The majority of the scripts are written from right to left [1]. In recent years, the growing use of physical papers has prompted the development of electronic documents to facilitate document interchange and storage. In the computer age, optical character recognition is an important and useful technique. Scanned photos are read and converted into a digital character-based format using optical character recognition (OCR) software [1][13].

Because of the wide range of languages, fonts, and styles in which text can be written, as well as the complexities of linguistic rules, OCR is a difficult challenge to solve. Gujarati script is used in a large variety of printed and handwritten documents. [1]From a historical and legal standpoint, as well as for effective transmission, such records must be preserved in digital format. Scanning is one of the most effective methods for converting paper documents to digital format. Editing, searching, and extracting information from scanned document pictures, on the other hand, is challenging [1]. As a result, obtaining information from a scanned page is a critical task. The recognition-based and recognition-free approaches to IR (Information Retrieval) from documents are the most common. The recognition-based method converts a document image into a text (ASCII) document using an OCR (Optical Character Recognition) system. The recognition-free technique treats a word image as a query image and performs the IR task by comparing the query word image to the document word pictures directly [1].

Overview of the Text Recognition system

There are two types of text analysis problems: 1) text recognition and 2) text matching. Recognizing the word/character from handwritten and printed documents is achievable in two ways in the text recognition task: 1) Word recognition in the offline world, 2) word recognition in the online world (Shown in fig.1).

Offline text recognition is concerned with the recognition of words after they have been written by individuals, usually on a piece of paper or a sheet of paper. Offline text recognition is the process of recognizing text that has been scanned from a piece of paper (or sheet) and saved digitally in formats like .pdf, .jpeg, .png, .bmp, and so on.

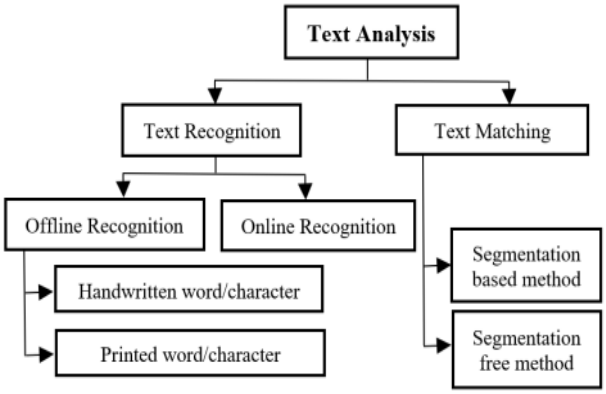

Fig -1: Types of word/character recognition [1]

The writing is done with a digital pen on an electronic notepad, tablet, or laptop in online text recognition (touch).In-text matching, information is retrieved without explicitly identifying characters or words [1].

The architecture of Optical Character Recognition

The scanned documents are converted into a machine-editable text format using the OCR process. The OCR system takes an image input and outputs machine editable, searchable, and translation formats. The process of recognizing text in an image entails several steps (Shown in fig.2),

Pre-processing is a step to remove the noise and correct the format of the documents. It is applied to characters before extracting the features and performing classification on them. It has many operations namely skew correction, normalization, noise removal, binarization, thinning, thickening, etc.

Binarization: Binarization is the conversion of a grayscale image to a binary (black and white) image with only 0 and 1 as images using the needed level of thresholding, the binarization technique is commonly used to separate foreground and background.

Noise Removal: Digital images consist of a variety of noises. These noises are required to be removed from an image for better processing. The morphological operation, Median filter, and Weiner filter are used to remove noise from an image. Median filter reduces blurring of edges.

Thinning and Filling: Smoothing implies both Filling and Thinning. Thinning reduces the width of character while Filling eliminates gaps, small breaks, and holes in digitized character

Normalization: To obtain characters of uniform size, rotation, and slant Normalization is applied to the image. To improve the accuracy of character recognition Normalization reduces shape variation.

Skew detection and correction: During the digitization of the document page it is often that image is not aligned correctly or it may be happening by a human while writing a document. To make incorrectly align Skew detection and correction technique is used. Skew detection technique can be classified: Analysis of Projection profile, Hough transform, clustering, connected component, and correlation between line techniques.

Segmentation A character in a document is segmented. Documents are first split byline, then by word, and finally by character.

Segmentation of Lines: The detection of text lines was done by scanning the input image horizontally. To create the row histogram, the frequency of black pixels in each row is counted. A boundary between two successive lines is defined as the point where the number of pixels in a row is zero[15].

Fig -2: Line segmentation [5]

Segmentation of words: To segment words, To create a column histogram, the number of black pixels in each column is determined. A word in a line is defined as the segment of the line with continuous black pixels. If no black pixels are identified in a vertical scan, the spacing between words is taken into account. As a result, distinct words in various lines are distinguished. As a result, the image file can now be thought of as a set of words [14].

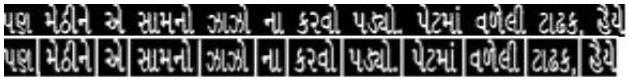

Fig -3: Word segmentation [5]

Segmentation of characters: On each of the divided words, a column histogram is used for character segmentation. In a nutshell, the separators between the characters are spaces between the characters. To create the column histogram, the frequency of black pixels in each column is counted [14]. A character boundary is defined as the space between two consecutive characters where the number of pixels in a column is 0. However, when using this method, a difficulty arises when a "g"-like the character is separated from the mantra (gy), resulting in a half-character in Gujarati. As a result, here are some things to take [15]:-

Separate the characters
Examine the following separated character's sizes.
If it's half the size of the character, merge the two characters and they'll become one.

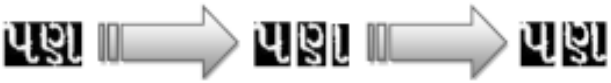

Fig -4: Character segmentation [5]

Feature extraction is employed, which makes pattern categorization simple using a formal procedure. There are specific types of parameters that may be derived from numbers and characters utilizing different feature extraction approaches [15].

The goal of principle component analysis (PCA) is to reduce the dimensionality of a data set with many connected variables while keeping as much information as possible.

$$HD(x,y) = \sum_{i=1}^{n} |Xi - Yi|$$

Classified characters:

Fig -5: Output of classification

An Optical Recognition System's classification step uses the feature gathered in the previous stage to identify the text segment according to the current rule. Decisions can be made based on the features collected using decision rules.

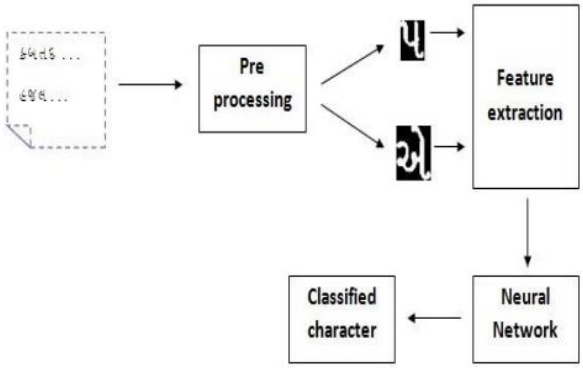

Fig -6: Classification process [5]

2. COMPONENT OF OCR SYSTEM

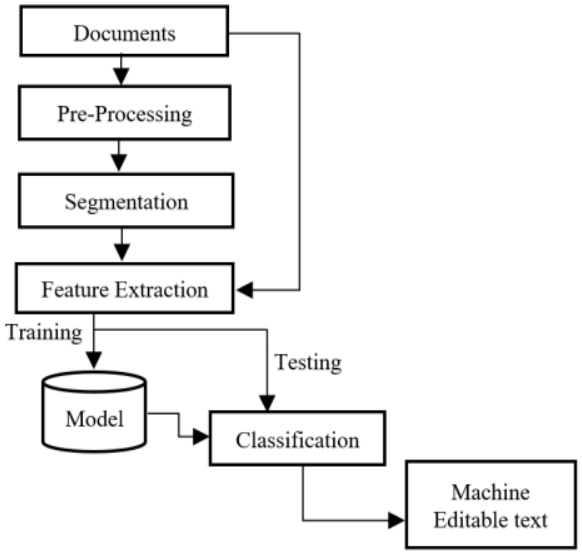

Fig -7: Components of the OCR system [1]

Properties of Gujarati characters

The Gujarati language is written in Gujarati script, which is written from left to right. The Gujarati language's character set consists of 34 constants, 10 numbers, and 12 vowels. Gujarati language also has conjuncts and join characters (Shown in fig.3)[12]

Fig -8: Gujarati character set: (a) Consonants, (b) Numbers, (c) Vowels, and (d) Conjuncts

Like other languages like Hindi and Sanskrit, Gujarati text is also divided into three zones: Upper, Middle, and Lower zone. The upper and lower zone contains the modifier symbols and the middle zone contains the basic character and conjuncts.

Fig -9: Zones of the words

Some of the work in the literature survey segments the character to the zone level where it segments both the basic characters as well as the modifier while some of the works segments the word only to the character level which considers modifiers apart of the character.

3. A LITERATURE SURVEY OF GUJARATI SCRIPTS

Vishal Naik and Apurva Desai [12] proposed a hybrid feature-based algorithm for online handwritten Gujarati character recognition. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Multi-layer Perceptron (MLP) classifiers were examined. They have gathered a training set of roughly 3000 samples for this purpose. They employed SVM with Radial basis function (RBF) kernel and SVM with the linear kernel to test their data set. Because the accuracy and execution times of both kernels differ. They got 91.63% accuracy and 0.063 seconds average execution time per stroke using SVM with RBF kernel, and 90.63 %accuracy and 0.056 seconds average execution time per stroke using SVM with Linear Kernel. Their suggested solution is eventually put to the test by 100 people.

Antani and Agnihotri [2] the data sets were created using 100 dpi scanned images of printed Gujarati text, as well as data from 15 font families found on the internet. They produced ten samples for each of the five fonts. The photos were scaled up and then down to a predetermined size to ensure that all of the samples were the same size, which was 30x20. It lacks skew correction and noise reduction capabilities. The author computed both invariant and raw moments for feature extraction. Image pixel values were also employed as features, resulting in binary feature space with 30x20=600 dimensions. For classification, the author used two classifiers: the K-NN classifier and the minimal Hamming distance classifier. For 600-dimensional binary features space, the best identification rate was 67 % for 1-NN. The recognition rate of 1-NN in regular moment space was 48%, while the recognition rate of the minimal distance classifier was 39%. Only 41.33% were detected by the Euclidean minimum distance classifier.

Jignesh Dholakia ET. al [3]have described an algorithm for identifying different zones in Gujarati printed text. They proposed using horizontal and vertical profiles in the algorithm. These zones have been defined by the slope of lines. The upper left corner of the rectangle and the borders of connected components from line-level determine the slope of the line. They employed three separate document images to extract 20 lines, 19 of which were detected with the right zone border. The line where it failed was skewed a lot.

M. Goswami and S. Mitra [4] published a paper in 2017 on their work on printed Gujarati characters with high-level strokes properties. The proposed method was evaluated using a dataset of printed Gujarati characters that included 12000 samples from 42 different classes. They employed K-nearest neighbor classifiers with shape similarity. On the combined dataset, the overall accuracy claimed was 94.97%.

P. Solanki and M. Bhatt[5] explained the work on Gujarati printed texts Their suggested solution uses PCA to extract features and Hopfield Neural Networks as classifiers. On a short dataset, the accuracy was found to be 93%.

N. Vyas and M. M. Goswami[6]published a paper in 2015 on their work on handwritten numerals for Gujarati scripts. The researchers gathered data samples from 300 participants of various ages, educational backgrounds, and genders. They gathered 3000 digit samples in all, with 900 digits written using a thick marker pen and the remaining 2100 digits written with a standard marker pen. All different kinds of fonts, writing styles, and digit sizes are available in the data sample they obtained. The authors used K-Nearest Neighbor, SVM (Support Vector Machine),

and Back propagation Neural Network to test three distinct classifiers. The recognition rates for KNN, SVM, and Back propagation NN are 91%, 93%, and 92%, respectively.

4. CHALLENGES IN GUJARATI OCR

Because the Gujarati alphabet is more complex than that of other regional languages, performing Gujarati OCR and obtaining a more effective result is more difficult. Some of the problems frequently affect the accuracy and reduce the result.

- The following are the criteria for quality:
- Characters that are similar for example ગ,ઞ
- Face, size, and style of the font constituents
- Conjuncts and diacritics
- Skewed Characters

Characters that are a mix of complicated half and full characters for example: Characters that are Brocken

ખ્ખ	ગ્ક	ઘ્ક	ચ્ક	ઞ્ક	ણ્ક	ત્ક	ધ્ક	ન્ક	પ્ક
khkha	gka	ghka	cka	ñka	ṇka	tka	dhka	nka	pka
બ્ક	ભ્ક	મ્ક	ય્ક	શ્મ	શ્લ	ષ્ટ	શ્ચ	ઙ્ક	ક્ક
bka	bhka	mka	yka	śma	śla	ṣṭa	śca	ṅka	kra
ખ્ર	ટ્ર	ર્ક	શ્ર	ત્ર	દ્ર	હ્ર	હ્ય	હ્મ	દ્વ
khra	ṭra	rka	śra	tra	dra	hra	hya	hma	dva
દ્ધ	દ્મ	દ્ય	ટ્ટ	ડ્ડ	ઠ્ઠ	ઢ્ઢ	ત્ત	દ્દ	
ddha	dma	dya	ṭṭa	ḍḍa	ṭṭha	ḍhḍha	tta	dda	

Zigzag line/word or letters, line spacing
These are just a few of the challenges that make Gujarati OCR tough.

Tesseract OCR engine

Tesseract is a popular OCR engine in comparison to others because it is open source, registered under the Apache license, and supports more than 100 languages [17]. It is compatible with all major computer operating systems. Apart from that, it generates output in a variety of forms, including Text, PDF, and other formats, using its own software. Or currently existing graphical user interfaces (GUIs) or APIs. It is maintained and renovated on a regular basis under the supervision of the owner. Google Inc.'s highly skilled crew [17].

Demonstrating how the Tesseract OCR engine works in practice. The first stage is to input a picture on which OCR is required, following which the image is sent through the "Adaptive Thresholding" step, where it is converted to Binary. Image [17]. Later, the Binary Image was processed into "Connected Component Analysis," which separated the text into words. Outlined characters are then put through a two-way pass to see if they can recognize words [17]. Pass-1 transforms recognized text/words into an adaptive classifier, which treats the data as training data. Now the text will be recognized for the second time, but this time it will utilize an adaptive classifier [17].

The rationale for the second recognition is because it is necessary to understand the context of the text from Pass-1 so that it may be identified easily in the second, third, and subsequent times[17].Character outlines were used to create the performance.

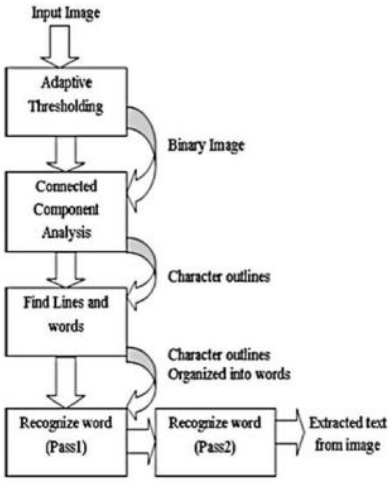

Fig -10: How Tesseract Works [17]

Dataset Details

For printed characters, there is a standard dataset available. The font styles Saral and shruti are included in this dataset. People can also improve the dataset by using various font styles and sizes. There are 46 characters in the dataset, with 34 vowels and 12 consonants. Each class has more than 100 photos. However, there are 385 classes in the dataset. All of the photos are 32*32 pixels in size [11].

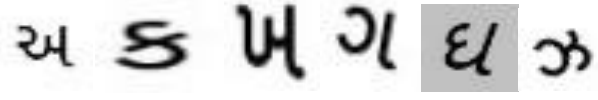

Fig -11: Dataset

5. CONCLUSION

An overview of several OCR approaches is offered in this publication. OCR is a multi-phase procedure that includes pre-processing, segmentation, feature extraction, classification, and post-processing. The OCR system can be utilized in a variety of real-time applications, including number plate recognition, smart libraries, and other real-

time applications. Despite a substantial amount of research in OCR, character recognition for regional languages like Arabic, Sindhi, and Urdu remains a challenge. Moreover, on Kaggle printed Gujarati character recognition dataset is available and it can be enhanced through different font styles and font sizes. The reviewed papers are not worked on the special characters and joint and half letters in Gujarati. So, that they need to work on that. Multilingual character recognition systems are another major field of research. Milind Kumar Aditya [17] developed an effective method of training the Tesseract engine with a high level of accuracy of 98 percent and 87 percent for both black and white and color images.

REFERENCES

[1] Kathiriya, K. B., & Goswami, M. M. (2019, March). Gujarati Text Recognition: A Review. 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). https://doi.org/10.1109/i-pact44901.2019.8960022.

[2] Antani, S., & Agnihotri, L. (1999). Gujarati character recognition. Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318). https://doi.org/10.1109/icdar.1999.791813

[3] Dholakia Jignesh, Negi Atul, S Rama Mohan, "Zone Identification in the Printed Gujarati Text", Proceedings of the Eighth International Conference on Document Analysis and Recognition(ICDAR '05) ISBN:0-7695-2420-6, pp.272-276.

[4] M. M. Goswami and S. K. Mitra, High-Level Shape Representation in Printed Gujarati Characters, vol. 1, SCITEPRESS, 2017, pp. 418-425.

[5] P. Solanki and M. Bhatt, "Printed Gujarati Script OCR using Hopfield Neural Network," International Journal of Computer Applications, vol. 69, pp. 33-37, 2013.

[6] N. Vyas and M. M. Goswami, "Classification of handwritten Gujarati numerals," in International Conference on Advances in Computing, Communications, and Informatics (ICACCI), Kochi, India, 2015.

[7] Patel Chhaya., Desai. A. Apurva, "Extraction of Characters and Modifiers from Handwritten Gujarati Words", International Journal of Computer Applications (0975 – 8887), Volume 73, Issue 3, pp 7-12.

[8] Nasir, T., Malik, M. K., & Shahzad, K. (2021). MMU-OCR-21: Towards End-to-End Urdu, Text Recognition Using Deep Learning. IEEE Access, 9, 124945–124962. https://doi.org/10.1109/access.2021.3110787.

[9] Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). IEEE Access, 8, 142642–142668. https://doi.org/10.1109/access.2020.3012542.

[10] Ubul, K., Tursun, G., Aysa, A., Impedovo, D., Pirlo, G., & Yibulayin, I. (2017). Script Identification of Multi-Script Documents: a Survey. IEEE Access, 1. https://doi.org/10.1109/access.2017.2689159

[11] https://www.kaggle.com/ananddd/gujarati-ocr-typed-gujarati-characters.

[12] Naik, V. A., & Desai, A. A. (2017, July). Online handwritten Gujarati character recognition using SVM, MLP, and K-NN. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2017.8203926.

[13] Goswami, M. M., Prajapati, H. B., & Dabhi, V. K. (2011, November). Classification of printed Gujarati characters using som based k-Nearest Neighbor Classifier. 2011 International Conference on Image Information Processing. https://doi.org/10.1109/iciip.2011.6108882

[14] Solanki, P., & Bhatt, M. (2013). Printed Gujarati Script OCR using Hopfield Neural Network. International Journal of Computer Applications, 69(13), 33–37. https://doi.org/10.5120/11905-7982

[15] Character Recognition of Gujarati and Devanagari Script: A Review. (2014). International Journal of Engineering Research & Technology (IJERT), 3(1).

[16] https://www.ijert.org/research/character-recognition-of-gujarati-and-devanagari-script-a-review-IJERTV3IS11115.pdf

[17] Zelic, F. (2022, February 10). [Tutorial] Tesseract OCR in Python with Pytesseract & OpenCV. AI & Machine Learning Blog. https://nanonets.com/blog/ocr-with-tesseract/

[18] Audichya, M. K. (2017). A Study to Recognize Printed Gujarati Characters Using Tesseract OCR. International Journal for Research in Applied Science and Engineering Technology, V(IX), 1505–1510.

[19] https://doi.org/10.22214/ijraset.2017.9219.