IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# A SURVEY ON PRINTED GUJARATI CHARACTER RECOGNITION FROM SCANNED IMAGE

[1]Anand J. Rathod, [2]Prof. Prashant B. Swadas, [3]Prof. Mosin I. Hasan

[1]M.tech Scholar, [2]Professor, [3]Professor

[1]Department of Computer Engineering,

[1]Birla Vishvakarma Mahavidyalaya, Vallabh Vidyanagar, India

***Abstract:*** Printed character recognition from a scanned document is a process of detecting characters of a specific language and recognizing them from a scanned document. It is useful in converting scanned document into some kind of editable document by extracting the recognized characters. Various OCR systems have been developed for recognition of various Indian languages such as Hindi, Kannada, Bangla, Odia, Urdu, Telugu, Gurumukhi and Gujarati. This paper presents some of the works that has been done in recognition of printed Gujarati characters from a scanned image/document using various feature extraction and classification techniques.

***Index Terms - Neural Network, Image Classification, Machine Learning, OCR, Gujarati.***

## I. INTRODUCTION

Optical character recognition is a technology that detects and recognizes the text that are present in any scanned document or any scanned image in handwritten or typewritten form. It is used to convert a scanned document into an editable/text document by recognizing the characters that can facilitate editing, searching, sorting or retrieval of documents. An OCR helps to improve the speed and decrease the human error in converting digital text to other digital file format and it also reduces the requirement of storage space as size of scanned documents are larger than text documents [2]. Major applications of an OCR are in data entry field, process automation, automatic license plate recognition and automatic cartography as well as an aid for visually disabled people by using OCR with speech synthesis [12]. An OCR faces the various challenges in recognition of the characters such as segmentation of words and characters, Skew-detection and correction, special characters, compound characters and similar looking characters. The efficiency of such systems very much depends upon the segmentation of single characters as the classification is implemented on a single characters rather than a complete word. Character recognition is mainly classified into online recognition and offline recognition. Online recognition is the type where the system recognizes the character in real-time when user is still writing [14]. Offline recognition takes an input of a scanned document or image after the user has completed the writing process.

Some work can be found in literature for various Indian languages like Hindi, Bangla, Odia, Urdu, Telugu, Gurumukhi as well as English however very little work can be found in recognition of printed Gujarati characters from a scanned document/image. Gujarati is a language spoken by around 60 million people in Gujarat (western India) as well as all around the world [6]. This survey is focused on (offline) recognizing printed Gujarati characters from a scanned document/image.

The rest of the paper divided as following: Section 2 describes the overview of an OCR. Section 3 describes the properties of the Gujarati script. Section 4 describes the Literature survey. Section 5 describes the analysis of literature survey and Section 6 describes the Conclusion.

## II. OVERVIEW OF AN OCR

An OCR can be implemented in four basic steps (shown in fig 1). The input to a basic OCR can be an image or a pdf document that contains only typed or handwritten text data without any images or figures or illustration. 1) Pre-processing: It involves the basic image processing operations on image to correct the input image before feature extraction and classification stage. Noise removal, binarization, skew correction are the common operations that are to be performed on the input. 2) Segmentation: It is the stage in which an input image is segmented into individual lines, lines are segmented into individual words and words are segmented into individual characters. It is the most important stage of an OCR since an OCR for almost any languages works on a single character rather than a whole word. 3) Feature Extraction: Features of the segmented characters are extracted using various feature extraction techniques that are to be used in the classification stage to classify different characters. 4) Classification: It is the final

stage of the system which classify the input character based on the extracted features using various classification models. After classification stage, once a character is recognized successfully, it is extracted to a text document using any mapping techniques, such as UNICODE mapping that is the most commonly used technique.

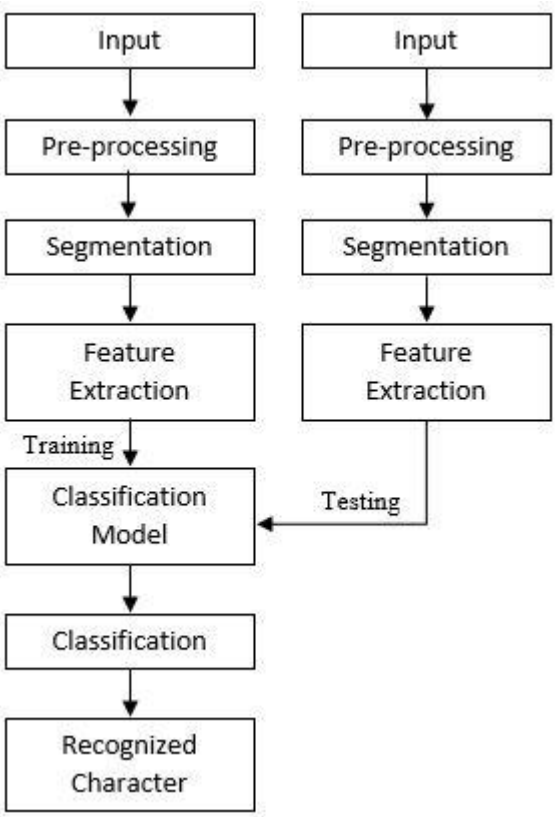


*Figure 1 Overview of an OCR*

## III. PROPERTIES OF GUJARATI SCRIPT

Gujarati script is a variant of the Devanagari script and is written from left to right. It contains 34 consonants, 12 vowels and 10 numbers [2] (shown in fig 2). It also contains various conjuncts and join characters (shown in fig 2). These conjuncts and join characters increases the difficulty of character segmentation which affects the classification. The main property of the script is that all the consonants can be combined with vowels to form a compound character (shown in fig 2) [13]. In a compound character, vowels work as modifiers or called 'Maatra' in Gujarati language with consonants. It can appear above, below, before or after the basic character.

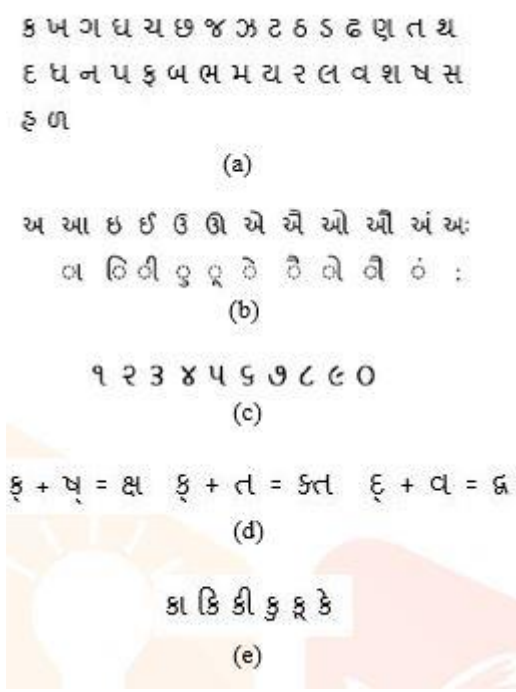


*Figure 2 (a) Consonants (b) Vowels and Modifiers (c) Numbers (d) Conjuncts (e) Compound Characters*

As other languages like Hindi and Sanskrit, Gujarati text is also divided into three zones [2]: Upper, Middle and Lower zone. The upper and lower zone contains the modifier symbols and the middle zone contains the basic character and conjuncts (shown in fig 3).


*Figure 3 Three Zones of script*

Some of the work in literature survey segments the character to the zone level where is 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 a part of the character.

## IV. LITERATURE SURVEY

M. Goswami, H. Prajapati, V. Dabhi [1] presented a method using Self Organizing Maps (SOM) with k-nearest neighbor for classification of printed Gujarati characters. SOM enables the pre-processing stage before k-nn in which data samples are reduced from n-dimension to 2-dimension while still preserving all the important information and then k-nearest neighbor classifier is used to classify the samples. The authors have achieved the accuracy of 82.36% for classification of test dataset of 533 characters collected from news-papers.

J. Dholakiya A. Negi, S. Mohan [2] proposed combinations of various feature extraction and classification techniques. Fringe map, Discrete Cosine Transform, Wavelet Transform and Zone Information is used for feature extraction while KNN and MLP such as Radial Basis Function network and General Regression Neural Network is used for classification. Wavelet transform with GRNN gives the highest accuracy of 97.59% in recognition of middle zone character and 96.32% for lower zone modifiers. The dataset contains 119 symbols of consonants, vowels, conjuncts and numbers.

M. Chaudhary, G. Shikkenawis, S. K. Mitra and M. Goswamiat [3] have proposed new mechanism of ESLPP (Extended Supervised Locality Preserving Projection) as feature extractor with BPNN to tackle the problem of classification of similar looking printed Gujarati characters. Six datasets each containing two or more similar looking printed Gujarati characters is used for experiment and the authors have achieved the minimum accuracy of 96.66% across all datasets.

P. Solanki and M. Bhatt [4] presented an OCR for recognition of basic printed Gujarati characters based on Principal Component Analysis (PCA) and Hopfield Network. The system is trained on 748 images of 34 consonants and 11 vowels and tested on 10 scanned documents with accuracy of 93.25%.

A. Sharma [5] presented a template matching technique to recognize the characters. A fringe map is created by replacing each non feature pixel by its distance from the feature pixel. A minimum distance between a template and input fringe map represents a recognized character. The system is able to recognize the basic character, upper and lower modifiers with an approx. accuracy of 72.3%.

M. Goswami and S. Mitra [6] presented a method based on the High Level Strokes (HLS) features of the characters. A HLS can be thought as a sequence of basic geometrical shapes like point, line, curve etc. and character can be thought as sequence of some HLS. Naïve Bayes Classifier and Hidden Markov Model are compared with HLS features. The system is tested on 16000 images of 48 characters and has achieved accuracy of 96.87% and 93.26% for HMM and NB respectively.

S. Chaudhary and R. Gulati [7] implemented a system that identifies and classifies English and Gujarati digits using KNN. After segmentation stage, a feature vector of size 35 is extracted from the input and is classified using a k-nn classifier with Euclidean distance. The system is implemented on 3500 samples of digits and has achieved an overall accuracy of 99.23%.

M. Goswami and S. Mitra [8] presented a method for character classification and word matching using High Level Strokes (HLS) features and KNN classifier using Shape Similarity as distance measure. Shape similarity is computed as a score by dividing the maximum matching subsequence of HLS between two characters by the minimum length out of the two sequences of the characters.The method achieves the accuracy of 94.97% on the samples of dataset from newspapers, books and scanned documents. The same method is experimented at word level also with the accuracy of 77.61%.

J. Prasad, U. Kulkarni and R. Prasad [9] presented a template matching technique for recognition of printed Gujarati characters. The input character is compared to pre-defined templates and cross correlation coefficient is calculated between input image and each templates and the highest coefficient represents the recognized character. The system is implemented for handwritten characters and can be implemented for using for printed character recognition. The overall accuracy achieved for recognition is 71.66%.

A. Choksi and S. Thakkar [10] proposed a system for recognition and to address the problem of classifying similar looking Gujarati typed characters using Wavelet Features and Fuzzy-Knn as classifier and also compared with other classifiers such as knn and GRNN.A fuzzy-knn assigns class membership of input to all samples rather than a particular class to an input. It gives degree of presence of a character to a class. The dataset contains 15 different font families with four sizes of 18 characters and 20 samples per character and the system has achieved accuracy of 67%, 97% and 100% with classifier knn, GRNN and fuzzy-knn respectively.

E. Hassan, S. Chaudhary and M. Gopal [11] presented a method based on Multiple Kernel Learnig (MKL) with Directed Acyclic Graph (DAG) for character classification with different features such as Feature Map, Shape Descriptor (SD), Histogram of Oriented Gradient (HoG).MKL SVM is used for classification.The system is implemented on dataset containing 5, 7 and 240 categories of lower, upper and middle zone primitives with 457, 1307 and 13083 number of images respectively and has achieved recognition accuracy between 96.37 and 97.14%

## V. Analysis of Literature Survey

This section describes the analysis of the literature survey based on the type of method used and based on the dataset.

### 5.1 Dataset:

Since there is not a universal dataset available for Gujarati language, almost all the works have developed a manual dataset of Gujarati printed characters in various fonts and for various number of classes/samples. The dataset samples are generated from various sources such as printed books, newspapers and laser printed documents [8]. The generated dataset is in various font size varying 12 to 16 and also in various types of font style such as Shruti, Saral. Though variations in font style is not very much but it still makes the dataset effective in classifying. If the letters are considered as a whole along with the modifiers, the dataset would contains around 374 classes which requires a very strong network model as well as results more in misclassification while segmenting the modifiers from the letters reduces the number of classes to 45 (34 basic letters + 11 modifiers). Thus it reduces the size of the dataset and number of classes as well as works more efficiently and accurately.

### 5.2 Methods and Approaches:

In a basic OCR, the input image/document should not contain any graphics or images or figures along with typed text since the system cannot differentiate between two of them. Skew detection and correction is done in the pre-processing stage to remove the skew of the input so it cannot decrease the recognition accuracy and Radon Transform is used for skew correction [4].

Segmentation is done to the zone level which segments the basic characters and the vowel modifiers that are used in upper and lower zones are segmented [2][8][1][3][6]. Slope information of imaginary lines between zones is used to separate the potential connected components of each zones [2]. Vertical histogram can be used for zone segmentation but since Gujarati language does not have ‘Shirorekha’(headline) over the words, there is no prominent peak differentiating upper and middle zone and more number of modifiers in upper and lower zones might be a problem as they can be counted as the middle zones because of their histogram peaks[2]. Line segmentation using vertical histogram gives efficient results [4]. Zone segmentation is not considered instead basic character with modifier is considered as a whole for recognition but it increases the number of classes [4].

In Feature extraction, Wavelet features are used for feature extractor where D-4 wavelet transform preserves all the important features after the first level of decomposition [2][10]. MLP with back propagation and RBF are iterative learning techniques which takes a large number of iterations to converge to a desired solution while GRNN is a single pass neural network that achieves a desired solution faster [2]. SLPP uses class labels of data to improve the discriminating along with inheriting the properties of LPP [3]. Characters represented by High Level Features gives reduced in size and high level shape description which are extracted using a junction-point-based algorithm [6][8]. Fuzzy-knn gives the most accurate results in recognizing similar looking characters with

wavelet features rather than structural features and also obtains the degree of membership of a character with a class [10]. Template matching gives acceptable results with any Euclidean or Manhattan distances of based on correlation coefficient [5][9].Hopfield Network which acts as an associative memory classifies the input character based on the stored pattern using hamming distance between the input and stored pattern[4]. Shape similarity is a score obtained between two characters, input and stored, and is used as a variation of distance measure in KNN [8]. MKL SVM is extended for multi class classification to decompose multiclass problem into multiple binary subclass problems. Binary MKL in DAG reduces the computational cost and chances of misclassification [11].

## VI. CONCLUSION

In this paper, a survey on recognition of printed Gujarati character from images is presented. A brief analysis of various techniques and methods of recognition system is done and an overview of a basic OCR is given. From the survey, it is found out that feature extraction techniques plays crucial role in recognition. If features are extremely localized, each sample might start representing a class which classifies character of same class into different classes and if feature extractor is weak, it may generate very similar features for characters in two different classes. Segmentation is also a challenging task since the script is different from other Indian scripts and modifiers are very difficult to segment. Character level recognition accuracy is better as compared to word level recognition. It is also found out that a standard dataset for such system is not available for Gujarati language either in printed or handwritten form. Further work can be carried out based on deep learning models and other methods for the recognition of printed Gujarati characters.

## REFERENCES

**[1]** M. M. Goswami, H. B. Prajapati and V. K. Dabhi, "Classification of printed Gujarati characters using som based k-Nearest Neighbor Classifier," 2011 International Conference on Image Information Processing, Shimla, 2011, pp. 1-5, doi: 10.1109/ICIIP.2011.6108882.

**[2]** Dholakia J., Negi A., Mohan S.R. (2009) “Progress in Gujarati Document Processing and Character Recognition.” In: Govindaraju V., Setlur S. (eds) Guide to OCR for Indic Scripts. Advances in Pattern Recognition. Springer,London. https://doi.org/10.1007/978-1-84800-330-9_4

**[3]** M. Chaudhary, G. Shikkenawis, S. K. Mitra and M. Goswami, "Similar looking Gujarati printed character recognition using Locality Preserving Projection and artificial neural networks," 2012 Third International Conference on Emerging Applications of Information Technology, Kolkata, 2012, pp. 153-156, doi: 10.1109/EAIT.2012.6407884.

**[4]** Prachi Solanki and Malay Bhatt. Article: Printed Gujarati Script OCR using Hopfield Neural Network. International Journal of Computer Applications 69(13):33-37, May 2013.

**[5]** Sharma, A., “Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching.” (2006).

**[6]** Goswami M.M., Mitra S.K. (2018) “Printed Gujarati Character Classification Using High-Level Strokes.” In: Chaudhuri B., Kankanhalli M., Raman B. (eds) Proceedings of 2nd International Conference on Computer Vision & Image Processing. Advances in Intelligent Systems and Computing, vol 704. Springer, Singapore. https://doi.org/10.1007/978-981-10-7898-9_16

**[7]** S. A. Chaudhari and R. M. Gulati, "An OCR for separation and identification of mixed English — Gujarati digits using kNN classifier," 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), Gujarat, 2013, pp. 190-193, doi: 10.1109/ISSP.2013.6526900.

**[8]** Goswami, M. and Mitra, S. “High Level Shape Representation in Printed Gujarati Character” In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017), pages 418-425 DOI: 10.5220/0006191104180425 ISBN: 978-989-758-222-6

**[9]** J. R. Prasad, U. V. Kulkarni and R. S. Prasad, "Template Matching Algorithm for Gujrati Character Recognition," 2009 Second International Conference on Emerging Trends in Engineering & Technology, Nagpur, 2009, pp. 263-268, doi: 10.1109/ICETET.2009.220.

**[10]** Amit H Choksi and Shital P Thakkar. Article: Recognition of Similar appearing Gujarati Characters using Fuzzy-KNN Algorithm. International Journal of Computer Applications 55(6):12-17, October 2012.

**[11]** Hassan, E., Chaudhury, S. &Gopal, M. “Feature combination for binary pattern classification.” IJDAR **17,** 375–392 (2014). https://doi.org/10.1007/s10032-014-0224-9

**[12]** A. Chaudhuri et al., Optical Character Recognition Systems for Different Languages with Soft Computing, Studies in Fuzziness and Soft computing 352, DOI 10.1007/978-3-319-50252-6_9

[13] C. Patel and A. Desai, "Segmentation of text lines into words for Gujarati handwritten text," 2010 International Conference on Signal and Image Processing, Chennai, 2010, pp. 130-134, doi: 10.1109/ICSIP.2010.5697455.

**[14]** S. R. Panda and J. Tripathy, "Odia Offline Typewritten Character Recognition using Template Matching with Unicode Mapping," 2015 International Symposium on Advanced Computing and Communication (ISACC), Silchar, 2015, pp. 109-115, doi: 10.1109/ISACC.2015.7377325.