# 3 OFFLINE GUJARATI HANDWRITTEN TEXT RECOGNITION (OGHTR)

The major objective of this research is to recognize text from the handwritten Gujarati text document and to model architectural solution to achieve it. Based on discussion given in section 1.3, 1.5 and 1.6 the architecture of offline Gujarati handwritten text recognition system framework is described with all its functions. Also, we can say that most of the research found in handwritten Gujarati is on isolated numerals recognition. The recognition rate of Gujarati alphabet and numerals is poor and in thirst of higher accuracy. This chapter propose an architectural model to design and develop offline HTR system for Gujarati script.

Based on literature review, characteristics of Gujarati script and its character set as discussed in section 1.3, we formulated an initial architectural model “Offline Gujarati Handwritten Text Recognition system (OGHTR)” consisting of preprocessing steps, segmentation, feature extraction and classification method to achieve objectives of developing HTR system stated in topic 1.6:

The objective of OGHTR module is to convert handwritten Gujarati text document image into editable UNICODE text. The purposes of this module is

- Preprocessing of document image and making it suitable for further process.
- Design segmentation algorithm of Gujarati text document image and extract symbols for purpose of recognition.
- Analysing and extracting features from the characters for recognition.
- Recognition of segmented unit as Gujarati alphabet includes consonant, vowels, numerals which are used in writing.
- Recognition of diacritic marks.

## 3.1 THE ARCHITECTURE OF OGHTR

The major objective of this research is to device appropriate segmentation process for handwritten Gujarati text document and to recognize it. The OGHTR follows analytical approach as it has benefit of unlimited vocabulary and it is well suited for Indic script [13,49,54,55,61,62,64]. The architecture of OGHTR module is shown as given below in Figure 8.

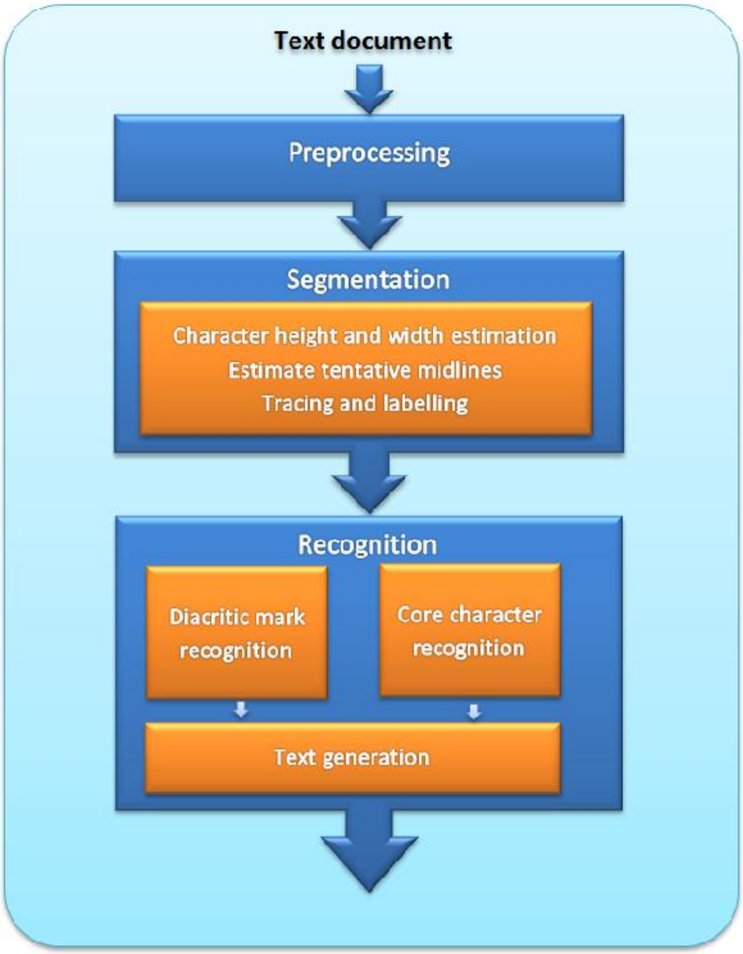

Text document
Preprocessing
Segmentation
Character height and width estimation
Estimate tentative midlines
Tracing and labelling
Recognition
Diacritic mark recognition
Core character recognition
Text generation

Figure 8 Offline Gujarati Handwritten Text Recognition System (OGHTR)

The OGHTR module is mainly divided into three phases, namely 1) preprocessing 2) segmentation and 3) recognition. The objective of preprocessing phase of OGHTR is to transform Gujarati handwritten text document in to a form which make it suitable for segmentation. The objective of segmentation phase is to isolate text into characters so that it can be feed into recognition phase for its recognition. The recognition phase of OGHTR extracts the character features which are used to recognize it to appropriate character class.

**Preprocessing:** The scanned image tends to have noises due to scanning device, pen, document skew and paper quality. The Gujarati text document image is inputted to the module and passed through the preprocessing steps. The preprocessing steps consist of processes like converting document to grayscale, noise removal, binarization, skew detection, corrections and cropping. The detail discussion on preprocessing phase is given in the next chapter. The preprocessing phase prepares document image to ready for segmentation.

**Segmentation**: The segmentation algorithm plays crucial role in recognition of handwritten text from document image in general [4,35,39,65,87,122] and Gujarati script in particular. The Gujarati script text lines have irregular shape of characters, diacritic marks, conjuncts and varying style of different writers increasing complexity of segmentation process [88].

As discussed earlier in section 1.4 the Gujarati text contains words and a word comprises of Akshara. It is further sub-divided into core character unit and diacritic mark attached to it. The main objective of OGHTR segmentation phase is to isolate these unit in such a way that it should sub-dived into appropriate categories i.e. core character and diacritic mark.

The detail discussion on segmentation phase of OGHTR is described in chapter 5.

**Recognition:** The result of segmentation process is isolated units, which are then classified to core characters and diacritics marks. We are using different approaches for recognition of core character and diacritic marks. The diacritic marks are few in numbers which can be easily recognize using binary tree classifier once its position is known. The output of segmentation phase for diacritic marks gives its position along with a symbol which allows us to use it.

For example, to identify right diacritic mark "*kāno (*ā, ɑ̈*)*" having structural properties are vertical line, two end points and position besides the character, similarly upper diacritic mark “*ek mātra (*e, ɛ*)”* having structural properties two end points and position above the character. Detail discussion on classification of diacritic mark is given in topic 6.1.

The core character can be a basic Gujarati character or conjunct. The core characters are 49 in numbers consist of 34 consonants, 10 numerals, 5 vowel symbol which are used in writing. As the core character classes are more in number we decided to use neural network for its classification. The recognition of conjunct is not in scope of this research.

To classify core character researcher proposed to build isolated Gujarati handwritten character recognition (IGHCR) system using neural network. The neural network requires training using existing character image data. The detail discussion on dataset is given in section 3.2.2 later in this chapter. The detail discussion on architecture of IGHCR system for core character recognition is given in section 6.2.

The text generation part in recognition module integrates recognized core character and diacritic mark to generate text using some linguistic rules. For example, if a character recognized as “ક્ષ kṣa kʃə” then requires to transform to three Unicode symbol as “□ ◌□□” which then form to character “ક્ષ kṣa kʃə”. This is due to ““ક્ષ kṣa kʃə”” is conjunct form of character “□” as half joined with second character or “□”.

## 3.2 DATASET

Standard datasets for English language are available for comparisons of results [6,63,75,108,123,124]. For example, CEDAR dataset consisting of 50000 segmented numerals from zip codes, 5000 city name, and 9000 state names, CENPARMI dataset contains 17000 numerals from zip codes and NIST contains more than 1000000 characters from forms, 91500 sentences with a dictionary.

The standard datasets for most Indian languages are not available [6,20,25,32,33,102]. The work on Gujarati character recognition is still at preliminary level with only Gujarati numerals and alphabets. As discussed in chapter 2, standard dataset for Gujarati text is not available, therefore we decided to create own dataset for Gujarati text document.

The OGHTR should recognize handwritten Gujarati text from Gujarati handwritten text document image containing basic Gujarati characters (34+2 consonants), numerals (10 numerals), and diacritic marks (11 matras except “◌▯ ḥ, ◌□ Â, and ◌▯” Ô). Also, we will discuss approach for join characters i.e. conjuncts, punctuation marks etc. which is right now not considered in the scope of research.

As per the discussion of the architecture model of OGHTR, we required two kinds of dataset,

i) Gujarati handwritten text dataset and,

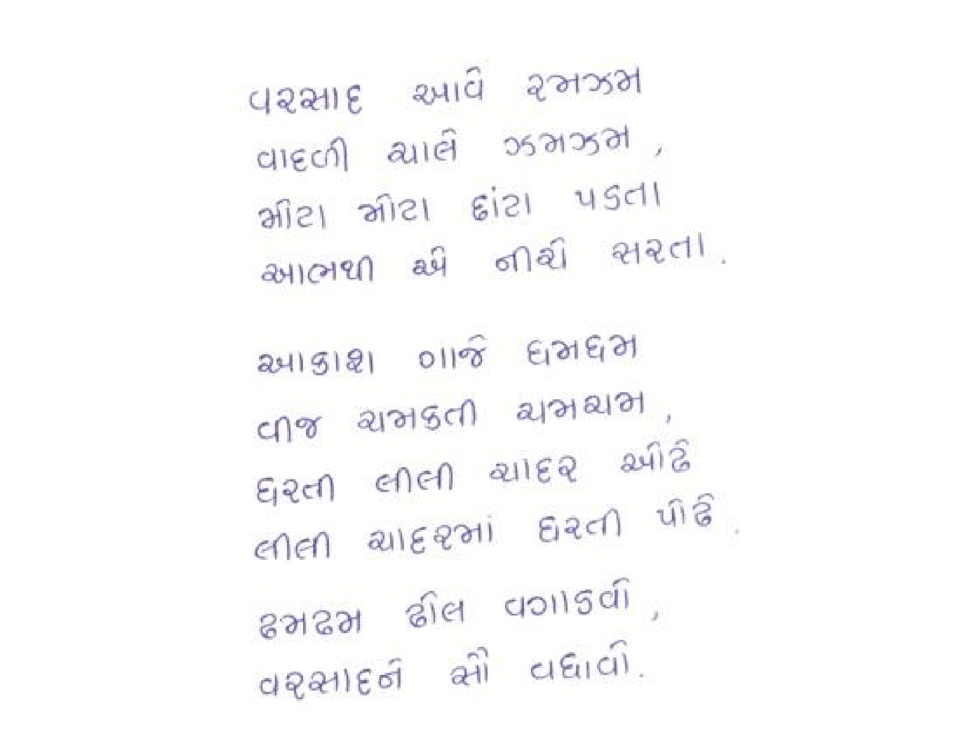
ii) Isolated Gujarati character dataset.

### 3.2.1 Gujarati handwritten text dataset

To fulfil the objective of segmenting document into core characters and diacritic mark, researcher collected handwritten text document written in Gujarati script. Since there is no standard benchmark text dataset available for Gujarati handwritten text, we need it for our experiment purpose.

Researcher collected 107 handwritten pages containing free flow Gujarati text on A4 size document. These pages are collected from twenty different writers of different age group ranging between 17 to 50 years, both male and female.

The pages are scanned using a flat-bed scanner at the resolution of 300 dpi into true colour images. Below Figure 9 shows handwritten sample text document images from handwritten text dataset. These handwritten text documents contain lines ranging from 5 to 21. The pages were written in single column. While writing document instruction is not given to any writer so that it contains naturally written text lines. Text document contains Gujarati alphabet, numerals as well as punctuation marks in its text.



વરસાદ આવે રમઝમ
વાદળી ચાલે ઝમઝમ ,
મોટા મોટા છાંટા પડતા
આભથી એ નીચે સરતા .

આકાશ ગાજે ધમધમ
વીજ ચમકતી ચમચમ ,
ધરતી લીલી ચાદર ઓઢે
લીલી ચાદરમાં ધરતી પોઢે .

ઢમઢમ ઢોલ વગાડવો ,
વરસાદને સૌ વધાવો .

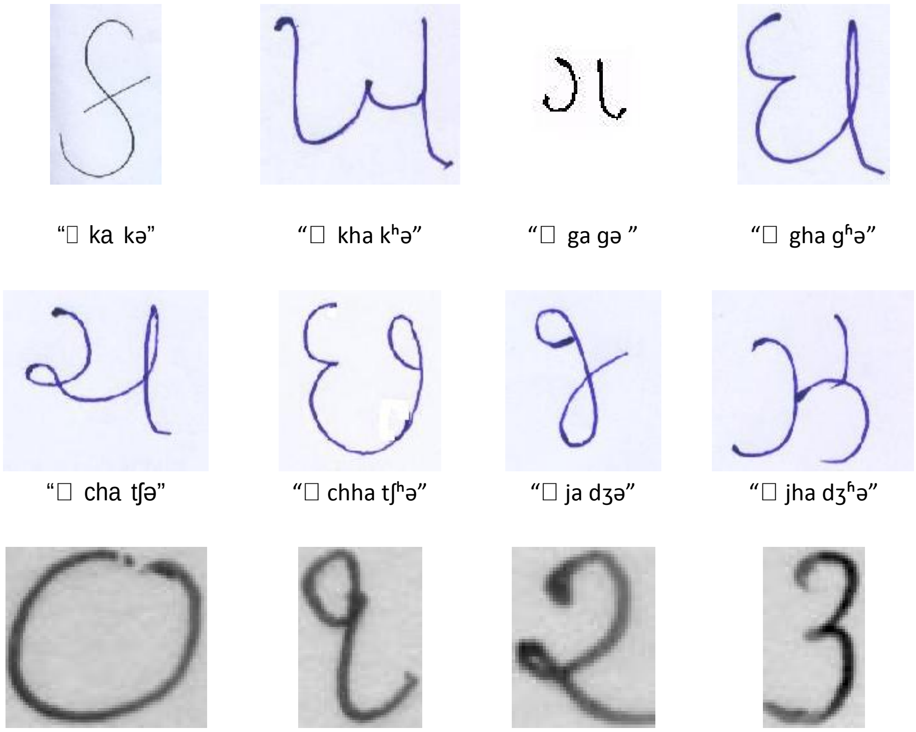
Figure 9 Handwritten Gujarati text document image

### 3.2.2 Isolated Gujarati character dataset

As the OGHTR model follows analytical approach, which is segmenting individual unit and recognizing it to generate text. To recognize the segmented unit, we need trained model from IGHCR module. The trained model is capable of recognizing Gujarati consonants, vowels (which are used in writing “□ a ə”, “□ i i”, “□ ī i”, “□ u u”, “□ u ũ”), and numerals. To implement IGHCR module and to obtain trained model from it, we required dataset of handwritten characters.

For training and testing, we used isolated Gujarati character data set obtained from T.D.I.L-Technology Development for Indian Languages [125]. TDIL Programme initiated by the Department of Electronics and Information Technology (DeitY), Ministry of communications & Information Technology, Government of India has the objective to develop information-processing tools & technologies to facilitate human machine interaction in Indian languages and to develop technologies to create & access multilingual knowledge resources.

Some of the character from the dataset is shown in Figure 10. The T.D.I.L. dataset consisting of 12860 samples of isolated Gujarati characters. Out of these, for each character we took 250 characters for our study.


“□ ka kə” “□ kha kʰə” “□ ga gə ” “□ gha gʱə”

“□ cha tʃə” “□ chha tʃʰə” “□ ja dʒə” “□ jha dʒʱə”

“□ shunya” “□ ekaṛo or ek” “□ bagaṛo or bay” “□ tragaṛo or tran”

Figure 10 Isolated Handwritten Gujarati Characters

The samples from TDIL contain 30 consonants and 12 vowels. Four characters which are not in the dataset are “□ ṇa ɳə” “□ la lə”, “□ śha ʃə”, “□ ha ɦə”. Furthermore, our objective is to identify numerals also we need samples for Gujarati numerals 0 to 9 in the dataset.

As these characters are used in writing, we decided to collect samples for it. We have collected 10 samples for each of these characters from 25 different writers. So, for one character 250 samples are added to a dataset. One such sample datasheet is shown in Figure 11.

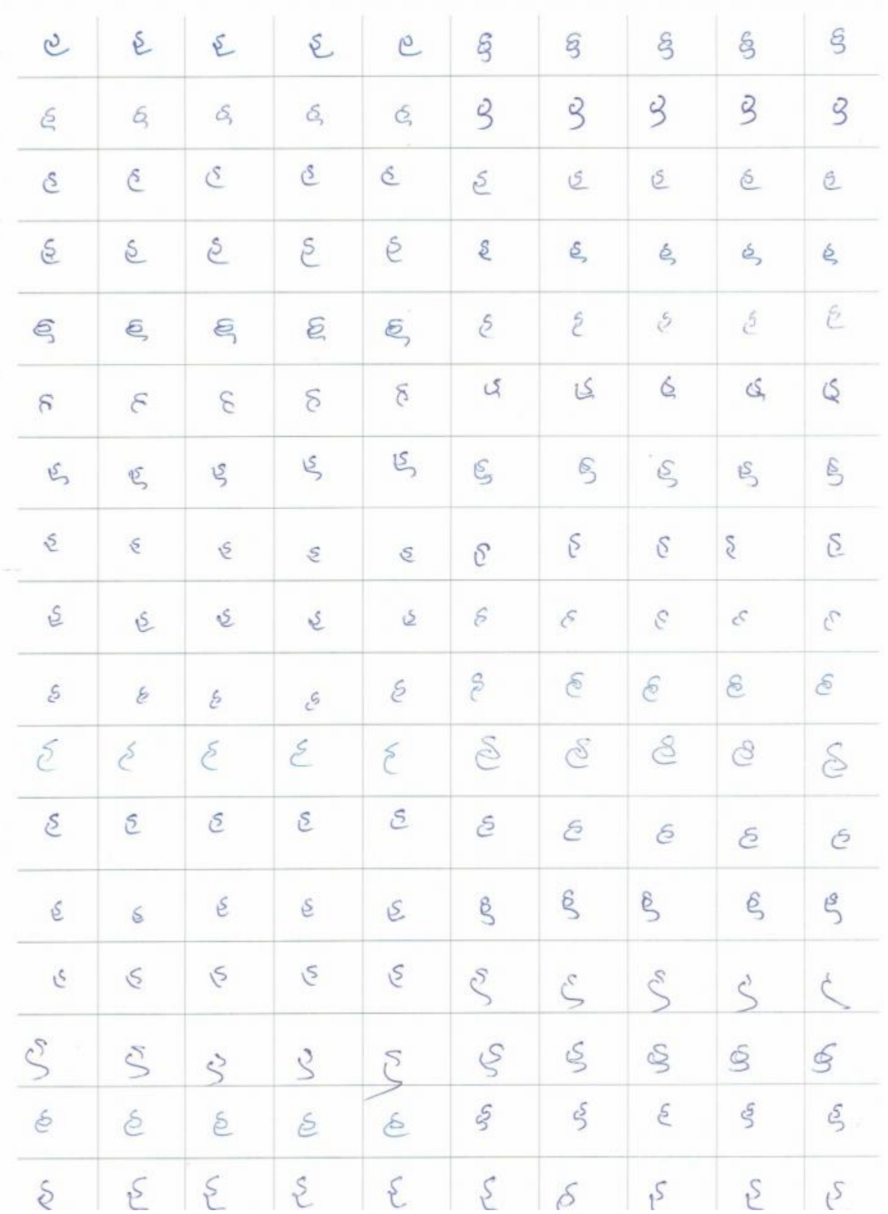


Figure 11 Sample datasheet

The dataset which is used to implement IGHCR module contains 49 handwritten isolated Gujarati character. The overall size of dataset is 12250 sample images as shown in Table 1.

Table 1 Sample size of Gujarati isolated characters

| Character | No of Samples | Character | No of Samples | Character | No of Samples |
|---|---|---|---|---|---|
| અ | 250 | □ | 250 | □ | 250 |
| ઇ | 250 | □ | 250 | □ | 250 |
| ઈ | 250 | □ | 250 | □ | 250 |
| ઉ | 250 | □ | 250 | [illegible] | 250 |
| ઊ | 250 | □ | 250 | [illegible] | 250 |
| ક | 250 | □ | 250 | □ | 250 |
| ખ | 250 | □ | 250 | □ | 250 |
| ગ | 250 | □ | 250 | □ | 250 |
| ઘ | 250 | □ | 250 | □ | 250 |
| ચ | 250 | □ | 250 | □ | 250 |
| છ | 250 | □ | 250 | □ | 250 |
| જ | 250 | □ | 250 | □ | 250 |
| ઝ | 250 | □ | 250 | □ | 250 |
| ટ | 250 | □ | 250 | □ | 250 |
| ઠ | 250 | □ | 250 | □ | 250 |
| ડ | 250 | □ | 250 | | |
| ઢ | 250 | □ | 250 | | |

### 3.3 DISCUSSION

In this chapter we have discussed proposed architecture of handwritten text recognition system for Gujarati script having preprocessing, segmentation and recognition phases. Our proposed architecture follows analytical approach to build HTR system, the output of segmentation process is individual segmented unit. These segmented units then labelled as Gujarati character or diacritic mark during segmentation process. We proposed different approaches for diacritic mark recognition and character recognition.

Two types of dataset are required to experiment the model. The handwritten text dataset is created for the experiment purpose and isolated character dataset is obtained from T.D.I.L. to train the recognition module for core character. In the next chapter we will discuss first step of our solution that is preprocessing.