Design and Development of Gujarati OCR

Technical Report · June 2012

CITATION
1

READS
996

2 authors, including:

A.G. Ramakrishnan
Indian Institute of Science
392 PUBLICATIONS **3,033** CITATIONS

SEE PROFILE

Design and Development of Gujarati OCR

A Project Report

Submitted in partial fulfillment of the

Requirement for the Degree of

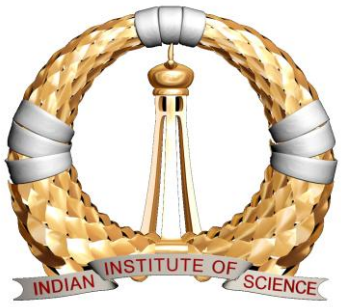
Master of Engineering

in

System Science and Automation

BY

Patel Pratik Shantilal



Electrical Engineering

INDIAN INSTITUTE OF SCIENCE

BANGALORE 560 012, INDIA

JUNE 2013

Acknowledgment

I acknowledge the support and encouragement provided by my loving parents, brother, my relatives and my friends.

I am deeply grateful to my guide, Dr. A.G.Ramakrishnan for his encouragement and guidance, and Helping me personally.

I would like to thank my lab mates Krunal, Suresh, and Himanshu for their help and support. Special thanks to Madhavraj, sandeep and shiva for helping me personally.

Finally, My heartfelt thanks to my IISc friends for supporting me and making my IISc life colorful and Unforgettable.

Abstract

This report presents the design and implementation of an OCR (Optical Character Recognition) system for printed Gujarati text.

The utilities provided by OCR systems are Postal address reading, Forms processing, back files of technical journals, Books, Reading aids for the blind (based on cell-phone camera), Personal OCR (including pocket computers with translator). Manual operation for the some of above applications is time consuming. Most of the research is done in OCR on Roman script. However, in the case of Indian scripts, the problem of automatic recognition is still an unsolved topic of considerable interest. In this report, the design of an OCR system for printed Gujarati text is presented. The machine recognition of Gujarati character is difficult due to similarity in the shapes of different characters, script complexity, non-uniqueness in the representation and very high number of symbols.

Initially, at M.S University, Vadodara effort is done by Prof.ramamohan, for classification of Gujarati script. They used Nearest Neighborhood classifier for reorganization.

In this present work, some of the existing algorithms have been used. Modifications have been done, wherever necessary, for improving system performance. For this work, a data set containing more than 14,000 samples of different font styles and sizes has been collected from scanned images of various Gujarati written pages. Total of 885 classes with 16 average numbers of samples have been used.

Contents

List of Figures

List of tables

Chapter 1

Introduction

Optical character reorganization (OCR) deals with the machine reorganization of character presented in a document image .there have been significant effort to develop a good machine reorganization system to recognize text from the scanned document image. The goal of the work is to convert paper manuscripts to digital documents.so text can be read and processed by the machines.

1.1 OCR work in Indian Script

For Indian script considerable efforts have been made and are still continuing towards the development of systems. A goal of this work to come out with a system which can deal with the all Indian scripts. The design of OCR system capable of converting multi-lingual manuscript to machine-readable digital document is one of the key steps in working towards the goal of machine translation.

Devanagari script the Most common script for the North Indian language, earliest attempts towards the development of OCR for this script is done by shethi and chatterjee [9],to recognize hand printed devenagri numerals. Srinivasan and ramakrishnan[10],analyzing the applicability of independent component analysis(ICA)[11,12] to OCR Devnagri, shown that the independent components of character are their strikes.they have subjected devnagari hand-written characters to ICA and repoted a high acuuracy. Dutta and chaudhary [13] have developed an algorithm to recognize the Bangla printed as well as hand written alphanumeric characters using a curvature features. This is the best efforts done for the Bangla Script. For machine reorganization of printed kannada characters, geometric moments have been extracted by Atul *et al*[14] and a neural network classifier has been used for classification. Kunte and Samuel [15] have developed a system for on-line reorganization of handwritten kannada characters. This system is able to

handle variation in size, speed and substantial deviations of the shape and reported the high success rate for kannada script.

For Tamil script shiromoney et al proposed an encoded character string dictionary to recognize of tamil characters. This principle is used by chandrashekaran et al. to for both printed and handwritten Tamil character reorganization. There are many other Indian scripts on which work has started in the recent past.

1.2 Properties of Gujarati script

Gujarati is the official language of West Indian state of Gujarat and is spoken by 60.3 million people. Like other Indian languages Gujarati is Different from the Roman Script.Unlike many north Indian languages Gujarati have no sirorekha on the top of the word and hence all the character are isolated. This creates problem in word segmentation and to segment the upper modifier.Gujarati script is more complicated for machine reorganization because of the compound character called *Jodakshar.* In Gujarati Script there is no upper case and lower case concept.

In Gujarati script there are 14 vowels (Table 1.1), 36 consonant (Table 1.2) and 10 Digits (Table 1.3). Consonant takes modified shapes when vowels are added to it. Vowels modifier can appear to the right, left, top, bottom and middle of the character. Such consonant vowels combination called modified characters. In addition two or three characters can make new shapes (*jodakshar*).

અ a	આ aa	ઇ e	ઈ Ee	ઉ u	ઊ oo	ઋ ru
ઌ Lu	એ ae	એ aee	ઐ Ai	ઓ o	ઓ ao	ઔ au

Table 1.1 Gujarati vowels

ક Ka	ખ kha	ગ Ga	ઘ Gha	ઙ na	ચ ca
છ Cha	જ ja	ઝ jha	ઞ Na	ટ ta	ઠ tha
ડ Da	ઢ dha	ણ na	ત Ta	થ tha	દ da
ધ Dha	ન na	પ pa	ફ Pha	બ ba	ભ bha
મ Ma	ય ya	ર ra	લ La	વ va	શ Sa
ષ Sa	સ sa	હ ha	ળ La	ક્ષ ksa	જ્ઞ gna

Table 1.2 Gujarati Consonant

૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine

Table 1.3 Gujarati digits

Vowel	Vowel Modifier	When attached to consonant (ક)
આ	ા	કા
ઇ	િ	કિ
ઈ	ી	કી
ઉ	ુ	કુ
ઊ	ૂ	કૂ
ઋ	ૃ	કૃ
ૠ	ૄ	કૄ
એ	ે	કે
ઐ	ૈ	કૈ
ઓ	ો	કો
ઔ	ૌ	કૌ

Table 1.4 Vowels and the shape of the consonant ક modified by vowel modifiers

Chapter 2

OCR System

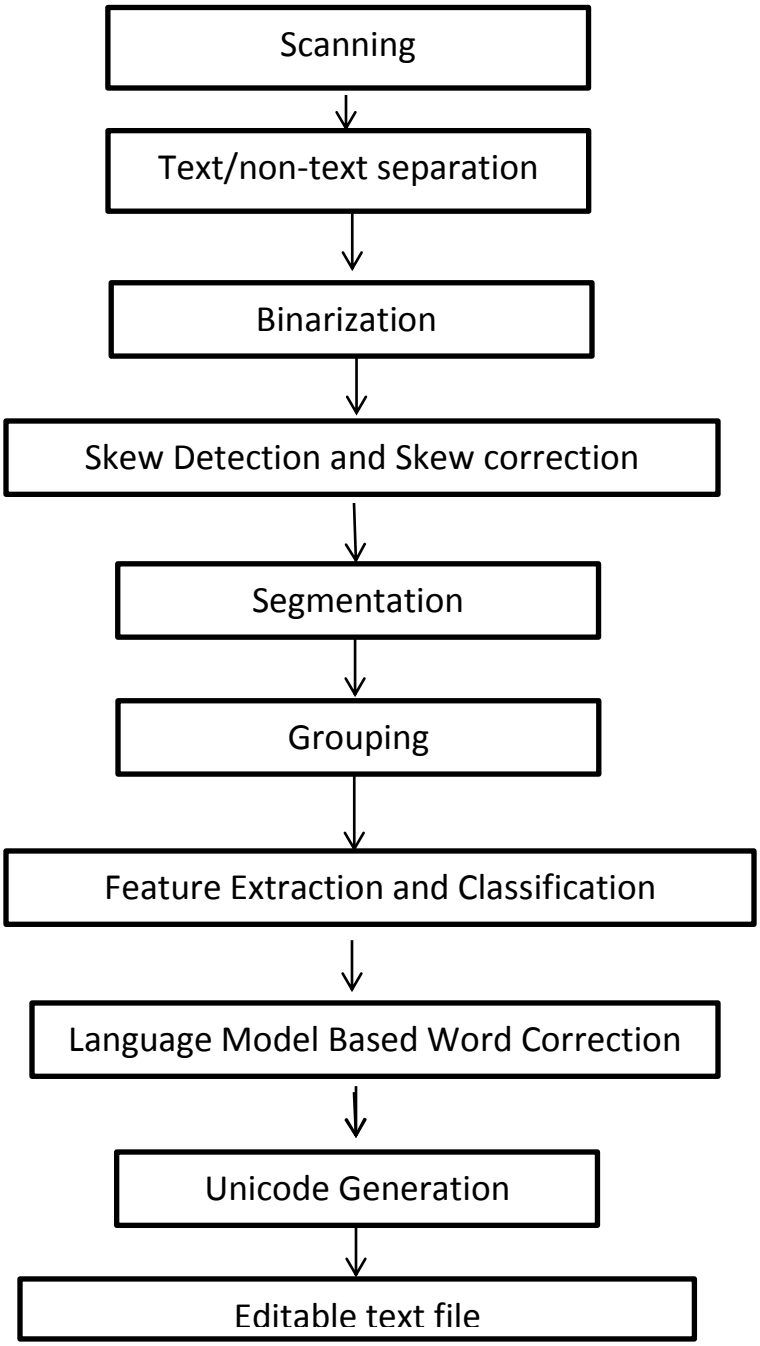


Fig 2.1 OCR System

Scanning: The hard copy of a text document is converted into a digital image file using a scanner interface.

Text/non-text separation: Separating a non-text from a text**.** This block will remove graphics, images, tables and lines from the text image.

Binarization: converting a gray level image into a binary image is called a binarization.

Segmentation: extracting the object of interest from the image is called segmentation.

Grouping: To reduce the confusing classes and search space, the classes are divided into a five different group based on height and width.

Feature Extraction: feature extraction is to extract information from image, which is most relevant for classification purposes.

Classification: using the extracted features, a decision is made by the classifier as to which character, the given sub block of image represents.

Post processing: Providing a priori information to the system about the script and the rules of the language to improve the performance of the OCR.

2.1 Digitization

Digitization is process of converting hard copy of text document into a gray scale digital image. Digitization is done using a flatbed scanner. Hp scanner is used with a scanning resolution of 300 dots per inch (dpi).Image are stored in tagged image file format (TIFF).

2.2 Preprocessing

The preprocessing includes all the processes, which remove noise, binarize the gray scale document image, detect the skew and correct it. The preprocessing also include process such as separation of text region from the rest of the contents like pictures and graphics in the page .segmentation of text region into line, word and characters belong to segmentation.

Preprocessing involves binarization, skew detection and skew correction.

2.2.1 Binarization

Image is obtained after digitization is a gray scale image.so we convert grayscale image in to a binary image, and the process is called binarization. Because of binarization it will reduce the space for recognition and also can cluster easily for classification. If binarization is not done carefully, binarization may lead to breaks in characters or merged the character. Traditionally two types of approach are used, global threshold method or local threshold method. Both approaches have advantages as well as disadvantages global threshold method will use global information and neglects the local information. Local threshold method will use local information and neglects the global information of the image. . Otsu's [1] binarization method is an example of global binarization approach. Niblack [2] and Sauvola [3] methods are local binarization methods. Therefore, binarization techniques proposed by Gorman [4] is used. In this method threshold is fixed globally, using local information. Thus, this method has the advantages of both global and local approaches. Using Gorman algorithm results are obtained. Fig 2.2 and fig 2.3 shows the result of binarization result.

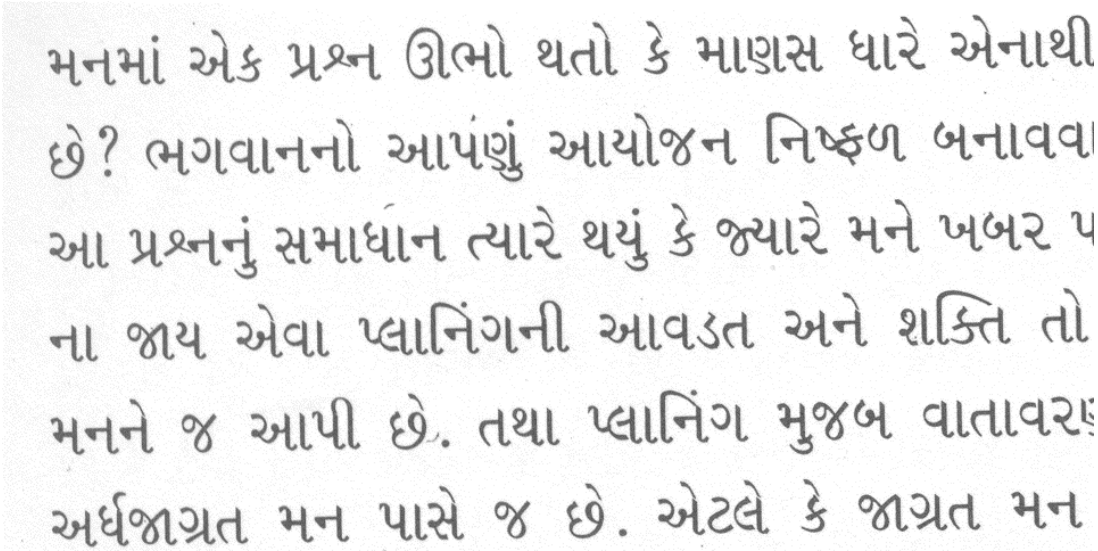

મનમાં એક પ્રશ્ન ઊભો થતો કે માણસ ધારે એનાથી
છે? ભગવાનનો આપંણું આયોજન નિષ્ફળ બનાવવા
આ પ્રશ્નનું સમાધાન ત્યારે થયું કે જ્યારે મને ખબર પ
ના જાય એવા પ્લાનિંગની આવડત અને શક્તિ તો
મનને જ આપી છે. તથા પ્લાનિંગ મુજબ વાતાવરણ
અર્ધજાગ્રત મન પાસે જ છે. એટલે કે જાગ્રત મન

Fig 2.2 Original Gujarati text image

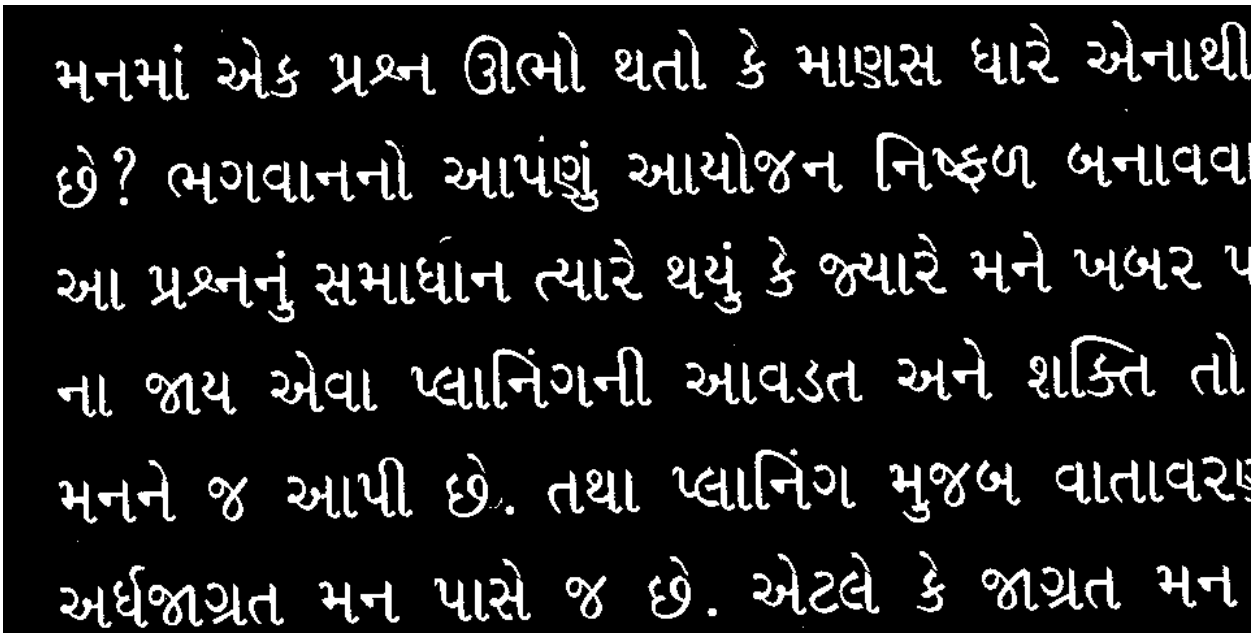

મનમાં એક પ્રશ્ન ઊભો થતો કે માણસ ધારે એનાથી
છે? ભગવાનનો આપંણું આયોજન નિષ્ફળ બનાવવા
આ પ્રશ્નનું સમાધાન ત્યારે થયું કે જ્યારે મને ખબર પ
ના જાય એવા પ્લાનિંગની આવડત અને શક્તિ તો
મનને જ આપી છે. તથા પ્લાનિંગ મુજબ વાતાવરણ
અર્ધજાગ્રત મન પાસે જ છે. એટલે કે જાગ્રત મન

Fig 2.3 Binarized image

2.2.2 Skew detection and correction

Misalignment of manuscript with respect to the axes of the scanner causes some tilt in the digital image. This is known as skew, and the angle of the tilt is called ***skew angle.*** It is essential to detect and correct the skew of the image. Skew of the image leads to erroneous line segmentation, word segmentation which leads to erroneous character segmentation and the recognition accuracy falls drastically. This process involves two steps (I) precise detection of the skew angle called skew detection, and (II) removing the misalignment called skew detection.

2.2.2.1. Skew Detection

Detection of the skew angle is one of the well-studied sub processes of document analysis. U. Pal *et al*.[5] proposes an algorithm which works well for devnagri script. This is due to the straight line present on the top of the words called ***shirorekha.***

The algorithm proposes by kaushik Mahata *at el* [6,7] works well for the Gujarati script. A course estimation of the skew with an accuracy of ±0.25° is achieved. The steps involved in coarse estimation of the skew angle are:

- Run-length smoothing: this process attempts to fill the space within characters and between characters in a word. This process is done to connect all characters of a particular word.
- Interim-line image formation: interim line is the line mediating the bag round gap between the two text lines. For finding the interim-line first perform Run length smoothing then find the mid-points of all pairs of consecutive words along the vertical direction and take those points as constituents of the interim line. Form an image with this line, this image is called interim image.

Apply Hough transform [16,17] on the interim image to find a coarse estimation of the skew angle.

The next step is to estimates the skew angle to a finer precision. The algorithm is as follows:

- Segment the image into individual lines using the rough estimate of skew angle.
- Create separate image for each segment line.
- Superimpose this entire line image by aligning their center to obtain a *scatter* image.
- The principle axis of the scatter image is obtained by the principle component analysis. The angle of the principle axis and the horizontal axis of the scattered image is called precise skew angle θ.

2.2.2.2 Skew correction

After finding the precise skew angle θ the image needs to be rotated in order to remove misalignment. This process involves three steps:

1. Up sample the image by a factor of two. Pass it through a low pass filter for smoothing.
2. Rotate the image by a precise angle θ. Due to the inability of discrete representation to represent fractions we are getting unfilled points.to avoid unfilled point a bilinear interpolation scheme is used.
3. Down sample the rotated image back to its original dimensions.

   After removing the misalignment the OCR will give much better performance.

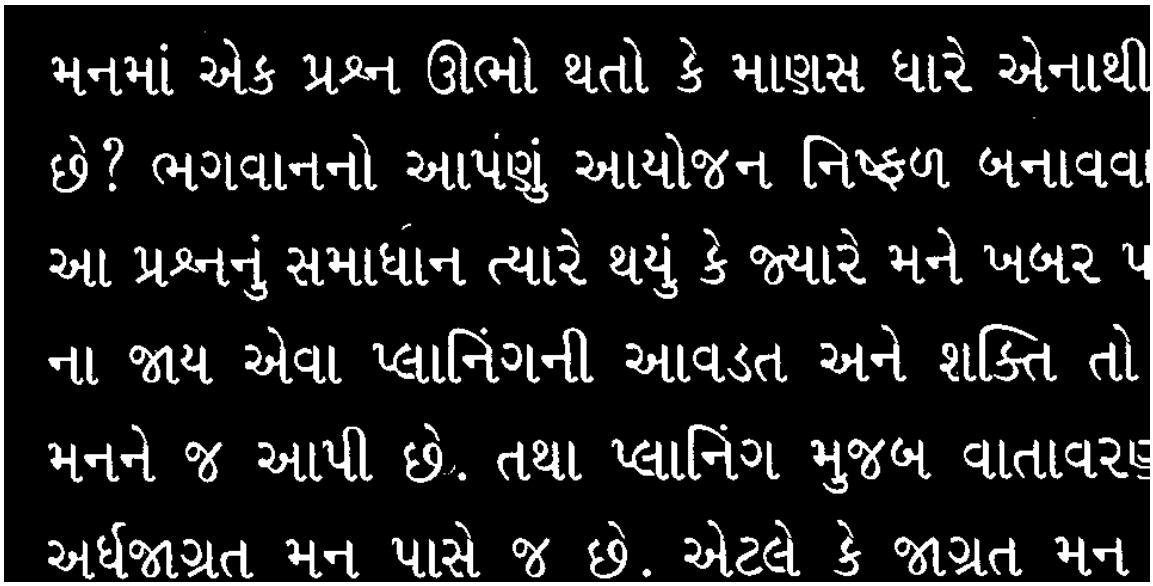


Fig 2.4 Skew corrected image

2.2.3 Segmentation

After correcting a skew of the binarized image we need to separate the separate character image for the classification. The technique of separating character is called segmentation. Segmentation is more important step because not properly segmented character images will not recognize properly and the OCR accuracy will fall drastically. Three types of segmentation are done on the image line segmentation, word segmentation and character segmentation.

2.2.3.1 Line segmentation

To segment the lines, the horizontal projection profile of the image is obtained. Horizontal projection profile is sum of the all ON pixels in the row. Therefore, this profile will have valleys of zeros height between the lines. This zero valley taken as the separators of the lines as shown in fig. .In the case of Gujarati script, sometimes the lower modifier of the upper line overlaps with the upper modifier of the bottom line. In this case non-zero valley is obtained between the two lines.

To segment such lines, the statistics of the height of the lines are found out from the projection profile. Then the threshold is fixed at 1.6 times the average line-height. Non-zero valley below the threshold indicates the text line and zero valleys above the threshold indicate the presence of a two lines. The midpoint of a non-zero valley of two text lines is the separator of the two lines.

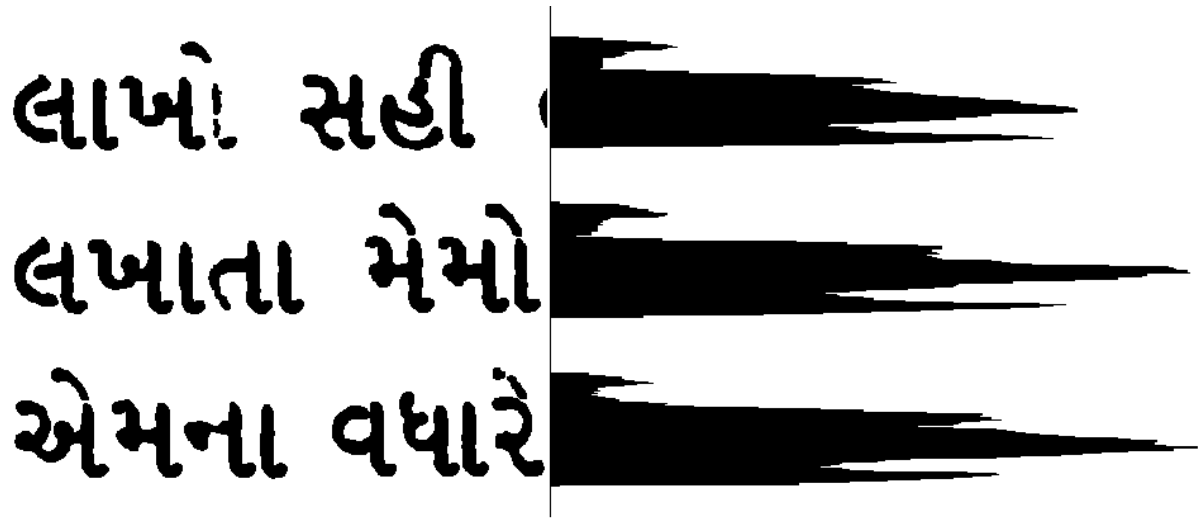


Fig 2.5 Line segmentation

2.2.3.2 Word segmentation

In Guajarati script there is no *sherorekha* like devnagri script. Therefore all character in a word is isolated. In Gujarati script the character spacing is non-uniform due to the presence of consonant conjuncts. Sometimes the spacing between the two characters is comparable to the word spacing. This could affect the word segmentation.

Therefore Morphological dilation is applied first and then vertical projection profile is obtained to segment the word. The objective of dilation is to connect all the characters in word. Each on pixel in image is dilated with the structuring element. Then the vertical projection profile of the dilated image is obtained. Vertical projection profile is the sum of all the ON pixels in column. The zero valued valley in the profile of the dilated image separate the words in the image. Word segmentation is shown in fig 2.6.

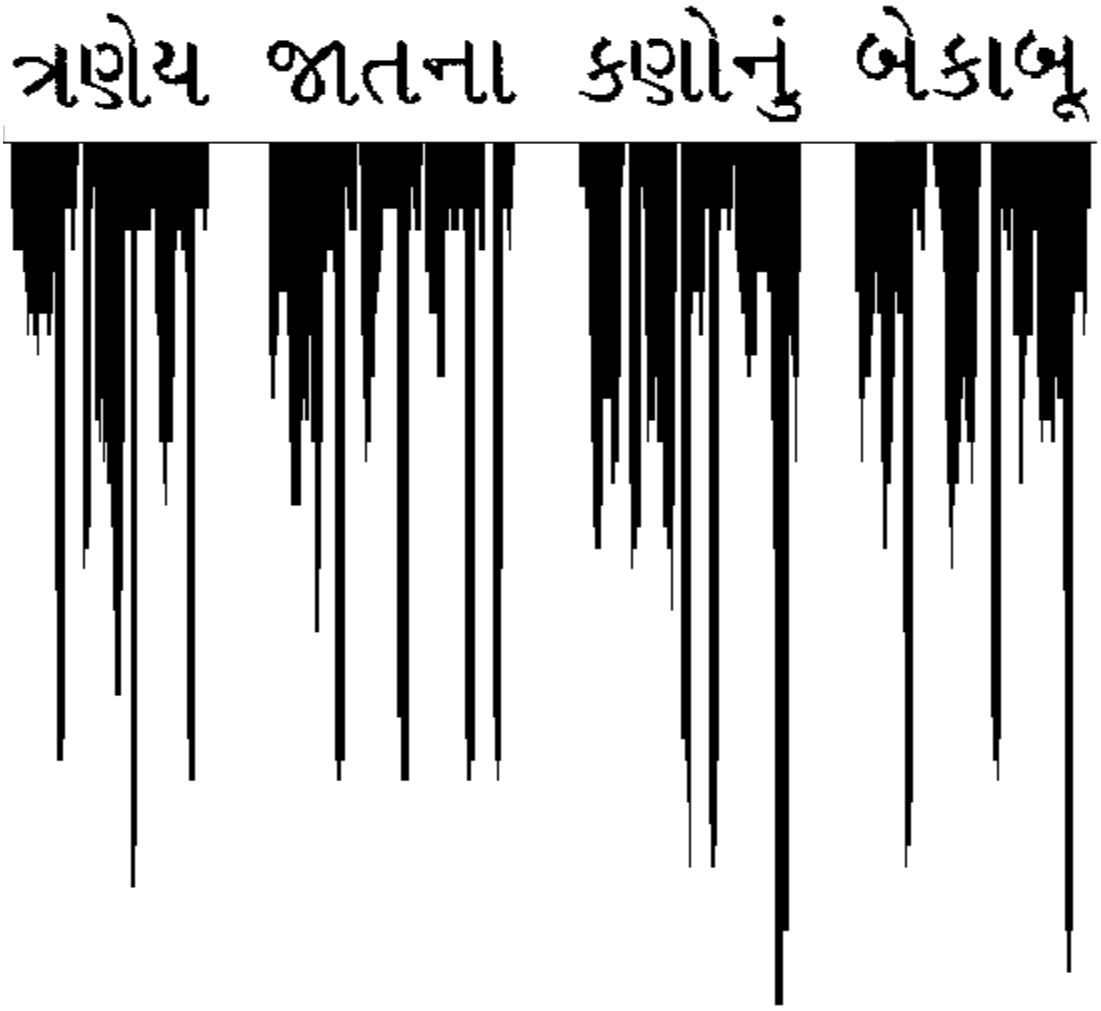


Fig 2.6 Word segmentation using VPP

2.2.3.3 Zone detection

Detecting the boundaries of the three vertical zones top zone, middle zone and the bottom zone helps in separating character and grouping the symbols. The horizontal projection profile of each segmented word is obtained. The line passing through the index corresponding to the maximum top half of the profile is the head line and the baseline corresponds to the maximum in the bottom half of the profile. Based on the profile each word can be partitioned in to three zones as shown in fig.2.7 the top zone contains the upper modifier, middle zone contain the basic character and compound character and the bottom zone contain lower modifiers.

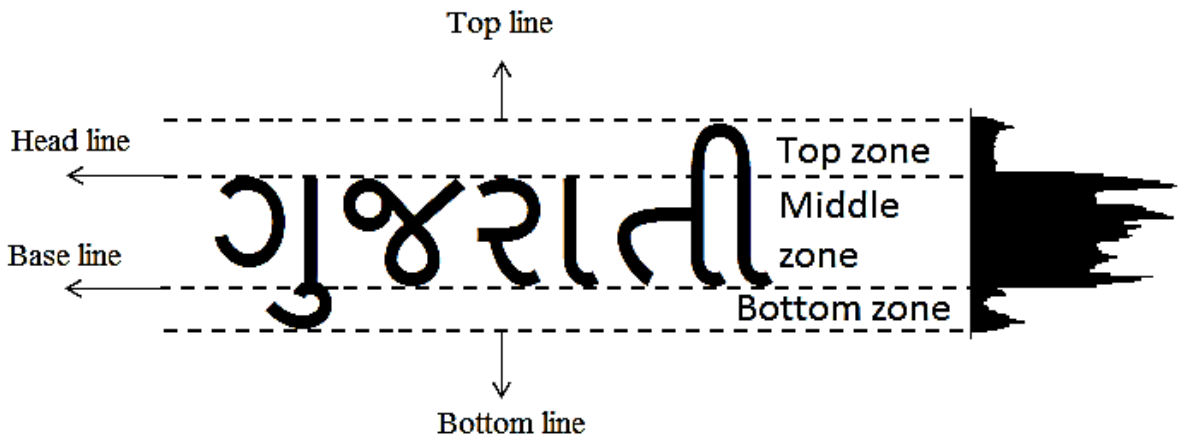


Fig 2.7 Zone Dectection

2.2.3.4 Symbol segmentation

After word segmentation the next step is to segment the symbol. Connected component algorithm is used to segment the symbol. The minimum box containing the connected component is called bounding box .each bounding box contain one symbol. After the symbols are segmented, every symbol is grouped into different five groups. Grouping is explaining in the next chapter.

2.3 Grouping

To reduce the search space for the classifier and to reduce the confusion pairs classes are grouped into five groups based on the height and width of the symbol .Different classifier is used for different group. Grouping is further explained in next chapter.

2.4 Normalization

After the individual symbols are segmented and grouped into respective groups, they are normalized before feature extraction. For every group different normalize size is taken and the below table shows the Normalization size for different group.

Group	**Normalized Size**
No Modifier	32x32
Upper Modifier	36x32
Lower Modifier	36x32
Both Modifier	40x32

Table 2.1 Normalization for Different Groups using bilinear interpolation

2.5 Feature Extraction

Feature is used to differentiate between patterns of different classes. The goal of feature Extraction is to extract the row data information which is most relevant for the classification.in other word feature extraction is a transformation from m-dimensional space to a smaller n dimensional feature space that retains most of the information needed for classification. Feature extraction should do in such way that patterns from different classes should be well separated in the feature space. Invariance to transform of the character such as scaling, translation and rotation. Features should robustness to noise like break or deformation.

There are advantages in working with features, rather than directly with the character bitmaps. Computational complexity of pattern classification is reduced by handling the data in lower dimensional feature space resulting in faster classification. Storing feature set takes less memory compared to storing the character bitmaps without feature extraction.

Some features are Structural features, DCT, DFT, KLT etc. In this work Template feature is used.

2.6 Classifications

The task of classification is to assign the test pattern x to unknown class label to one of the predefined set of classes using some similarity measurement between predefined classes and the text sample. The classifier evaluates the features provided by the feature extractor and takes the final decision about the label of the test pattern.

There are so many classifiers are used for pattern reorganization. Some of them are: Nearest neighbor classifier (NN), Hidden Markov model (HMM) and Support vector machine (SVM). In this work SVM is used as a classifier.

2.7 Unicode generation

Unicode, or the Universal Character Set (UCS), was developed after all the words are recognized. After the word is decided then the respective Unicode for the word is generated. The Unicode range for Guajarati is 0A81 to 0AF1.

Chapter 3

Issue in Gujarati OCR

3.1 Vertical overlap

In Gujarati script some symbol like ૬, ૯, ૬ made up using a two different connected component.

If we segment such character using a connected component analysis then we will get two different symbols which will increase the no of classes. The total number of classes in Gujarati language is 902. Dealing with those many numbers of classes is difficult. But the property of these characters is that the two connected components are in vertical overlap. So by combining connected components having vertical overlap, we can make a one symbol and no of classes will reduce. The reduced number of classes, using this approach, comes out to be 892. The percentage overlap is calculated from the below equation.

$$percentage\ overlap = \max\left(\frac{w_{overlap}}{w_{cc1}}, \frac{w_{overlap}}{w_{cc2}}\right)$$

Where,

$w_{overlap}$ = Width of the vertical overlapping area

w_{cc1} = Width of the connected component 1

w_{cc2}= Width of the connected component 2

Using a different overlapping threshold, the algorithm tested on 120 pages and the 0.29 is selected as optimum threshold. The following table contains the statistical data.

Threshold	Number of pages tested	Number of samples getting merged	Number of Samples getting split
0.25	120	53	0
0.26	120	46	0
0.27	120	40	0
0.28	120	14	2
0.29	120	8	4
0.30	120	5	10

Table 3.1 Statistics for vertical overlap threshold

3.2 Grouping

SVM is considered to be a reliable performer among many classifiers. But SVM classifier can become unwieldy for 892 classes. If classes are divided into some groups then we can use separate SVM classifier for each of these groups. For this I have used the property of Gujarati script. The basic property of the script used is the height of the character.

For grouping I have used three properties of the Gujarati Script.

H_L- Height of the line.

W_{cc}- Width of connected component

H_{cc}- Height of the connected component

3.2.1 Grouping the special symbols:

For grouping the special symbol height and the width of the symbol is used. The width of the special symbol is compare to very less than the other base symbol.so the ratio of the width to

height (W_{cc}/H_{cc}) is calculated. And using these ratio special symbols is grouped. Another approach for grouping the special symbol is that, while scanning the each row of the segmented symbol image the number of crossing from off pixel to on pixel and on pixel to off pixel is less for the special symbols and it is very high for the base symbols. . So, if the number of these crossing is less than 4 then the segmented component is considered as a special symbol otherwise a base symbol.

Some times because of different style font, some special symbols have more crossing than four. Therefore such symbols are grouped into a special symbol and also into a base symbol. This type of grouping is called soft grouping.

3.2.2 Grouping of base Symbols

After grouping special symbol and base symbol, we can further divide base symbol into four group using height of connected component and height of the line. To grouped base symbol ratio of height of the connected component to height of the line (H_{cc} /H_L) is calculated.

If the ratio is less than some threshold then symbols are included grouped in pure consonant group or no modifier group. Because the height of the connected component is near to the line height.

If the ratio is above the threshold and connected component spans more in vertical direction above the middle of the line then it grouped into upper modifier group.

If the ratio is above the threshold and connected component spans more than the line in the vertical direction below the middle line then symbol grouped into lower modifier group.

If the ratio is much higher than the threshold and the connected component spans more both above and below the middle then it can be grouped into a group with both the modifiers group.

Based on the above approach, I have grouped the classes into 5 groups.

List of Groups:

Group1: Consonant only – single and connected character (e.g. ક ક્ય)

Group2: Consonant with upper modifier – single and connected character (e.g. કે ક્યે)

Group3: Consonant with lower modifier – single and connected character (e.g. કુ ક્યુ)

Group4: Consonant with both modifiers – single and connected character (e.g. કું ક્યું

)

Group5: Special symbols (e.g. [{ !)

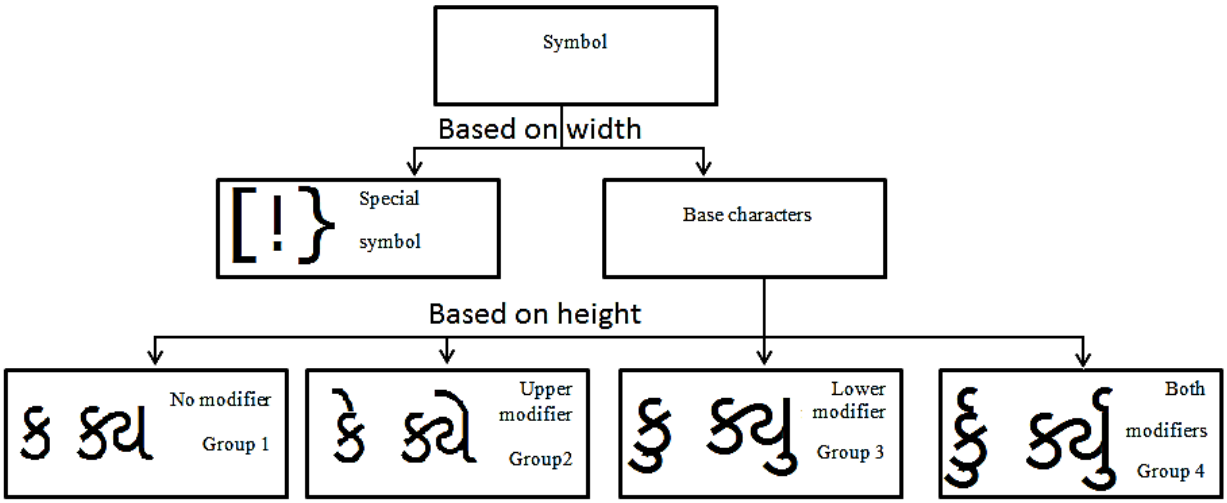


Fig 3.1 Grouping based on Height and width

3.3 Classification

SVM is very reliable classifier among the other classifiers therefore In this work SVM is used for classification.

The cross Validation Accuracy of Gujarati using SVM with linear Kernel for template as a feature is shown in Table 3.1.

Group ID	**Cross Validation Accuracy (%)**	**Value of parameter "c"**
Group 1	89.8	16
Group 2	96.3	16
Group 3	89.5	1
Group 4	93.3	16
Group 5	89.8	4

Table 3.2. Cross Validation Accuracy for Grouped Model

The weighted average Cross Validation Accuracy of the grouped data is 93.7%.

3.4 Discussion

For the ungrouped model there are 1014 Confusion pairs. Some of them are shown in Table 3.2.

Actual Character	Recognized Character
ચે	ઌે
ા	ઠ
બા	ક્યુ
ક	ઞે
ચ્ય	ઋજ

Table 3.3 Confusion pairs comparison for ungrouped model

From above table it is while using a grouped model many of the confusion pairs will be in different group. Thus, the number of confusion sets decreases in Grouped model compares to ungrouped model. Number of classes getting confused for group model is given in below table.

Group ID	**# classes confused**
Group 1	417
Group 2	376
Group 3	167
Group 4	22
Group 5	5

Table 3.4 Number of Confusion pairs for 5 Groups

Total number of confusion class's pairs for ungrouped model is 1024 and for grouped model the numbers of confusion class's pairs are 987.

3.5 collected Database

Total no of samples and the average no of samples per class for every group is shown in table 3.5. As discuss earlier some of the classis may go to two or more group.so this grouping is soft grouping. Because of soft grouping number of classes increased to 938while without grouping number of classes are 885.

The below grouping is soft grouping. As there are some classes which are in two or more groups. Because of this the total number of classes is 938 while it is 885 without grouping.

Group	Number of Classes	Total number of samples	Average number of samples per Class
Group 1 (no modifier)	312	16843	54
Group 2 (upper modifier)	382	10158	27
Group 3 (lower modifier)	196	7625	39
Group 4 (both modifiers)	48	2325	48

3.5 Total number of class per group and Average number of samples per class

3.6 Unicode Generation

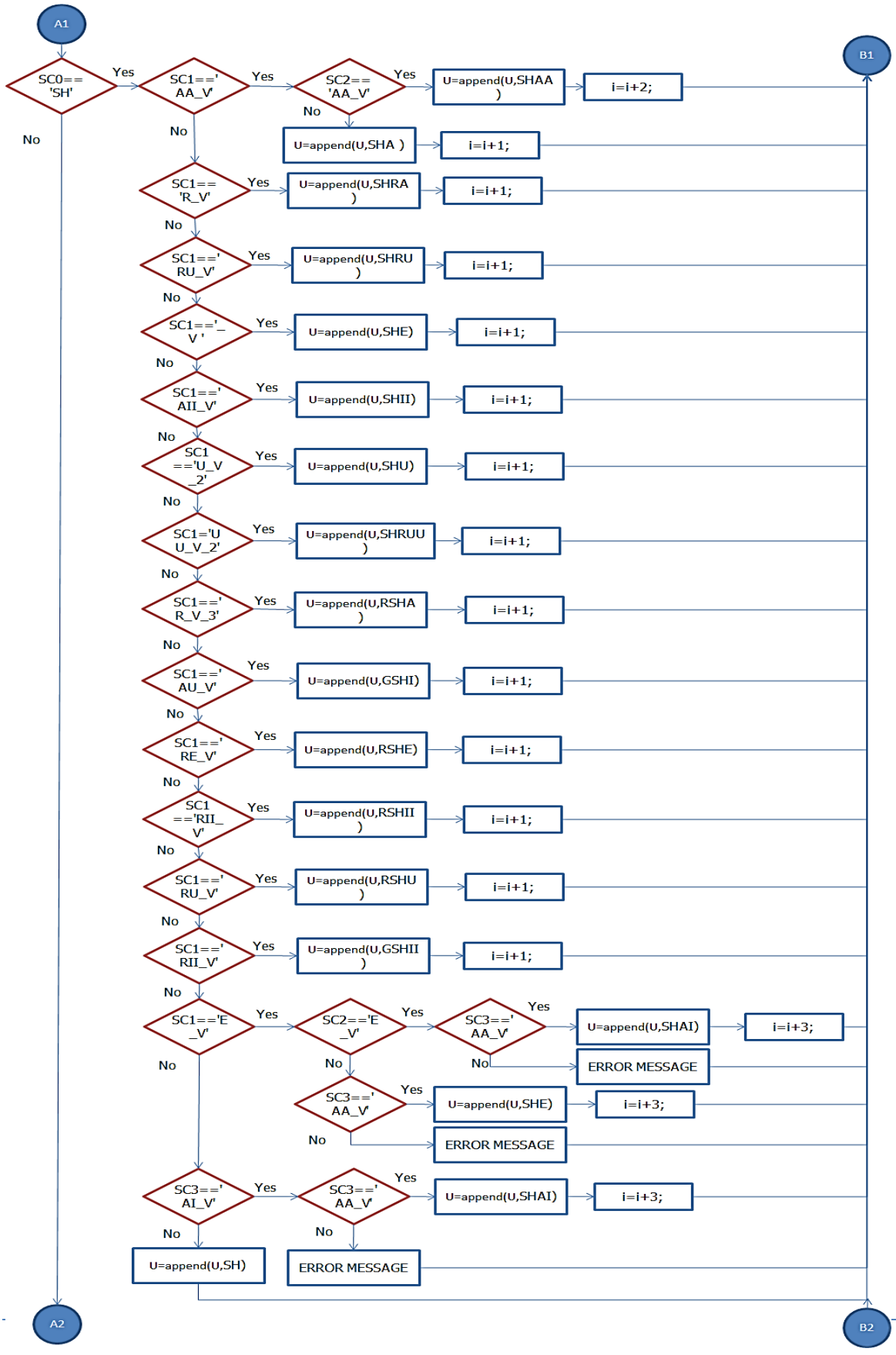
3.6.1 Unicode mapping file

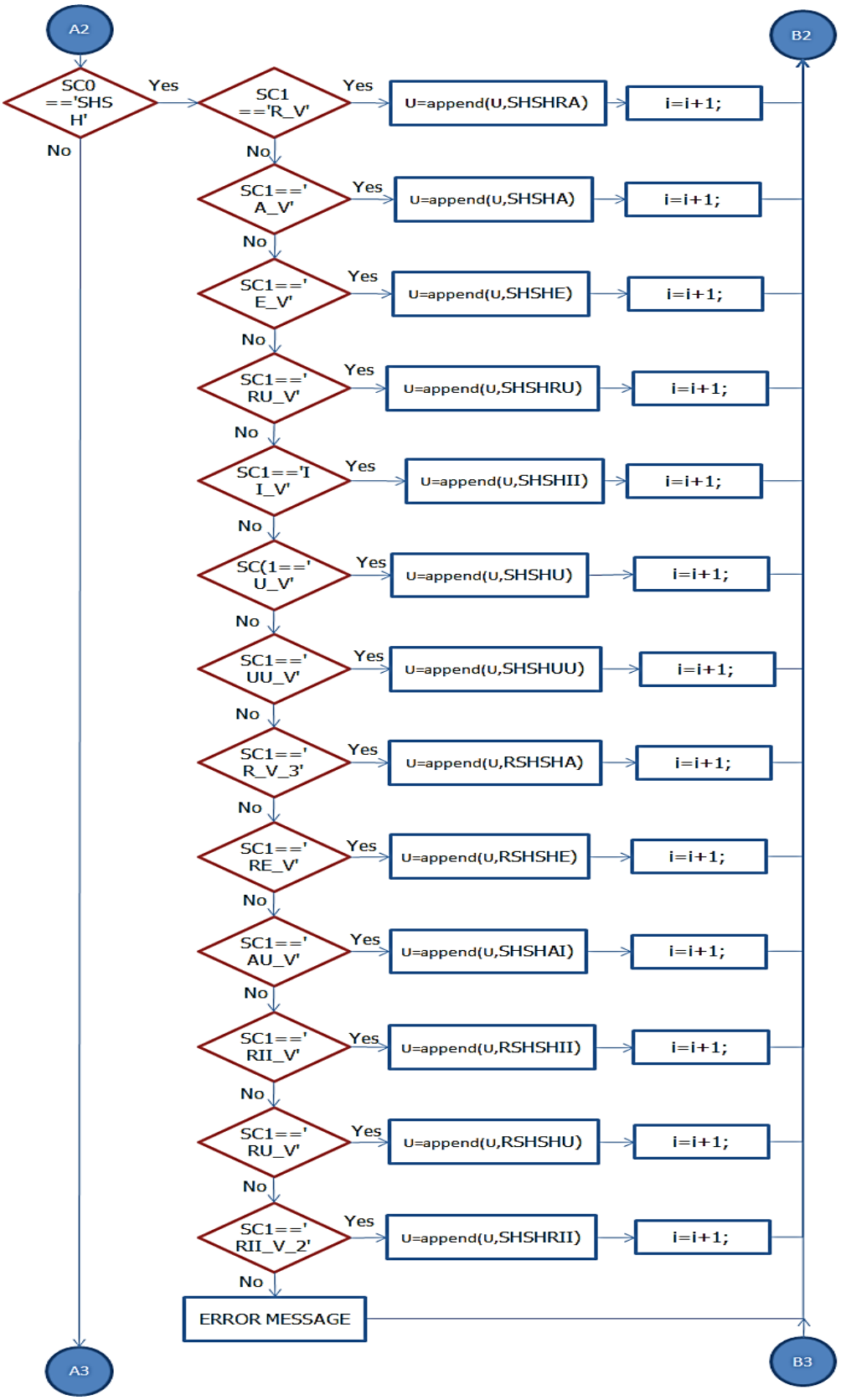
The next stage after classification is to generate the Unicode for recognized classes The Unicode mapping file gives the respective Unicode to the classes. This mapping file contains the Unicode for every Gujarati classes and its corresponding Guajarati Unicode. When SVM classified the

label, the label is converted to respect Unicode by looking up in the mapping file. Unicode mapping file is created and included in Appendix I. For Unicode generation some rules are used which is explained in next section.

3.6.2 Unicode Generation Rules

For some character are having more than one segmented, therefore before generating of Unicode, those needed to be combined. So some rules are used based on the script.following flow chart explain the rules. First flow chat explains rules for character "ર્લ" ,when first segmented component is classified .And second flow chart for jodakhar "ર્ર્લ" when first two symbol of the *jodakhar* classified consecutively.


A1
SC0== 'SH'
Yes
SC1==' AA_V'
Yes
SC2== 'AA_V'
Yes
U=append(U,SHAA)
i=i+2;
No
U=append(U,SHA)
i=i+1;
No
SC1== 'R_V'
Yes
U=append(U,SHRA)
i=i+1;
No
SC1==' RU_V'
Yes
U=append(U,SHRU)
i=i+1;
No
SC1==' _V '
Yes
U=append(U,SHE)
i=i+1;
No
SC1==' AII_V'
Yes
U=append(U,SHII)
i=i+1;
No
SC1 ==U_V _2'
Yes
U=append(U,SHU)
i=i+1;
No
SC1='U U_V_2'
Yes
U=append(U,SHRUU)
i=i+1;
No
SC1==' R_V_3'
Yes
U=append(U,RSHA)
i=i+1;
No
SC1==' AU_V'
Yes
U=append(U,GSHI)
i=i+1;
No
SC1==' RE_V'
Yes
U=append(U,RSHE)
i=i+1;
No
SC1 ==RII_ V'
Yes
U=append(U,RSHII)
i=i+1;
No
SC1==' RU_V'
Yes
U=append(U,RSHU)
i=i+1;
No
SC1==' RII_V'
Yes
U=append(U,GSHII)
i=i+1;
No
SC1=='E _V'
Yes
SC2=='E _V'
Yes
SC3==' AA_V'
Yes
U=append(U,SHAI)
i=i+3;
No
ERROR MESSAGE
No
SC3==' AA_V'
Yes
U=append(U,SHE)
i=i+3;
No
ERROR MESSAGE
No
SC3==' AI_V'
Yes
SC3==' AA_V'
Yes
U=append(U,SHAI)
i=i+3;
No
U=append(U,SH)
No
ERROR MESSAGE
A2
B1
B2


A2
SC0 =='SHS H'
Yes
No
SC1 =='R_V'
Yes
No
U=append(U,SHSHRA)
i=i+1;
SC1=='A_V'
Yes
No
U=append(U,SHSHA)
i=i+1;
SC1=='E_V'
Yes
No
U=append(U,SHSHE)
i=i+1;
SC1=='RU_V'
Yes
No
U=append(U,SHSHRU)
i=i+1;
SC1=='I I_V'
Yes
No
U=append(U,SHSHII)
i=i+1;
SC(1=='U_V'
Yes
No
U=append(U,SHSHU)
i=i+1;
SC1=='UU_V'
Yes
No
U=append(U,SHSHUU)
i=i+1;
SC1=='R_V_3'
Yes
No
U=append(U,RSHSHA)
i=i+1;
SC1=='RE_V'
Yes
No
U=append(U,RSHSHE)
i=i+1;
SC1=='AU_V'
Yes
No
U=append(U,SHSHAI)
i=i+1;
SC1=='RII_V'
Yes
No
U=append(U,RSHSHII)
i=i+1;
SC1=='RU_V'
Yes
No
U=append(U,RSHSHU)
i=i+1;
SC1=='RII_V_2'
Yes
No
U=append(U,SHSHRII)
i=i+1;
ERROR MESSAGE
A3
B2
B3

Chapter 4

Result of OCR System and Discussion

4.1 Results

Any OCR system should perform well for wide range of fonts, printing ink and for different types of paper used.in this project I have calculated the performance evaluation of the Guajarati OCR with different 17 books with different fonts and print types. For all the books annotation are available for every page.

Annotation is ground truth text typed for all the text blocks available in the pages.

Following figure shows the output of a Gujarati OCR system which is integrated in IISc print to Braille Tool. The left image is scanned image of a text document as a input to the OCR system and the right image is the output of the Gujarati OCR system.

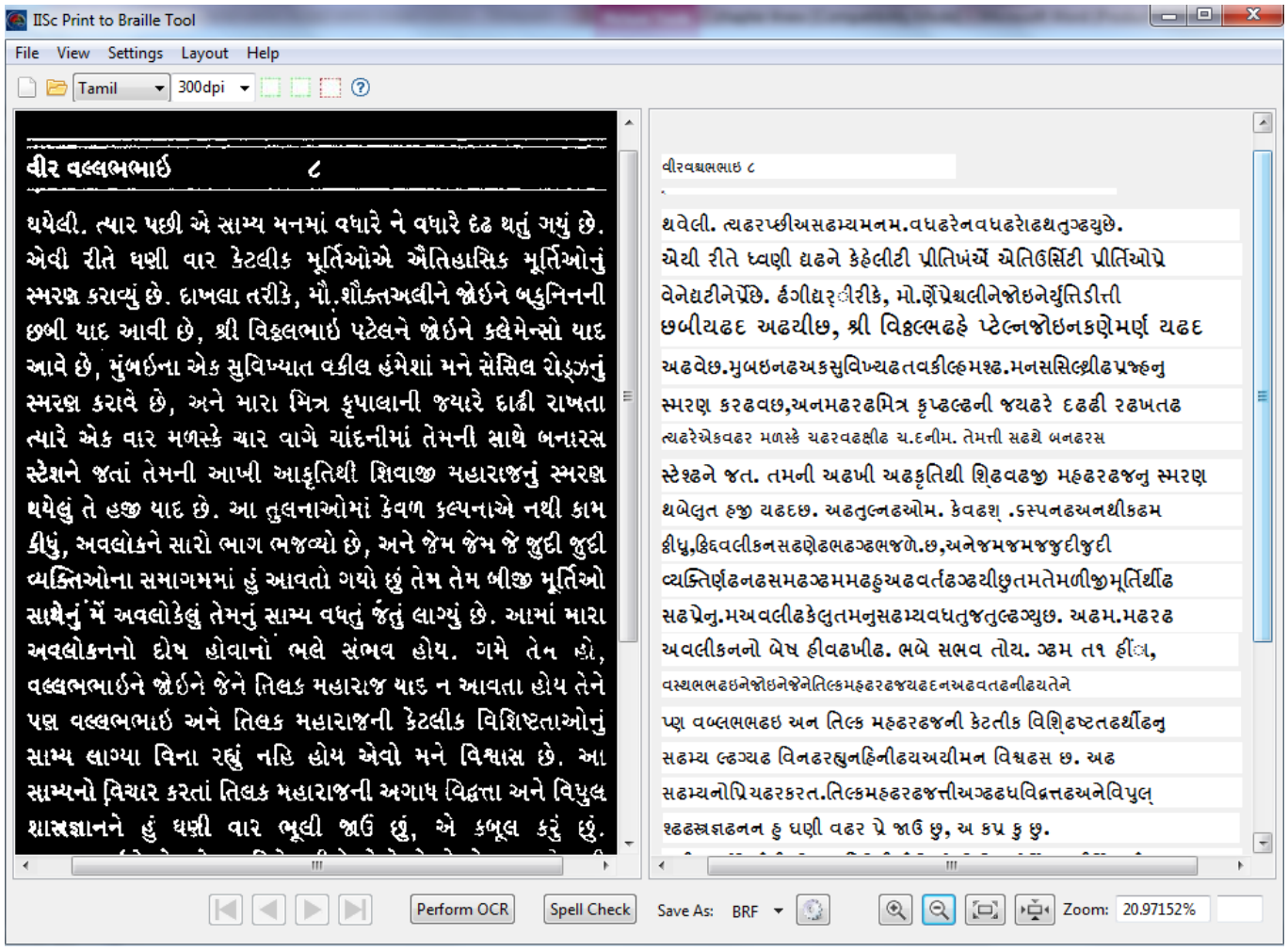


Fig.4.1 scanned image and its OCR output

OCR system is tested on the 17 books and the data is given in the table 4.1..

In this table third column contains annotation size of the book which is nothing but the Total number of Unicode for the book, fourth column is the Levenhsten edit distance between ground truth text and OCR output text, fifth column shows the mismatch distance which is nothing but the number of incorrectly recognized character. Sixth column shows the substitution error rate which is the ratio of mismatch distance to the annotation size. And the last column shows the total error rate which is the ratio of Levenhstein edit distance to the annotation size.

Book Name	Total pages	Annotation Size	Edit Distance	Mismatch Distance	Substitution Error Rate	Total Error Rate
1. Veer Vallabhbhai	66	65447	9954	4145	6.33	15.21
2. Chinma Choppan Divas	94	148459	5743	37470	25.24	40.91
3. Aakhare Usha Ugikhar	201	52781	43211	23387	44.31	81.87
4. Manav Adhikar	184	396682	112937	57085	14.39	28.47
5. Panchamruth	413	318524	91596	34394	10.8	28.76
6. Atmakatha	459	546684	308455	136792	25.02	56.42
7. Maun Man	257	200730	83598	50111	24.96	41.65
8. Prem Ane Purusharth	427	446211	388515	185151	41.49	87.07
9. Vishvakarma Bhajan	212	169528	120937	79807	47.08	71.34
10. Pushthi Ras Madhurya	283	208822	37765	20475	9.81	18.08
11. Saradar Patel	513	1033200	344101	162293	15.71	33.3
12. Panch Katha	200	112369	52497	7294	6.49	46.72
13. Champak	28	22380	5411	3434	15.34	24.18
14. Gujarat No Jay	430	449376	129083	83040	18.48	28.72
15.Samaragan	245	248986	75544	43032	17.28	30.34
16. Sinhasan	160	79760	19052	11987	15.03	23.89
17. Hitopdesh Ni Vato	141	149726	46396	26883	17.95	30.99

Table 4.1 Performance Evaluation Result

4.2 Discussion

The total error rates for the books numbered 3, 6, 8 and 9 are more than 50%. This is because of the broken characters, merged characters etc. In this work, noise removal is not implemented and so the presence of noise merges consecutive characters because of printing and scanning.

In this work OCR system is developed only for Guajarati script and therefore OCR system unable to handle the Multi-language books so for book numbered 9 containing devenagri scripts and book no 5 containing English text script, OCR system given higher error rate.

The poor performance for book numbered 8 is because of the improper line segmentation, as the lines are very closely printed and overlaps with one another.

Chapter 5
Work on Print to Braille tool

IISc Print to Braille Tool is a Graphical User Interface (GUI) frontend for OCR systems, designed to significantly reduce the time spent by users in correcting the OCR output, as well as converting the text output into Braille or other text formats.
As shown in fig 5.1, the input page is in the left half and its OCRed text output is in the right half of the Print to Braille Tool. The text font size and the position of text lines are so adjusted that the text displayed on right-half and image displayed on left-half are aligned with each other, line by line. There are zoom in and zoom out options are available on the right half of bottom toolbar below text region. Zooming will simultaneously zoom both the image region as well as the text region. At higher zoom levels, scroll-bars (horizontal and vertical) appear for both images and text regions. Scrolling from either the text region or image region will scroll the other region automatically, thereby always keeping the image and text regions in sync. When user clicks the buttons for navigating to next or previous page, both the left-side image and right-side text output are updated to display the image and OCR text of the new page.

The text lines displayed on the right side are editable. User can quickly compare the OCR output text displayed on right side with the original printed text displayed on left side and if there are any corrections needed, directly edit the right side text. Print to Braille tool automatically
saves all edits immediately, and hence even if user navigates to some other page or closes the tool and reopens to the same page, the earlier corrections made are kept as it is and shown to user.

Once all corrections are made, user can save output in RTF for normal text and in Braille formats for braille cell. Users have also option to save one page or all the pages in desired formats.

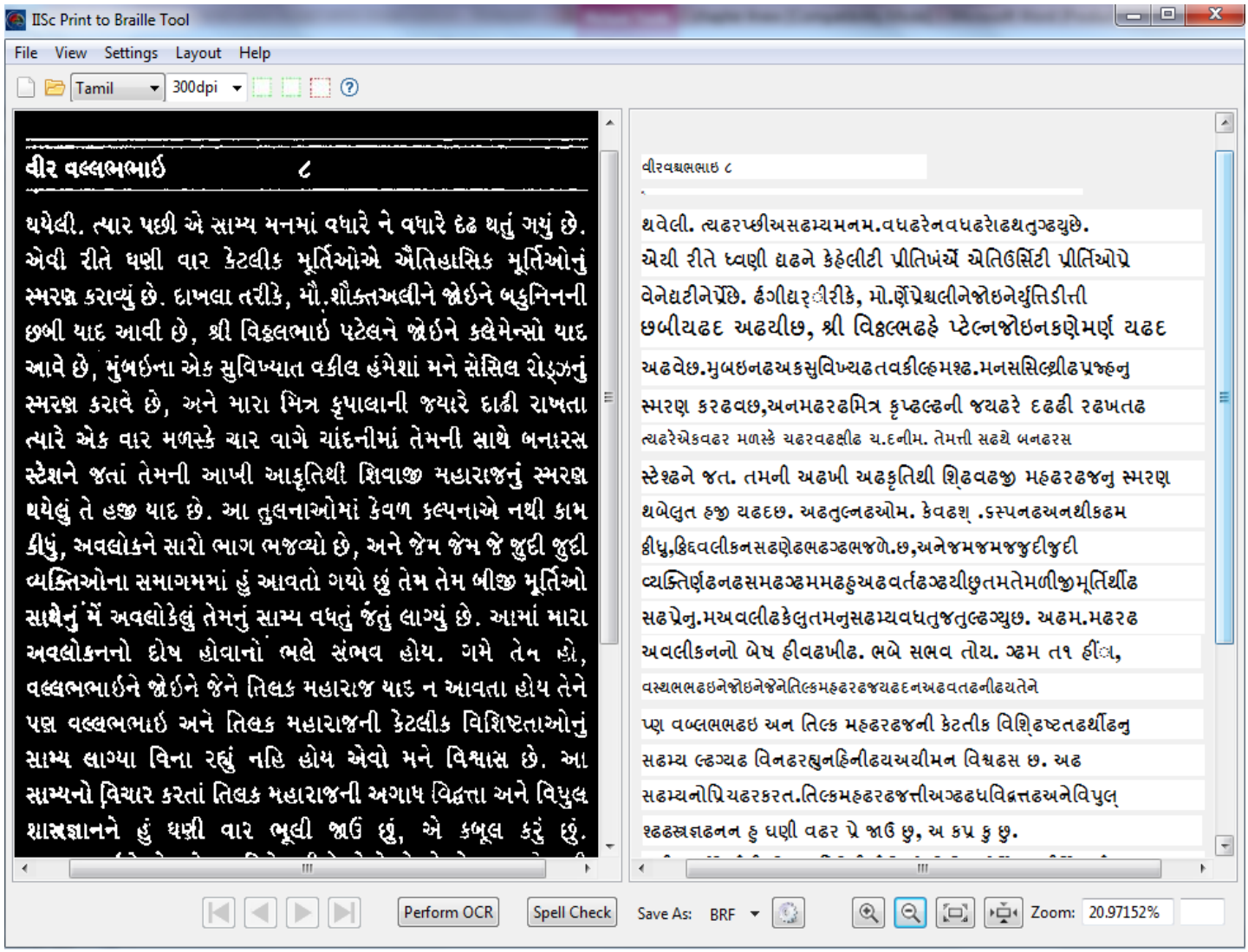


Fig.5.1 print to braille tool output

5.1 Gujarati Unicode to Braille cell

After performing the OCR system on scanned image of Guajarati page, the next step is to convert the Gujarati text file into a corresponding braille Unicode. Below image show the character and corresponding representation in braille Lippi. To convert the Gujarati character into a braille ASCII, mapping file is created which is included in appendix II. This mapping file

contains the Unicode for every Gujarati symbols and its corresponding Braille ASCII code. When OCR output text is available, the Unicodes in them are converted to Braille codes by looking up in the mapping file.

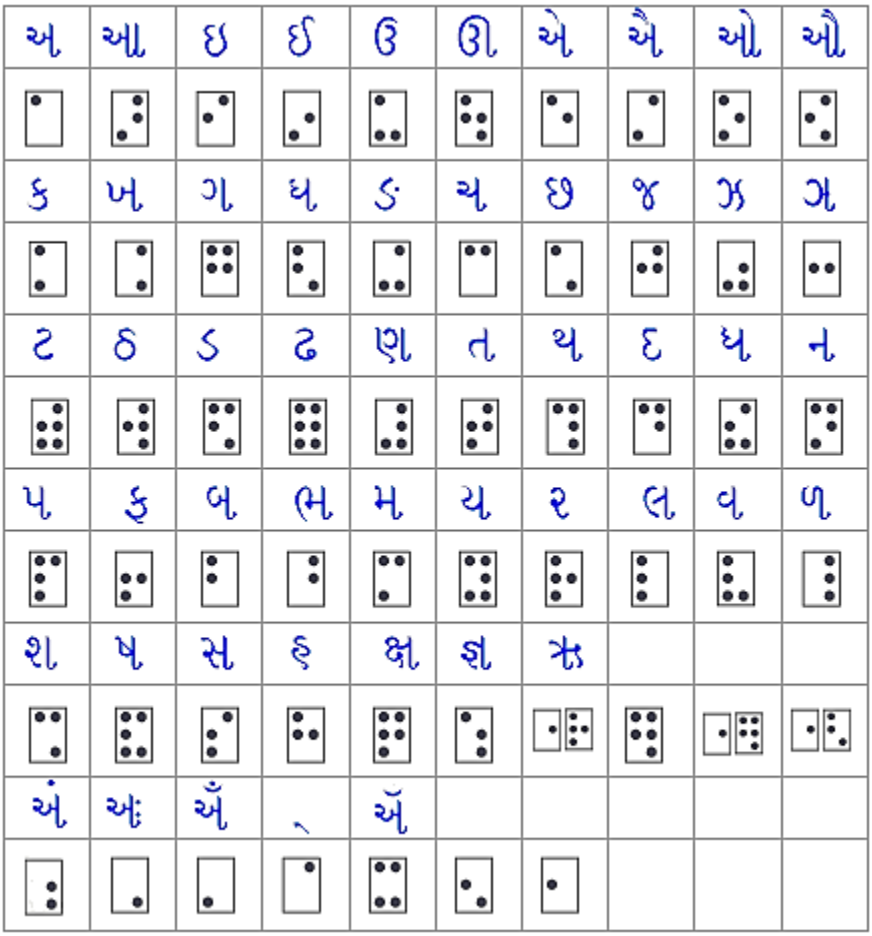
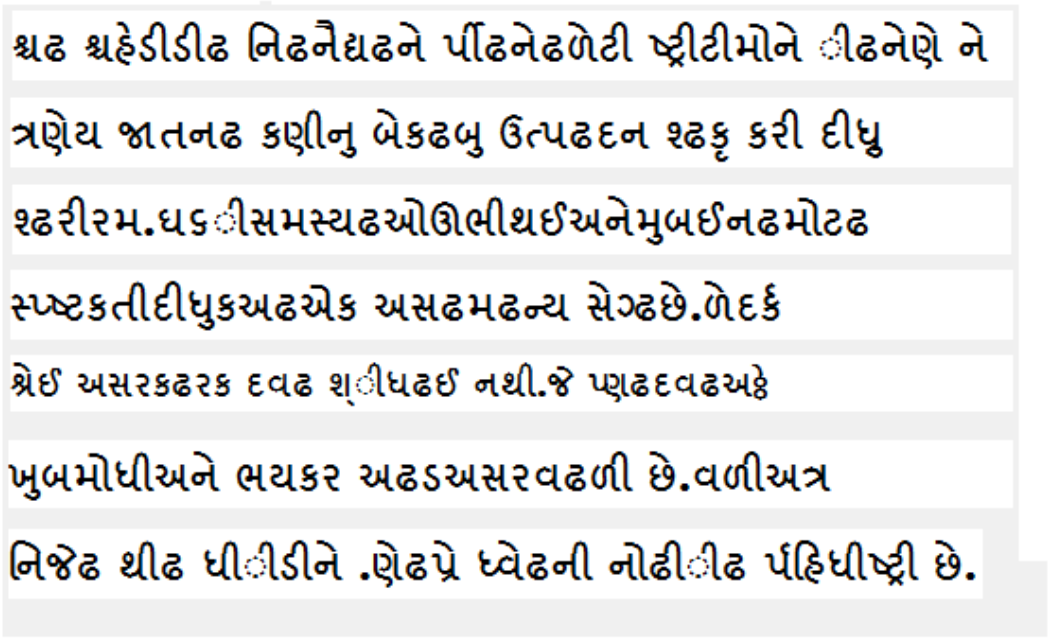


Table 5.1 Guajarati character and corresponding braille representation.

Below figure shows the result of a print to braille tool. Upper image is the output of a Gujarati OCR and the bottom image shows the braille image.



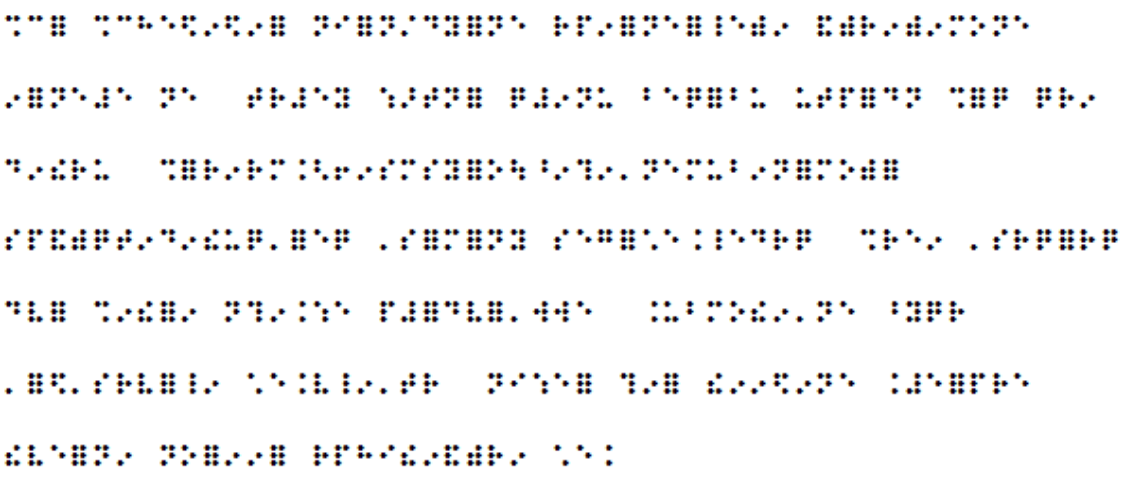


Fig 5.2 Gujarati text Unicode converted into braille cell

5.2 Gujarati Dictionary for spell check

After performing the OCR system on scanned image of Guajarati page, In Print to Braille tool "spell Check" option is available, below the output. Which will check the spell of a word using a dictionary and if word is not found in the dictionary then it will suggest the words from the Dictionary. Therefore, form 25 books annotated data 1,15,070 unique Gujarati words are extracted for spell check.

5.3 Image comparison for word correction

In Print to Braille tool there is an option for spell checking which displays suggestion for every word image. This feature can be used to correct the OCR output without the need for user to type the correct word (if the suggested word is in the dictionary). This spell check has an option of "replace all" which replaces all the occurrences of a particular word to the suggested word. However, there arises a problem, where even correctly recognized words might be replaced when blindly do the "replace all" option. So there has to be a way where the spell checker uses some sort of input from the word image also, rather than only using the input from the OCRed text. In the next section, we will discuss about a word image comparator to check whether the two given word images are similar or not, based on which the performance of spell checker can be improved.

5.3.1 Image comparisons

In this work for comparing the images i have used algorithm described by Zhou Wang [8].

The system is shown in Fig. 5.4. Suppose x and y are two images to be compared, which have been aligned with each other (e.g., spatial patches extracted from each image). If we consider one of the images to have perfect quality, then the similarity measure can serve as a quantitative measurement of the quality of the second image. The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure.

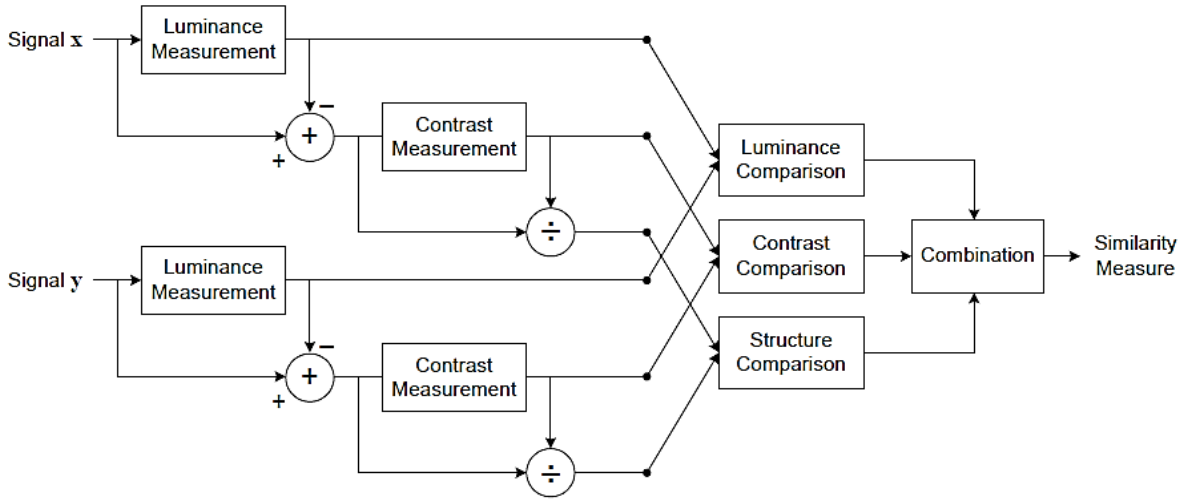


Fig 5.3 Diagram of the structural similarity (SSIM) measurement system.

First, the luminance of each image is compared. Assuming discrete images, this is estimated as the mean intensity. Luminance comparison function is given by

$$l(x,y) = \frac{2u_x u_y + c_1}{u_x^2 + u_y^2 + c_1}$$

Where u_x and u_y are the mean intensity given by

$$u_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad u_y = \frac{1}{N}\sum_{i=1}^{N} y_i$$

One constant c_1 is included to avoid instability when $u_x^2 + u_y^2$ is very close to zero.

In this work c_1=$(k_{1,}L)^2$,where L is the dynamic range of the pixel values (255 for 8-bit grayscale images) and $k_{1,}$=0.01.

Contrast comparison function is given by

$$C(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

Constant $c_2$is included to avoid instability when $\sigma_x^2 + \sigma_y^2$ is very close to zero and c_2=$(k_{2,}L)^2$

Where σ_x and σ_y are the standard deviation calculated by below equations.

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - u_x)^2\right)^{\frac{1}{2}} \qquad \sigma_y = \left(\frac{1}{N-1}\sum_{i=1}^{N}(y_i - u_y)^2\right)^{\frac{1}{2}}$$

Structure comparison function is given by

$$l(x,y) = \frac{2\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

Where c_3=c_2/2 and σ_{xy} is correlation coefficient given by

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - u_x)(y_i - u_y)$$

Structural SIMilarity (SSIM) index between signals x and y is calculated using below equation.

$$SSIM(x,y) = \frac{(2u_xu_y + c_1)}{(u_x^2 + u_y^2 + c_1)}\frac{(2\sigma_x\sigma_y + c_2)}{(\sigma_x^2 + \sigma_y^2 + c_2)}$$

5.3.2 Image Quality Assessment using SSIM index

For image quality assessment, it is useful to apply the SSIM index locally rather than globally, The local statistics u_x, u_y and σ_{xy} are computed within an 11 X 11circular symmetric Gaussian window, which moves pixel-by-pixel over the entire image.

Gaussian weighting function $W = \{W_i | i = 1,2,3 \ldots, N\}$ with standard deviation of 1.5 samples, normalized to unit sum ($\sum_{i=1}^{N} W_i = 1$).The estimates of local statisticsu_x, $u_y$and σ_{xy} are then modified accordingly as

$$u_x = \frac{1}{N}\sum_{i=1}^{N} W_i x_i$$

$$\sigma_x = \left(\sum_{i=1}^{N} W_i (x_i - u_x)^2 \right)^{\frac{1}{2}}$$

$$\sigma_{xy} = \sum_{i=1}^{N} W_i (x_i - u_x)(y_i - u_y)$$

We require a single overall quality measure of the entire image. We use a mean SSIM (MSSIM) index to evaluate the overall image quality:

$$MSSIM(X, Y) = \frac{1}{M} \sum_{i=1}^{M} SSIM(X_j, Y_j)$$

Where X and Y are the reference and the distorted images, respectively; X_j and Y_j are the image contents at the *j*-th local window, and *M* is the number of local windows in the image.

5.4 Result and discussion

I have tested this algorithm on different 110 pairs of images and calculated MSSIM. Some of the compared image is shown in table 5.1. When x and y are same MSSIM will be 1.so using a statistical data I have chosen 0.37 as a threshold. When MSSIM will above 0.37 both image are taken as similar image and for below 0.37, images are different.

Image X	Image Y	MSSIM
		0.0736
		0.5565
		0.2690
		0.6355
		0.3163

Table 5.2 compared images and their result

Appendix I

Unicode Mapping

The following list shows the Unicode mapping from class label to respective Unicode.

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
1	KA	ક	1		24	RKHA	ખ્	2
2	KII	કી	2		25	KHRA	ખ્ર	1
3	KI	કિ	2		26	KHRE	ખ્રે	2
4	KU	કુ	3		24	RKHA	ખ્	2
5	KUU	કૂ	3		25	KHRA	ખ્ર	1
6	KE	કે	2		26	KHRE	ખ્રે	2
7	KAI	કૈ	2		24	RKHA	ખ્	2
8	KRU	કૃ	3		25	KHRA	ખ્ર	1
9	KRA	ક્ર	1		26	KHRE	ખ્રે	2
10	RKA	ક્	2		27	KHRAI	ખ્રૈ	2
11	KTA	ક્ત	1		28	KHRU_2	ખ્રુ	3
12	KKA	ક્ક	1		29	KHRUU	ખ્રૂ	3
13	KYA	ક્ય	1		30	KHKHA	ખ્ખ	1
14	KKII	ક્કી	2		31	KHKHE	ખ્ખે	2
15	KYU	ક્યુ	3		32	KHKHU	ખ્ખુ	3
16	KHA	ખ	1		33	KHKHUU	ખ્ખૂ	3
17	KHII	ખી	2		34	KHYA	ખ્ય	1
18	KHI	ખિ	2		35	KHYE	ખ્યે	2
19	KHU	ખુ	3		36	KHYU	ખ્યુ	3
20	KHUU	ખૂ	3		37	KHYUU	ખ્યૂ	3
21	KHE	ખે	2		38	KHTA	ખ્ત	1
22	KHAI	ખૈ	2		39	KHTE	ખ્તે	2
23	KHRU	ખૃ	3		40	KHTU	ખ્તુ	3

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
41	KHTUU	ખ્તૂ	3		68	NNAA	ણા	1
42	KHL	ખ્લ્	1		69	GDA	ગદ	1
43	KHLE	ખ્લે	2		70	GDE	ગદે	2
44	KHLU	ખ્લુ	3		71	RMAA	ર્મા	2
45	KHVA	ખ્વ	1		72	SAA	સા	1
46	KHVE	ખ્વે	2		73	AA_V_2	ા	1
47	KH	ખ્	1		74	GG	ગ્ગ્	1
48	KHVU	ખ્વુ	3		75	GGU	ગ્ગુ	3
49	KHDDA	ખ્ડ	1		76	RTAA	ર્તા	2
50	KHDDE	ખ્ડે	2		77	RYAA	ર્યા	2
51	KHDDAI	ખ્ડૈ	2		78	RYO	ર્યો	2
52	KHDDU	ખ્ડુ	3		79	GDHA	ગધ	2
53	KHDDUU	ખ્ડૂ	3		80	BO	બો	2
54	GRII	ગ્રી	2		81	SSO	ષો	2
55	GRI	ગ્રિ	2		82	DO	દો	2
56	GRA	ગ્ર	1		83	GNA	ગ્ન	1
57	G	ગ્	1		84	GRE	ગ્રે	2
58	GA	ગ	1		85	GRO	ગ્રો	2
59	R_V	્ર	1		86	VO	વો	2
60	LLYAA	લ્યા	1		87	DHO	ધો	2
61	CCH	ચ્છ	1		88	GRUU	ગ્રૂ	3
62	RR	ૠ	1		90	GLA	ગ્લ	1
63	RAA	રા	1		89	GLE	ગ્લે	2
64	PAA	પા	1		91	G_CE	ગૅ	2
65	SSAA	ષા	1		92	HO	હો	2
66	GAA	ગા	1		93	RU_V	ૃ	3
67	DAA	દા	1		94	GTTHA	ગઠ	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
95	KHO	ખો	2		123	CYU	ચ્યુ	3
96	CO	ચો	2		124	CYUU	ચ્યૂ	3
97	GHO	ઘો	2		125	RCA	ર્ચ	2
98	VYO	વ્યો	1		126	CYII	ચ્ચી	2
99	SAU	સૌ	2		127	CYI	ચ્ચિ	2
100	RVAA	ર્વા	2		128	CYE	ચ્ચે	2
101	GYA	ગ્ય	1		129	CYAI	ચ્ચૈ	2
102	GYE	ગ્યે	2		130	CCE	ચ્ચે	2
103	GYU	ગ્યુ	3		123	CYU	ચ્યુ	3
104	YO	યો	2		124	CYUU	ચ્યૂ	3
105	GDDA	ગ્ડ	1		125	RCA	ર્ચ	2
106	PO	પો	2		126	CYII	ચ્ચી	2
107	TO	તો	2		127	CYI	ચ્ચિ	2
108	MO	મો	2		128	CYE	ચ્ચે	2
109	THO	થો	2		129	CYAI	ચ્ચૈ	2
110	TYAA	ત્યા	1		130	CCE	ચ્ચે	2
111	GVA	ગ્વ	1		123	CYU	ચ્યુ	3
112	GVE	ગ્વે	2		124	CYUU	ચ્યૂ	3
113	NO	નો	2		125	RCA	ર્ચ	2
114	GVI	ગ્વિ	2		126	CYII	ચ્ચી	2
115	GVII	ગ્વી	2		127	CYI	ચ્ચિ	2
116	VYAA	વ્યા	1		128	CYE	ચ્ચે	2
117	M	મ્	1		129	CYAI	ચ્ચૈ	2
118	DHYAA	ધ્યા	2		130	CCE	ચ્ચે	2
119	STHAA	સ્થા	1		131	CYA	ચ્ય	1
120	GHII	ઘી	2		132	CCA	ચ્ચ	1
121	GHA	ઘ	1		133	CCHA	ચ્છ	1

Class No.	Class Name	Unicode	Group no	Class No.	Class Name	Unicode	Group no
134	CCHU	ચ્છુ	3	161	JU	જુ	3
135	CI	ચિ	2	162	JE	જે	2
136	CII	ચી	2	163	JUU	જૂ	3
137	CUU	ચૂ	3	164	JO	જો	2
138	CA	ચ	1	165	JA	જ	1
139	RCYU	ર્ચ્યુ	4	166	JAU	જૌ	2
140	C_CE	ચે	2	167	J_CE	જે	2
141	CRU	ચૃ	3	168	JHA	ઝ	1
142	CRA	ચ્ર	1	169	JHE	ઝે	2
143	CRE	ચ્રે	2	170	JHAI	ઝૈ	2
144	CRAI	ચ્રૈ	2	171	JHO	ઝો	2
145	CE	ચે	2	172	STAA	સ્તા	1
146	CHAI	છૈ	2	173	JHJHA	ઝ્ઝ	1
147	CHU	છુ	3	174	JHRUJA	ઝૃજ	3
148	CHII	છી	2	175	TTYU	ટ્યુ	3
149	CHE	છે	2	176	TTYE	ટ્યે	2
150	CHA	છ	1	177	TTYAI	ટ્યૈ	2
151	CHCHA	છ્છ	1	178	TTRA	ટ્ર	1
152	CHYA	છ્ય	1	179	TTRE	ટ્રે	2
153	CHUU	છૂ	3	180	KYAA	ક્યા	1
154	CHRU	છૃ	3	181	TTRU	ટ્રુ	3
155	CH_CE	છે	2	182	TTTTA	ટ્ટ	1
156	CHI	છિ	2	183	TTE	ટે	2
157	CHYU	છયુ	3	184	TTII	ટી	2
158	JYA	જ્ય	1	185	TTI	ટિ	2
159	JAA	જા	1	186	TTA	ટ	1
160	JII	જી	1	187	RTTA	ર્ટ	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
188	TTTTII	ટ્ટી	2		215	DDHA	ઢ	1
189	TTU	ટ્ટુ	3		216	DDHYA	ઢ્ય	1
190	TTKA	ટ્ક	1		217	NN	ણ્	1
191	TTLA	ટ્લ	1		218	TRI	ત્રિ	2
192	TT_CE	ટ	2		219	TRII	ત્રી	2
193	TTVA	ટ્વ	1		220	TTI_2	ત્તિ	2
194	TTHTTHA	ઠ	1		221	TSA	ત્સ	1
195	TTHTTHII	ઠ્ઠી	2		222	TYA	ત્ય	1
196	TTHTTHI	ઠ્ઠિ	2		223	RTA	ર્ત	2
197	TTHTTHE	ઠ્ઠે	2		224	TVA	ત્વ	1
198	TTHTTHAI	ઠ્ઠૈ	2		225	TTA_2	ત્ત	1
199	TTHA	ઠ	1		226	TTE_2	ત્તે	2
200	TTHI	ઠિ	2		227	TTII_2	ત્તી	2
201	TTHII	ઠી	2		228	RTI	ર્તિ	2
202	TTHU	ઠુ	3		229	TNII	ત્ની	2
203	TTHUU	ઠૂ	3		230	TU	તુ	3
204	TTHE	ઠે	2		231	TE	તે	2
205	DDE	ડે	2		232	TII	તી	2
206	DDII	ડી	2		233	TI	તિ	2
207	DDI	ડિ	2		234	TA	ત	1
208	DDU	ડુ	3		235	TUU	તૂ	3
209	DDUU	ડૂ	3		236	TRU	તૃ	3
210	DDA	ડ	1		237	TRA	ત્ર	1
211	DDYA	ડ્ય	1		238	TNA	ત્ન	1
212	DDYU	ડ્યુ	3		239	RTU	ર્તુ	4
213	DDHII	ઢી	2		240	TRUU	ત્રૂ	3
214	DDHE	ઢે	2		241	TRE	ત્રે	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
242	TRU_2	ત્રુ	3		269	DRII	દ્રી	2
243	TYU	ત્યુ	3		270	DHYE	ધ્યે	2
244	THRII	થ્રી	2		271	DHYA	ધ્ય	2
245	THRI	થ્રિ	2		272	DHI	ધિ	2
246	RTHE	ર્થે	2		273	DHU	ધુ	3
247	RTHA	ર્થ	2		274	DHII	ધી	2
248	THTHA	થ્થ	1		275	DHA	ધ	2
249	THII	થી	2		276	DHUU	ધૂ	4
250	THE	થે	2		277	DHDHA	ધ્ધ	2
251	THA	થ	1		278	RDHU	ર્ધુ	4
252	TH	થ્	1		279	NME	ન્મે	2
253	THU	થુ	3		280	NJA	ન્જ	1
254	DDII_2	દ્દી	2		281	NYA	ન્ય	1
255	DVA	દ્વ	1		282	NTA	ન્ત	1
256	DRA	દ્ર	1		283	NNA_2	ન્ન	1
257	DDHA_2	દ્ધ	1		284	NE	ને	2
258	DYA	દ્ય	2		285	NI	નિ	2
259	DUU	દૂ	3		286	NU	નુ	3
260	DII	દી	2		287	NII	ની	2
261	DI	દિ	2		288	NA	ન	1
262	DE	દે	2		289	NDHA	ન્ધ	2
263	DU	દુ	3		290	NTHRA	ન્થ્ર	1
264	DA	દ	1		291	NRU	નૃ	3
265	D	દ્	1		292	RE	રે	2
266	DAI	દૈ	2		293	NKHA	ન્ખ	1
267	DMA	દ્મ	1		294	NSA	ન્સ	1
268	RDAI	દૅ	2		295	NNA_2	ન્ન	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
296	RA	ર	1		323	PSH	ષ્	1
297	NVA	ન્વ	1		324	PCA	પ્ચ	1
298	NKA	ન્ક	1		325	PKA	પ્ક	1
299	NTTA	ન્ટ	1		326	PTTA	પ્ટ	1
300	NDDA	ન્ડ	1		327	MNA	મ્ન	1
301	RTHYA	ર્થ્ય	2		328	PNA	પ્ન	1
302	NTHA	ન્થ	1		329	PYA	પ્ય	1
303	NDA	ન્દ	1		330	PAI	પૈ	2
304	MPRA	મ્પ્ર	1		331	PPA	પ્પ	1
305	NMA	ન્મ	1		332	PL	પ્લ્	1
306	NMU	ન્મુ	3		333	PI	પિ	2
307	NYE	ન્યે	2		334	PRII	પ્રી	2
308	NYU	ન્યુ	3		335	PHE	ફે	2
309	NSH	ન્શ્	1		336	PHU	ફુ	3
310	PRE	પ્રે	2		337	PHUU	ફૂ	3
311	PRI	પ્રિ	2		338	PHA	ફ	1
312	PTTE	પ્ટે	2		339	BDII	બ્દી	2
313	PYU	પ્યુ	3		340	BDHA	બ્ધ	2
314	PTA	પ્ત	1		341	BRA	બ્ર	1
315	RPA	પ્ર	2		342	BDA	બ્દ	1
316	PRA	પ્ર	1		343	BII	બી	2
317	PU	પુ	3		344	BUU	બૂ	3
318	PRU	પૃ	3		345	B	બ્	1
319	PE	પે	2		346	BU	બુ	3
320	PII	પી	2		347	BE	બે	2
321	PUU	પૂ	3		348	BA	બ	1
322	PA	પ	1		349	BRU	બૃ	3

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
350	BBA	બ્બ	1		377	BHA	ભ	1
351	RBA	બ્ર	2		378	BHDDA	ભ્ડ	1
352	BDE	બ્દે	2		379	BHRA	ભ્ર	1
353	RME	મ્રૅ	2		380	BHL	ભ્લ્	1
354	MYA	મ્ય	1		381	BHYA	ભ્ય	1
355	MPHA	મ્ફ	1		382	BHRU	ભૃ	3
356	MBA	મ્બ	1		383	RYU	ર્યુ	4
357	MTTHA	મ્ઠ	1		384	RYA	ર્ય	2
358	MDDA	મ્ડ	1		385	YA	ય	1
359	MNN	મ્ણ્	1		386	YU	યુ	3
360	MCHA	મ્છ	1		387	THAI	થૈ	2
361	MKA	મ્ક	1		388	YRU	યૃ	3
362	MKHA	મ્ખ	1		389	YE	યે	2
363	MLLA	મ્ળ	1		390	YYA	ય્ય	1
364	RMA	મ્ર	2		391	YRA	ય્ર	1
365	MBHE	મ્ભે	2		392	RE	રે	2
366	ME	મે	2		393	RUU	રૂ	3
367	MI	મિ	2		394	RU	રુ	3
368	MU	મુ	3		395	RII	રી	2
369	MII	મી	2		396	NTRA	ન્ત્ર	1
370	MUU	મૂ	3		397	YU_V	્યુ	3
371	MA	મ	1		398	YA_V	્ય	1
372	MMA	મ્મ	1		399	AUU_V	ૂ	3
373	BHE	ભે	2		400	LLA_2_v	્લ્	1
374	BHUU	ભૂ	3		401	AII_V	ી	2
375	BHU	ભુ	3		402	U_V_2	ુ	3
376	BHII	ભી	2		403	A_V_2	્	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
404	L_2	લ્	1		431	VI	વિ	2
405	L	લ્	1		432	VII	વી	2
406	CA_V	્ચ	1		433	VE	વે	2
407	CII_V	્ચી	2		434	VU	વુ	3
408	CI_V	્ચિ	2		435	VRU	વૃ	3
409	CU_V	્ચુ	3		436	VAI	વૈ	2
410	LYUU	લ્યૂ	3		437	VA	વ	1
411	TTA_V	્ટ	1		438	VVA	વ્વ	1
412	TTII_V	્ટી	2		439	VVII	વ્વી	2
413	LTTI	લ્ટિ	2		440	VVI	વ્વિ	2
414	TTU_V	્ટુ	3		441	VRA	વ્	1
415	TTUU_V	્ટૂ	3		442	VTTA	વ્ટ	1
416	DDA_V	્ડ	1		443	VTTI	વ્ટિ	2
417	TA_V	્ત	1		444	VTTII	વ્ટી	3
418	DA_V	્દ	1		445	VTTHA	વ્ઠ	1
419	PA_V	્પ	1		446	VTTHII	વ્ઠી	2
420	BA_V	્બ	1		447	VTTHI	વ્ઠિ	2
421	LRA	લ્	1		448	VKA	વ્ક	1
422	VA_V	્વ	1		449	VL	વ્લ્	1
423	SA_V	્સ	1		450	SHRII	શ્રી	2
424	HA_V	્હ	1		451	SHL	શ્લ્	1
425	LRU	લૃ	3		452	SCA	શ્ર	1
426	R_V	ર્	1		453	SCHWA	શ્ર	1
427	VYU	વ્યુ	3		454	SRA	શ્ર	1
428	RVE	વ્રે	2		455	SRNA	શ્ન	1
429	RVA	વ્ર	2		456	KHYAA	ખ્યા	1
430	VYA	વ્ય	1		457	SH	શ્	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
458	SHU	શુ	3		485	SSTTA	ષ્ટ	1
459	STRI	સ્ત્રી	2		486	SSKA	ષ્ક	1
460	STTRII	સ્ટ્રી	2		487	RSSA	ર્ષ્	2
461	STRA	સ્ત્ર	1		488	SSTTHA	ષ્ઠ	1
462	SKU	સ્કુ	3		489	SSI	ષિ	2
463	SVII	સ્વી	2		490	SSA	ષ	1
464	SMRU	સ્મૃ	3		491	SSTTI	ષ્ટિ	2
465	STU	સ્તુ	2		492	SSNN	ષ્ણ્	1
466	STTA	સ્ટ	1		493	RSSE	ર્ષે	2
467	STHA	સ્થ	1		494	SSRU	ષૃ	3
468	STA	સ્ત	1		495	SSYA	ષ્ય	1
469	STHYA	સ્થ્ય	1		496	SSPA	ષ્પ	1
470	SNA	સ્ન	1		497	HMII	હ્મી	2
471	SYA	સ્ય	1		498	HMI	હ્મિ	2
472	SMA	સ્મ	1		499	HMU	હ્મુ	3
473	SVA	સ્વ	1		500	HMUU	હ્મૂ	2
474	SHA_2	સ્હ	1		501	HU	હુ	3
475	SSA_2	સ્સ	1		502	HE	હે	2
476	SPHA	સ્ફ	1		503	HI	હિ	2
477	SUU	સૂ	3		504	HII	હી	2
478	S	સ્	1		505	HA	હ	1
479	SU	સુ	3		506	HRA	હ્ર	1
480	SII	સી	2		507	HYAA	હ્યા	1
481	SE	સે	2		508	HYII	હ્યી	2
482	SA	સ	1		509	HYI	હ્યિ	2
483	SSTTRII	ષ્ટ્રી	2		510	HYUU	હ્યૂ	3
484	SSTTRA	ષ્ટ્ર	1		511	HYU	હ્યુ	3

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
512	HMA	હ્મ	1		539	II	ઈ	2
513	HUU	હૂ	3		540	A	અ	1
514	KSYA	ક્ષ્ય	1		541	E	એ	2
515	KSII	ક્ષી	2		542	U	ઉ	2
516	KSA	ક્ષ	1		543	I	ઇ	1
517	KSI	ક્ષિ	2		544	E_V	ે	2
518	LLYU	લ્યુ	3		545	O_V	ો	2
519	LLYE	લ્યે	2		546	QUESTIONMARK	?	1
520	LLYA	લ્ય	1		547	COMA	,	1
521	LLII	ળી	2		548	AII_V	ી	2
522	LLU	ળુ	3		549	U_V_2	ુ	3
523	LLE	ળે	2		550	II_V_2	ી	2
524	LLI	ળિ	1		551	I_V	િ	2
525	LLA	ળ	1		552	SINGLEQUOTATION	'	1
526	JNII	જ્ઞી	2		553	LEFTPARANTHESIS	(	1
527	JNA	જ્ઞ	1		554	RIGHTPARANTHESIS	)	1
528	RR_2	ૠ	1		555	AI_V	ૈ	2
529	ZERO	૦	1		556	GHE	ઘે	2
530	ONE	૧	1		557	NNII	ણી	2
531	THREE	૩	1		558	GHUU	ઘૂ	3
532	FOUR	૪	1		559	GHU	ઘુ	3
533	SIX	૬	1		560	NNU	ણુ	3
534	SEVEN	૭	1		561	UU_V_2	ૂ	3
535	EIGHT	૮	1		562	R_V	્ર	1
536	NINE	૯	1		563	RE_V	્રે	2
537	UU	ઊ	2		564	DOT	.	1
538	AA_V	ા	1		565	HYPHEN	-	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group no
566	MRU	મૃ	3		594	PRAA	પ્રા	1
567	MRA	મ્ર	1		595	R_V_2	ર્‌ે	2
568	RKHE	ખ્ખે	2		596	SHI_V	શિ્	2
569	FULLSTOP	.	1		597	NNA	ણ્	1
570	O	ઓ	2		598	AU_V	ૌ	2
571	AA	આ	1		599	CU	ચુ	3
572	AI	ઐ	2		600	RDA	દ્	2
573	AU	ઔ	2		601	KSMII	ક્ષ્મી	2
574	UU_V	ૂ	3		602	RAI	રૈ	2
575	U_V	ુ	3		603	TTAA	ટા	1
576	SHA	શ	1		604	NNYU	ણ્યુ	3
577	LEFTDOUBLE QUOTATION	"	1		605	JNAA	જ્ઞા	1
578	RIGHTDOUBLE QUOTATION	"	1		606	YI	યિ	2
579	GII	ગી	2		607	THI	થિ	2
580	GI	ગિ	2		608	RTHII	ર્થી	2
581	GI_HALANT	ગ્	1		609	YII	યી	2
582	EXCLAMATIONMARK	!	1		610	CCII	ચ્ચી	2
583	V	વ્	1		611	SHKA	શ્ક	1
584	BHI	ભિ	2		612	STTRE	સ્ટ્રે	2
585	TR	ત્ર્	1		613	STII	સ્તી	2
586	JJA	જ્જ	1		614	TTYA	ટ્ય	1
587	NNYA	ણ્ય	1		615	SRU_2	શ્રુ	3
588	RCII	ર્ચી	2		616	LU	લુ	3
589	KH	ખ્	1		594	PRAA	પ્રા	1
590	HYAA	હ્યા	1		595	R_V_2	ર્‌ે	2
591	RBHA	ર્ભ	2		596	SHI_V	શિ્	2
592	BHYU	ભ્યુ	3		597	NNA	ણ્	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
598	AU_V	ૌ	2		625	TMA	ત્મ	1
599	CU	ચુ	3		626	TTVA_2	ત્વ	1
600	RDA	દ્	2		627	PHI	ફિ	2
601	KSMII	ક્ષ્મી	2		628	SBA	સ્બ	1
602	RAI	રૈ	2		629	VHA	વ્હ	1
603	TTAA	ટા	1		630	SRU	સૃ	3
604	NNYU	ણ્યુ	3		631	N	ન્	1
605	JNAA	જ્ઞા	1		632	T	ત્	1
606	YI	યિ	2		633	RPU	ર્પુ	4
607	THI	થિ	2		634	JVA	જ્વ	1
608	RTHII	ર્થી	2		635	KSU	ક્ષુ	3
609	YII	ચી	2		636	KSNNA	ક્ષ્ણ	1
610	CCII	ચ્ચી	2		637	DHRYA	ધ્ર્ય	2
611	SHKA	શ્ક	1		638	SCI	શ્ચિ	2
612	STTRE	સ્ટ્રે	2		639	DDE_2	દ્દે	2
613	STII	સ્તી	2		640	RII_V	ર્ી	2
614	TTYA	ટ્ય	1		641	TTU_2	ત્તુ	3
615	SRU_2	શ્રુ	3		642	PHRA	ફ્ર	1
616	LU	લુ	3		643	BYA	બ્ય	1
617	THYA	થ્ય	1		644	HAI	હૈ	2
618	NN	ણ્	1		645	PHII	ફી	2
619	RI	રિ	2		646	JHUU	ઝૂ	3
620	C	ચ્	1		647	STE	સ્તે	2
621	BI	બિ	2		648	STTE	સ્ટે	2
622	SNE	સ્ને	2		649	SKRU	સ્કૃ	3
623	SS	ષ્	1		650	BH	ભ્	1
624	SSU	ષુ	2		651	RJA	જ્	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
652	SYU	સ્યુ	3		679	DHRU_2	ધ્રુ	4
653	DHDHI	ધ્ધિ	2		680	RNSA	ર્ન્સ	2
654	STHI	સ્થિ	2		681	DDRA	ડ્ર	1
655	MHA	મ્હ	1		682	CCHE	ચ્છે	2
656	SMI	સ્મિ	2		683	RSVA	શ્વ	2
657	NDRI	ન્દ્રિ	2		684	TYE	ત્યે	2
658	KRI	ક્રિ	2		685	RHU	હ્યુ	3
659	LI	લિ	2		686	SHE	શે	2
660	SAI	સૈ	2		687	RYE	ર્યે	2
661	RJE	જ્યે	2		688	SPE	સ્પે	2
662	BBE	બ્બે	2		689	DDHI	ઢિ	2
663	BHAI	ભૈ	2		690	KTI	ક્તિ	2
664	RNNA	ર્ણ	2		691	SPA	સ્પ	1
665	NNE	ણે	2		692	RDHI	ર્ધિ	4
666	TAI	તૈ	2		693	STHUU	સ્થૂ	3
667	JRA	જ્ર	1		694	NVE	ન્વે	2
668	MYU	મ્યુ	3		695	RJAA	જ્ઞા	2
669	JJE	જ્જે	2		696	TPRE	ત્પ્રે	2
670	RTRU	ર્તૃ	4		697	PNU	પ્નુ	3
671	DDHUU	ઢૂ	3		698	THVII	થ્વી	2
672	NNI	ણિ	2		699	RMII	ર્મિ	2
673	SKRA	સ્ક	1		700	TPA	ત્પ	1
674	SKUU	સ્કૂ	3		701	NTE	ન્તે	2
675	SKA	સ્ક	1		702	CH	છ્	1
676	TTHYA	ટ્થ	1		703	SI	સિ	2
677	MA_V	◌્મ	1		704	JHU	ઝુ	3
678	JHII	ઝી	2		705	DHE	ધે	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
706	NTII	ન્તી	2		733	NNE_2	ન્ને	2
707	STTII	સ્ટી	2		734	KSMA	ક્ષ્મ	1
708	DDHYU	ઢ્યુ	3		735	KSE	ક્ષે	2
709	THUU	થૂ	3		736	BHYE	ભ્યે	2
710	TVE	ત્વે	2		737	STTUU	સ્ટ્ર	3
711	SSII	ષી	2		738	RTHI	ર્થિ	2
712	BRE	બ્રે	2		739	RSA	ર્સ	2
713	RDDA	ર્ડ	2		740	RBHI	ર્ભિ	2
714	JHI	ઝિ	2		741	PHYU	ફ્યુ	3
715	SSN	ષ્ણ્	1		742	TKA	ત્ક	1
716	DRU	દ્ર	3		743	KVE	ક્વે	2
717	KL	ક્લ્	1		744	DDAI	ડૈ	2
718	SSMA	ષ્મ	1		745	DD	ડ્	1
719	DDYE	ડ્યે	2		746	DHVII	ધ્વી	2
720	DHVA	ધ્વ	2		747	SSE	સ્સે	2
721	JAI	જૈ	2		748	SSPI	ષ્પિ	2
722	PHRI	ફ્રિ	2		749	DHRUU	ધ્રૂ	4
723	KHRI	ખ્રિ	2		750	K_CE	કૅ	2
724	VYE	વ્યે	2		751	SPI	સ્પિ	2
725	RTII	ર્તી	2		752	HYE	હ્યે	2
726	KRUU	ક્રૂ	3		753	STTRI	સ્ટ્રિ	2
727	BRU_2	બ્રુ	3		754	KTTA	ક્ટ	1
728	NYUU	ન્યૂ	3		755	STHII	સ્થી	2
729	4MBU	મ્બુ	3		756	NCA	ન્ચ	1
730	RSSI	ર્ષિ	2		757	RTTE	ર્ટ	2
731	TTUU	ટ્ટ	3		758	CCHII	ચ્છી	2
732	BRI	બ્રિ	2		759	KHKHII	ખ્ખી	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
760	NSYU	ન્યુ	3		787	SLII	સ્લી	2
761	PSA	પ્સ	1		788	SSPHA	ષ્ફ	1
762	NDDII	ન્ડી	2		789	RCE	ર્ચ	2
763	TTYUU	ટ્યૂ	3		790	RMI	ર્મિ	2
764	NSTTII	ન્સ્ટી	2		791	RNA	ર્ન	2
765	NTTE	ન્ટે	2		792	RTTU	ર્ટુ	4
766	MPE	મ્પે	2		793	RTTII	ર્ટી	2
767	MAI	મૈ	2		794	THTHU	થ્થુ	3
768	RCI	ર્ચિ	2		795	NDRA	ન્દ્ર	1
769	NAI	નૈ	2		796	PHO	ફો	2
770	PHRE	ફ્રે	2		797	SHRUU	શ્રૂ	3
771	DDHE_2	દ્ધે	2		798	CHUM	છું	4
772	RKE	ર્કે	2		799	DHVE	ધ્વે	2
773	BLA	બ્લ	1		800	RGA	ર્ગ	2
774	DHYU	ધ્યુ	4		801	SHAI	શૈ	2
775	NSUU	ન્સૂ	3		802	RDII	ર્દી	2
776	VSA	વ્સ	1		803	TTRI	ટ્રિ	2
777	NG	ન્ગ્	1		804	STTRA	સ્ટ્ર	1
778	SCA_2	સ્ચ	1		805	TTRYA	ત્ર્ય	1
779	MPA	મ્પ	1		806	NDRE	ન્દ્રે	2
780	NKRII	ન્કી	2		807	SBII	સ્બી	2
781	NTTRE	ન્ટ્રે	2		808	RVI	ર્વિ	2
782	CAI	ચૈ	2		809	RBU	બ્રુ	4
783	NSU	ન્સુ	3		810	LEFTSQUAREBRACKET	[	1
784	SL	સ્લ્	1		811	RIGHTSQUAREBRACKET	]	1
785	DRE	દ્રે	2		812	K	ક્	1
786	KSA_2	ક્સ	1		813	JHA	જ્હ	1

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
814	NTU	ન્તુ	3		841	HII_V	◌ૃહી	2
815	DHRA	ધ્ર	2		842	THVA	થ્વ	1
816	PHPHA	ફ્ફ	1		843	NMYU	ન્મ્યુ	3
817	TT	ટ્	1		844	NMII	ન્મી	2
818	E_V	◌ે	2		845	NHA	ન્હ	1
819	SSTTII	ષ્ટી	2		846	THRU	થ્રુ	3
820	H	હ્	1		847	RDHVA	ર્ધ્વ	2
821	MSA	મ્સ	1		848	HIIM	હીં	2
822	M	મ્	1		849	MRYU	મ્ર્યુ	4
823	DHRU	ધૃ	4		850	BHEM	ભેં	2
824	DDHU	ઢુ	3		851	NNE_2	ઞે	2
825	SRE	શ્રે	2		852	MMII	મ્મી	2
826	RPI	પ્રિ	2		853	RTE	ર્તે	2
827	LII	લી	2		854	DDHDDHA	ઢ્ઢ	1
828	SSTTHI	ષ્ઠિ	2		855	TTHA_2	ત્થ	1
829	NUU	ન્	3		856	KHAM	ખં	2
830	GHNA	ઘ્ન	1		857	SSTTHII	ષ્ઠી	2
831	JAU_2	જ◌ૌ	3		858	TTHUM	ઠું	4
832	NDU	ન્દુ	3		859	VYUU	વ્યૂ	3
833	SSE	ષ્વે	2		860	SVE	શ્વે	2
834	VUU	વ્	3		861	RUM	રું	4
835	DRUU	ડ્રૂ	3		862	KHKSA	ખ્ક્ષ	1
836	DDA_2	ઌ	1		863	NNII_2	ઞી	2
837	RU_V	ર◌ુ	3		864	RJU	જ્ર્યુ	4
838	_CE	◌ૅ	2		865	DVE	દ્વે	2
839	LL	ળ્	1		866	B_CE	બૅ	2
840	NNU_2	ઞુ	3		867	MBE	મ્બે	2

Class No.	Class Name	Unicode	Group no		Class No.	Class Name	Unicode	Group
868	SRRU	શૃ	3		881	YAA	ચા	1
869	NDI	ન્દિ	2		882	LLAA	ળા	1
870	TRYE	ત્ર્યે	2		883	TTAAA	ટા	1
871	TAA	તા	1		884	T	ત્	1
872	MAA	મા	1		885	KSNNA	ક્ષ્ણ	1
873	BHAA	ભા	1		886	STTI	સ્ટિ	2
874	KHAA	ખા	1		890	RII_V	◌્રી	2
875	VAA	વા	1		892	H	હ્	1
876	BAA	બા	1		893	A_Halant	અ્	1
877	DHAA	ધા	2		894	DA	દ	1
878	NAA	ના	1		895	DH	ધ્	2
879	HAA	હા	1		896	PA	પ્	1
880	CAA	ચા	1					
881	YAA	ચા	1					

Appendix II

Gujarati Unicode to Braille ASCII mapping

The following list shows the Gujarati Unicode mapping from OCRed output to respective Braille ASCII.

Gujarati Unicode	ASCII HEX	ASCII CHAR	Braille cell	Braille Meaning
ં	3B	;	⠒	Anusvara
ઃ	2C	,	⠄	Visarga
અ	41	A	⠁	A
આ ા	3E	>	⠜	AA
ઇ િ	49	I	⠌	I
ઈ ી	39	9	⠔	II
ઉ ુ	55	U	⠥	U
ઊ ૂ	5C	\	⠳	Ou
ઋ ૠ	22 52	“R	⠲⠗	Ru
એ ે	45	E	⠑	E
ઐ ૈ	2F	/	⠌	Ai
ઓ ો	58	X	⠭	O
ઓ ો	4F	O	⠕	O
ઔ ૌ	5B	[	⠪	Ow
ક	4B	K	⠅	Ka
ખ	2E	.	⠲	Kha
ગ	47	G	⠛	Ga
ઘ	3C	<	⠣	Gha
ઙ	2B	+	⠖	Nga
ચ	43	C	⠉	cha
છ	2A	*	⠡	Chha
જ	4A	J	⠚	Ja
ઝ	30	0	⠴	Za

ઞ	33	3	⠒	Nya
ટ	29	29	⠳	Ta
ઠ	57	W	⠺	Tha
ડ	24	$	⠹	Da
ઢ	3D	=	⠿	Dha
ણ	23	#	⠻	Na
ત	54	T	⠾	Ta
થ	3F	?	⠽	Tha
દ	44	D	⠸	Da
ધ	21	!	⠮	Dha
ન	4E	N	⠼	Na
પ	50	P	⠯	Pa
ફ	36	6	⠖	Fa
બ	42	B	⠨	Ba
ભ	5E	^	⠘	Bha
મ	4D	M	⠬	Ma
ય	59	Y	⠵	Ya
ર	52	R	⠗	Ra
લ	4C	L	⠇	La
વ	56	V	⠧	Va
શ	25	%	⠩	Sha
ષ	26	&	⠦	Sha
સ	53	S	⠎	Sa
હ	48	H	⠓	Ha
◌્	40	@	⠁	Halant
૧	31	1	⠂	One
૨	32	2	⠆	Two
૩	33	3	⠒	Three
૪	34	4	⠲	four
૫	35	5	⠢	Five
૬	36	6	⠖	Six

૭	37	7	::	Seven
૮	38	8	:.	Eight
૯	39	9	.·	Nine
૦	30	0	.:	Zero
ऽ	31	2	:	Gujarati sign avagraha
અઁ	27	‘	`	am
।	34	4	·:	

References

[1] Otsu, N., "A Threshold Selection Method from Gray-level Histogram," IEEE Trans. SMC, 9, 62–66 (1979).

[2] Niblack, W., [An Introduction to Digital Image Processing], Englewood Cliffs, N.J. Prentice Hall, 115–116 (1986).

[3] Sauvola, J.J. and Piet¨aikinen, M., "Adaptive document image binarization," Pattern Recognition, vol. 33,no. 2, 225–236, (2000).

[4] L. O. Germon. Binarization and Multithresholding of Document Images Using connectivity.computer vision,Graphics and Image Processing,56 (6):494-506,1994.

[5] U. Pal and B. B. Chaudhuri. Skew anlge detection of digitized Indian Script Document. *IEEE transaction on pattern Analysis and Machine Intelligence.*

[6] K.Mahata and A. G. Ramakrishnan. Precision skew detection through principle axis. In *International Conference on Multimedia Processing and Systems,2000.*

[7] K, Mahata.Optical Character Recognition for printed Tamil script, Mastersthesis, Indian Institute of Science,Bangalore,2000 .

[8] Image Quality Assessment: From Error Visibility to Structural Similarity,Zhou Wang, *IEEE transactions on image processing, vol. 13, no. 4, april 2004*

[9] I.K. Sethi and B. Chatterjee. Machine recognition of hand printed Devanagari numerals. Journal of the IETE (India), 22(8):532-535, 1976.

[10] S.H. Srinivasan and K.R. Ramakrishnan. The independent components of characters are 'strokes'. In proceedings of the International Conference on Document Analysis and Recognition, pages 414-417, 1999.

[11] P. Comon. Independent Component Analysis, A new concept. Signal Processing, 36:287-314, 1994.

[12] A. Hyvarinen and E. Ja. A fast fixed point algorithm for independent component analysis. Neural Computation, 9:1483-1492, 1997

[13] A.K. Dutta and S. Chaudhuri. Bengali alpha-numeric character recognition using curvature features. Pattern Recognition, 26:1757-1770, 1993.

[14] A. Negi, B. Phanikumar, and B.K. Trinadh. Offline printed Kanada script recognition system. In International Workshop on Performance Evaluation Issues in Multi-lingual OCR, Sept. 1999.

[15] R.S.R. Kunte and R.D.S. Samuel. On-line recognition of handwritten Kanada characters employing wavelet features. In Conf. on Communication, Control and Signal Processing, pages 393-396, 2000.

[16] R.C. Gonzalez and R. E. Woods. Digital Image Processing. Addision Wesley Publishing Co., new York, 1993.

[17] A. K. Jain, Fundamentals of Digital Image Processing. Prentice Hall of India Pvt. Ltd. 1995.

[18] K. Mahata and A. G. Ramakrishnan. Precise skew detection through principal axis. In International conf. on Multimedia Processing and Systems, 2000.

[19]