IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 3 Issue 4 April, 2015

A New Approach for Zone Identification on Printed Gujarati Text: Vertical Bar Method

Author: Shweta Agravatt[1]; Mukesh Goswami[2]; Hardik Agravatt[3]

Affiliation: ITM Universe, Vadodara, India[1]; Dharmsinh Desai University-India, Nadiad, India[2]; Sardar Vallabhbhai Institute of Tehnology, Vasad, India[3]

E-mail: Shweta.agravat@gmail.com[1]; mgoswami.it@ddu.ac.in[2]; hmavt11@gmail.com[3]

ABSTRACT

Gujarati language belongs to Indo-Aryan family of languages, widely spoken in western Indian state - Gujarat. Gujarati is a multilevel script comprises of 3 different zones- Upper Zone, Middle Zone and Lower Zone. Several characters in Middle Zone have modifiers attached with it at upper or lower region. To discriminate between character and modifier, Zone boundaries between Upper, Middle and Lower Zones are needed to be identified. Some Devnagari languages like, Hindi and Bangla have existing OCR systems. There is existence of Shirorekha in Devanagari languages. Shirorekha can be considered as a reference for Zone Segmentation, thus making Zone boundary Identification and separation easy. But in absence of Shirorekha, Zone boundary Identification is a challenging task, and the methods used in other Devnagari script OCR systems will not be useful for Gujarati Script. Hence Zone Boundary Identification becomes a difficult task for Gujarati script. This paper proposes a new Zone Boundary Identification Approach based on the Vertical Bar present in Gujarati script. This method is tested on 250 Machine Printed lines and 200 Laser Printed Lines, and accuracy achieved is 80% and 93% respectively. It is expected that this method will reduce the complexity of the Zone Identification methods and provide efficient results.

Keywords: Zone Identification, Upper Zone, Lower Zone, Middle Zone, Vertical Bar.

1. INTRODUCTION

Optical Character Recognition (OCR) plays a major role in digitization of literature and documents for Indian Scripts. Gujarati is also one of the 22 official languages recognized by the Government of India. Apart from Gujarat, this language is also spoken in adjacent union territories of Daman and Diu and Dadra and Nagar Haveli. Gujarati language stands at the 26th position among the most spoken native language in the world and nearly 50 million people speak Gujarati throughout the world [1].

Some promising work is going on in the field of OCR for Indic Scripts. In [2] Bansal V. and Sinha R.M.K. have presented complete OCR system for printed Hindi text. Negi A., Bhagvati C. and Krishna B. have provided various techniques for the South Indian Telugu Script in [3]. Chaudhuri B. B. and Pal U. in [4] have proposed complete Bangla OCR system. In [5] Dholakia J., Negi A. and Rama Mohan S. have introduced a new approach of Imaginary Line Identification for Zone Detection for Gujarati Printed Script. Patel C., Desai A. in [6] have focused on Zone Boundary Identification for handwritten Gujarati Script. From the analysis of literature available, it is evident that in comparison with other Devnagari scripts, less amount of work has been proposed for Gujarati Language. Gujarati due to its particular characteristics is needed to be treated differently from other Devnagari Scripts.

2. CHARACTERISTICS OF GUJARATI SCRIPT

Gujarati script is derived from ancient Devnagari script with some degree of modifications. Gujarati is a phonetic language where, characters are written and spoken in a same way. Gujarati Script comprises of 34 consonants and 12 vowels as indicated in Figure 1.

Out of 34 Gujarati consonants, 22 consonants are having vertical bar at the right end of the character. As Gujarati script also includes existence of Joint Characters, the consonant with vertical bar can be divided into half and

joined with another full character forms a Joint character.

In Gujarati Script, a word is formed by combining multiple basic consonants together and basic consonants may have one or more modifier(s)/vowel(s) attached with it. Multiple words can form a single line and multiple lines may create entire text.

Consonants
k K g G c C j z T q D Q N t Y d F n p f b B m y r l v s S P h L x X
Vowels
a ai e E u U a[a] ai[ai] a:
Conjunct Characters
OT J At H V M W H Rv AT
Vowel Modifiers
I Z [ ] < ^ & * ( )
Numerals
0 1 2 3 4 5 6 7 8 9

Fig 1: Consonants & Vowels of Gujarati Script

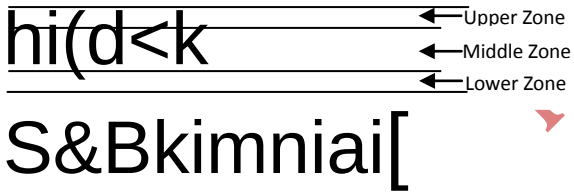

Fig 2. Zones in Gujarati Script

Fig 3. Presence of Shirorekha in Hindi Script

Gujarati is a multi-level script. A line can be divided into 3 different horizontal parallel regions as shown in Figure 2.

1. *Upper Zone*: The top part of a text line where dependent modifier(s) can be written.
2. *Middle Zone*: The region below the upper zone and above lower zone, where basic consonants and independent vowels are written.
3. *Lower Zone*: The region below the Middle layer, where dependent modifier(s) are written.

Unlike other Devnagari Scripts like Bangla and Hindi, Gujarati does not have intra word characters connected with a horizontal line called Shirorekha as shown in Figure3. Shirorekha can directly be used as a reference line to identify the separation line between Upper and Middle zone of the word. So, the presence of shirorekha makes segmentation of line and Zone Boundary Identification easy. But this is not the case with Gujarati Script. In absence of shirorekha, zone boundary identification becomes a tough task to accomplish.

3. ZONE BOUNDARY INDENTIFICATION

The accuracy of an OCR system depends upon the recognition of each individual glyph/consonant. Mainly there are two approaches for recognition of the script,

1. Identify entire word from a script as a cluster of consonants and vowels, i.e. a word is identified as a separate entity.
2. Segment the word and then separate dependent modifier(s)/vowel(s) form the consonant and then recognize them separately, i.e. each consonant/modifier is identified as a separate entity as indicated in Fig: 4.

If we are following the first approach, then we have to identify all words possible in Gujarati script. By combining basic consonants, modifiers and joint characters, approximately 2.5 million words are possible [6]. So, this approach is very tedious and not prominent for recognition part.

For second approach, we have to segment lines from a document firstly, and then segment all the words from the lines. And then segment consonants and modifiers from the word in next stage. Each separated symbol is then recognized. This approach is widely used by the researchers for Indian scripts like Hindi, Bangla etc. as described in [2], [4]. Various methods for character segmentation are described in [8].

From above explanation and previous work by Bansal and Sinha for Devnagari script [2], Pal and Chauduri for Bangla script [4], it is concluded that the second approach is more prominent then the first approach. The approaches discussed in [5] and [6] perform upper zone identification by using special characteristic of Devnagari script-Shirorekha. By computing simple Horizontal Project

Profiling (HPP), we can get prominent pick of Shirorekha, which makes separation of upper zone very easy.

But in absence of Shirorekha, Zone separation becomes quite difficult for Gujarati script. If Horizontal Projection Profile is used to identify the upper zone, then it may happen that in certain cases like - skewed document, presence of misalignment in words, less number of modifiers present, most of the modifiers are connected with the consonant, HPP method may not give required accuracy. So, for Gujarati text, a robust algorithm is needed for Zone Boundary Separation.

4. PROPOSED APPROACH

An important characteristic of Gujarati script is presence of curvature/arcs in a consonant. Many of the characters are formed by combining curvatures and vertical bar. Some of the characters like 'g, N, S' have two types of shapes in formation of the character, one is the curvature part and the other is vertical bar, not touching to the curvature part as indicated in Figure 4. A modifier/vowel 'i'- (a+i = ai) is also represented as a vertical bar. All these vertical bars represent the height of the Middle Zone. So if, the height of a vertical bar is obtained, then we can separate Upper and Lower Zones from Middle zone. Our new approach for Zone identification is based on this idea. Maximum height of a basic consonant represents the height of the middle zone. The vertical bar present in consonants represents the maximum height for all the consonants. So, by identifying the vertical bar present in the form of a part of a consonant or a modifier, we can obtain the height of middle zone. Once the middle zone is identified, the upper and lower part of this region is separated, i.e. Upper and Lower zones are separated from the middle zone.

A survey to find the minimum occurrence of vertical bar present in a line has been done before proceeding for this new approach based on Vertical Bar. There are 100 Gujarati machine printed documents analyzed and it is concluded that because of the characteristics of Gujarati Script, every single text line of a document contains at least one vertical bar in form of a modifier 'i' or separated vertical bar present in characters like 'g, N, S'. Figure 5 shows an example of Vertical bar identified in a text document.

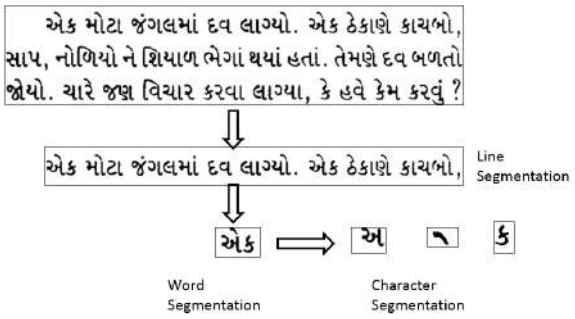

Fig 4. Segmentation of Line, Word & Character

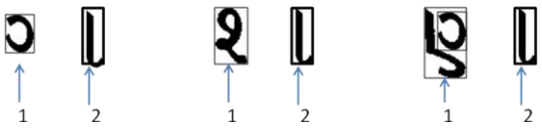

Fig 5. Separate Vertical Bar present in Gujarati Consonants. 1 represents the curvature part of the character, 2 represents the untouched vertical bar

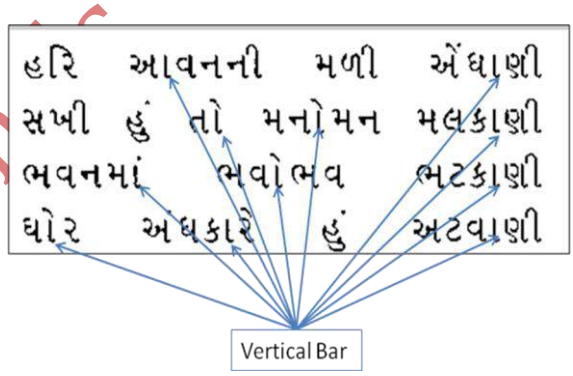

Fig 6. Presence of Vertical bars in a text line

So, using vertical bar as a reference for Zone Boundary Identification is reliable.

Here we present our new simple and less complex Algorithm for Zone Boundary Identification: Vertical Bar Method for printed Gujarati Text. The pre-processing steps are needed to be performed before applying this algorithm. Pre-processing stage includes- Binarization, de-skewing on and line segmentation.

Input: Image of text line of Gujarati script

Output: Separate Images of line with Upper, Middle and Lower zone.

Step 1: Perform labeling operation (using bwlabel() inbuilt function in Matlab) on text line to identify the Connected Components (CC) separately.

Step 2: Compute total number of black pixels in each CC.

Then derive the height and width of each CC.

Step 3: Compute Aspect ratio for each CC.

Aspect ratio = (height/width) of CC

Step 4: Identify the CC that gives highest Aspect ratio.

The vertical bar is identified based on a fact that, the height of Vertical bar is nearly three times of its width. So, the Aspect ratio of Vertical bar will be highest in comparison with all other consonants, modifiers and symbols.

Step 5: Now, based on the height of the Vertical bar, separate Upper and Lower zones and store Upper, Middle and Lower zones as separate image.

After following this algorithm, three zones will be separated and we will get 3 images of lines- first one with Upper zone, second one with Middle zone and the third one with Lower zone. The results are shown in Figure 6. Figure 6(a) shows an image of a text line, that is the input for Zone Identification Algorithm. Figure 6(b) displays the output of Zone Identification, where three zones are separated and stored as a separate image.

5. RESULTS & DISCUSSIONS

Proposed Zone Boundary Identification Algorithm has been tested on two different types of documents- Machine printed documents and Laser Printed Documents. The implementation of algorithm is done in Matlab (2012) platform.

For machine Printed documents, we have considered 250 text line from 10 documents. These documents are scanned at 300 dpi scale. We have achieved 80% accuracy for Machine Printed Document. Table I indicates the results obtained on Machine Printed Documents. Here, partially correct results indicate that because of presence of some characters like – 'U, f, h, N, J' etc., that are quite large in size and causing improper Zone separation. Some other factors affecting the accuracy of Zone separation are like- Skew, low quality printed documents and broken characters.

For Laser Printed Documents we have taken 200 text lines as an input for Zone Boundary Identification. These documents are scanned at 300 dpi scale. 93% of accuracy has been achieved for Laser Printed documents. Here, the skew present in image is a major factor that is affecting the accuracy of the results.

Our new approach based on Vertical Bar works well with both Machine and Laser Printed Documents. But as Machine Printed Documents have existence of broken characters, misalignments between intra-word, the accuracy is less in comparison with Laser Printed Document.

Our Vertical Bar based Zone Identification Approach is less complex and provides good results in comparison with other methods available for Gujarati Text.

Table 1. Results for Zone Boundary Identification on Machine Printed Document

	Results		
	Correctly Detected Zones	**Partly Correct Detected Zones**	**Incorrectly Detected Zones**
No of text lines	200	13	37
Accuracy	80%	5.2%	14.8%

Table 2. Results for Zone Boundary Identification on Laser Printed Document

	Results		
	Correctly Detected Zones	**Partly Correct Detected Zones**	**Incorrectly Detected Zones**
No of text lines	186	06	08
Accuracy	93%	3%	4%

6. ACKNOWLEDGMENTS

Our special thanks to Mr. Fedrick Macwan who have contributed towards development of this approach.

7. REFERENCES

[1]. http://www.indianmirror.com/languages/gujarati-language.html

[2]. Veena Bansal and R. M. K. Sinha, A complete OCR for printed Hindi text in Devanagari script, Proceedings of Sixth International Conference on Document Analysis and Recognition, no. d, pp. 800–804, 2001.

[3]. Atul Negi, Chakravarthy Bhagvati, and B. Krishna. An OCR System for Telugu. In Proc. 4th ICDAR, pp. 1110-1114, 1997

[4]. B. B. Chaudhuri and U. Pal, A Complete Printed

Bangla OCR system, vol. 31, no. 5, 1998.

[5]. Jignesh Dholakia, Atul Negi and S. Rama Mohan, Zone Identification in the Printed Gujarati Text, Proc. of 8th ICDAR, Vol.1, pp. 272-276, 2005.

[6]. Chhaya Patel and Apurva Desai, Zone Identification for Gujarati Handwritten Word, Second International Conference on Emerging Applications of Information Technology, pp. 194-197, 2011.

[7]. U. Pal, B.B. Chaudhuri: "Automatic Machine-Printed and Hand-Written Text Lines", Proc. 6th ICDAR, pp. 645-648, 1999.

[8]. Casey, R.G.; Lecolinet, E.; "A survey of methods and strategies in character segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 18,Issue 7, pp. 690–706, ISSN: 0162- 8828 , 1996.

