International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 1, January - 2014

Segmentation Problems in Handwritten Gujarati Text

Shailesh Chaudhari
M.Sc.(I.T.) Programme, Veer Narmad South Gujarat University, Surat, Gujarat

Dr. Ravi Gulati
Dept. of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat

Abstract

Segmentation plays a very crucial role, for any handwritten Optical Character Recognition (OCR) system. The handwritten text is separated into lines, lines into words and words into characters. Incorrect segmentation of line, word, or character decreases the recognition accuracy. Segmentation of handwritten script in general and Gujarati script in particular is a difficult task due to the curvature shapes of characters and varying writing style of different writers. Furthermore, the frequent appearance of vowel modifiers makes the text segmentation a challenging task. A good segmentation technique can improve the recognition rate. This paper deals with the problems that occur in segmentation of handwritten Gujarati text. This paper also explains the main reasons for some of these problems.

1. Introduction

Prior to invention of computer, important documents were created mainly by way of writing on a piece of paper either by handwritten or by typewriter. As a result massive volume of paper documents were generated. Further, someone has to preserve such documents for long time usage. It is necessary to preserve those documents by converting them into some other form such as in digital form. By scanning one can convert documents into digital form. The method that is used to convert scanned document into identifiable and editable form is known as Optical Character Recognition (OCR). The field of OCR has been widely researched since last 60 years, and due to its vast application environment, it continues to be an interesting area for active research. Very little work is found in the literature for recognition of handwritten Indian language scripts.

Gujarati is the official regional language of Gujarat state in India. It is a language from the Indo-Aryan family of languages, used by about 50 million people in the western part of India. Gujarati character is cursive in nature and cursive characters are normally composed of curvilinear strokes and connected successive strokes, relaxes the input constructs and permits greater variability in stroke, order and stroke numbers. Different writing styles, different sizes of characters and different shapes of characters in texts written by different people makes the job of segmentation very challenging. The technique used to segment the printed characters cannot be applied to handwritten documents due to variation in text written by varying people. The problems in segmentation depend upon the text written by a writer. A good or clearly written text has fewer problems in segmentation as compared to badly written text.

2. Related Work

A comprehensive survey of OCR is given in [1]. To the best of our knowledge, no commercial OCR for handwritten Gujarati text is available till today. The earlier work on Gujarati OCR for printed Gujarati text is presented in [2-3]. The papers dealing with handwritten Gujarati text segmentation are referenced in [4-5]. Many algorithms have been developed for segmenting of touching characters in Indian scripts, but most of them are for printed text. Line segmentation in handwritten documents is referenced in [6-8]. The papers dealing with segmentation of overlapping lines is referenced in [9].

Jindal et al. [10] have segmented the touching characters in middle zone and upper zone of printed Gurmukhi script using structural properties of the script. Chaudhuri et al. [11] have used the principle of water overflow from a reservoir to segment touching characters in Oriya script. The work on line segmentation, consonant segmentation, upper modifier segmentation and lower modifier segmentation and half character segmentation in Handwritten Hindi text are explained in [12, 13, 14]. The main objective of this paper is to find different character segmentation

problems which may occur during handwritten Gujarati script.

3. CHARACTERISTICS OF GUJARATI LANGUAGE

The basic direction of writing Gujarati is from left to right and top to bottom. Gujarati alphabets utilize 94 symbols altogether, which can be categorized into the different groupings. Gujarati character set provides 34 (+2 compound *ksha, gna*) consonants, 14 vowels which are represented by a single symbol, and 10 numerals as shown in Figure 1(a, b, c, d).

ક	ka	[kə]	ખ	kha	[kʰə]	ગ	ga	[gə]	ઘ	gha	[gʱə]	ઙ	ṅa	[ŋə]
ચ	ca	[ʧə]	છ	cha	[ʧʰə]	જ	ja	[ʤə]	ઝ	jha	[ʤʱə]	ઞ	ña	[nə]
ટ	ṭa	[ʈə]	ઠ	ṭha	[ʈʰə]	ડ	ḍa	[ɖə]	ઢ	ḍha	[ɖʱə]	ણ	ṇa	[ɳə]
ત	ta	[tə]	થ	tha	[tʰə]	દ	da	[də]	ધ	dha	[dʱə]	ન	na	[nə]
પ	pa	[pə]	ફ	pha	[fə]	બ	ba	[bə]	ભ	bha	[bʱə]	મ	ma	[mə]
ય	ya	[jə]	ર	ra	[rə]	લ	la	[lə]	વ	va	[ʋə]			
શ	śa	[ʃə]	ષ	ṣa	[ʃə]	સ	sa	[sə]						
હ	ha	[ɦə]	ળ	ḷa	[ɭə]	ક્ષ	kṣa	[kʃə]	જ્ઞ	jña	[ʤnə]			

Figure 1a. Gujarati consonants

Figure 1b. Some conjunct consonants

Vowels

અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ		એ	ઐ		ઓ	ઔ	અં	અઃ
a	ā	i	ī	u	ū	ṛ	e	ē	ai	o	ō	au	aṃ	aḥ

Figure 1c. Gujarati vowels

Digits

૧	૨	૩	૪	૫	૬	૭	૮	૯	૦

Figure 1d. Gujarati digits

There are 3 other symbols used for representing fractions. These are called 'pa' (*One Fourth*), 'adadho' (*Half*) and 'poNo' (*Three Fourth*). Gujarati consists of a special symbol called *Maatra*, corresponding to each vowel, which are attached to consonants to modify their sound. A character is said to be simple if it is a consonant alone or with a *maatra*. A character is said to be conjunct if it is a half consonant along with other consonant. There are many possibilities for the conjunct consonants that increase difficulties in segmentation and identification of the characters. The vowels (modifiers) can be placed at the left, right, top or bottom (or both) of the consonant. Gujarati word is divided into three regions-upper region, middle region and lower region. The upper and lower region includes vowels and middle region includes consonants.

4. Segmentation Problems

There are many problems encountered in the segmentation procedure. The poorly written text can lead to decrease in segmentation rate and hence recognition rate. This can be broadly divided into two categories:

1) The problems that can be avoided.
2) The problems which cannot be avoided.

Some of the problems in the text cannot be avoided due to writer's natural way of writing the text. The problems related with writer's natural handwriting i.e. the way of writing different characters creates problems in data which are difficult to overcome. This leads to decrease in recognition rate. The problems that can be avoided occur due to bad quality of material, bad scanning and most important factor is speed of writing. If a writer uses the gel pen for writing then chances are more for touching of characters as compared to thin tip ball point pen. The bad quality of material like paper and pen creates fewer problems as compared to problems created by speed of writing the text. The major problems in same text written by a single writer in different situations occur due to his natural handwriting and speed of writing. The problems due to speed of writing the text can be avoided. Problems in handwritten text can be divided into three categories:

1) Problems in Line Segmentation
2) Problems in Word Segmentation
3) Problems in Character Segmentation

4.1 Segmentation Problems in Line

The problems in line segmentation can occur due to following reasons:

4.1.1 The lower modifier of one line overlaps with the upper modifiers of lower line. In figure 2, upper modifier of lower line overlaps with lower modifier of upper line. Due to overlapping of pixels of two lines it is not possible to segment the two lines with horizontal projection technique.

4.1.2 Zigzag lines of the text and Zigzag words of the same line. This creates curvature in the lines. Due to curvature in the lines as shown in Figure 3, it is very difficult to determine the proper base line. In such cases the segmentation of two lines is very challenging.

4.1.3 Unusual space between lines. It also creates line segmentation problems as shown in Figure 4.

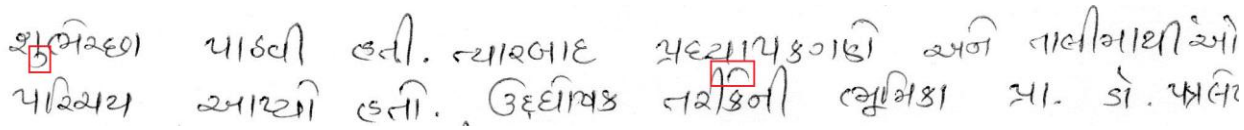

Figure 2. Modifier overlapping

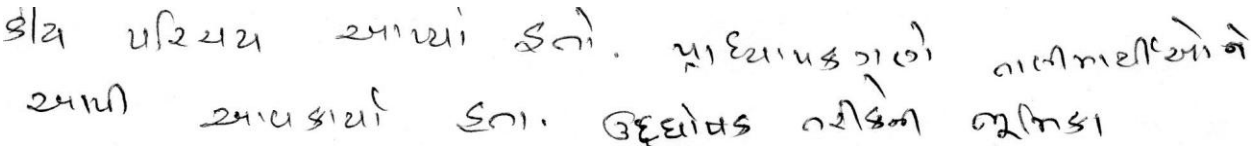

Figure 3. Zigzag line and zigzag word

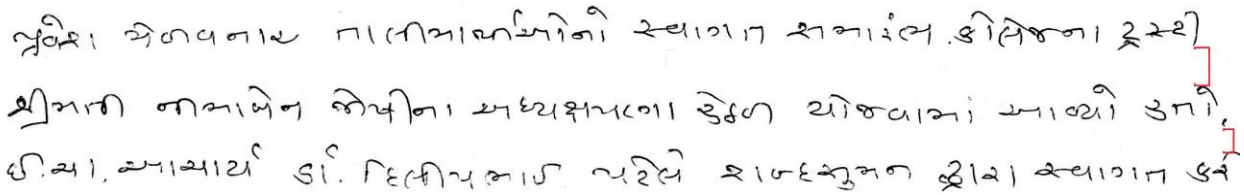

Figure 4. Unusual line spacing

4.2 Segmentation Problems in Word

Gujarati is a curvature language, unlike many other Devnagari languages, as it does not have *Shirolekha* (Headlines) over characters of a word. This makes word segmentation in Gujarati little more difficult. The problems in word segmentation are very less. Some problems occur due to improper writing style of writer. Sometimes writer does not form uniform character spacing between characters of a single word and unusual spacing between words in the same line as shown in Figure 5. So it leads to over segmentation of words.

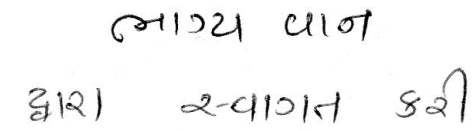

Figure 5. Unusual spacing in inter-word and intra-word

4.3 Segmentation Problems in Character

Maximum number of problems occurs in character segmentation. The problems in character segmentation can be further divided into following categories:

1) Problems in upper region
2) Problems in lower region
3) Problems in middle region

4.3.1 Problems in upper region. The problems in upper region can be further divided into two categories:

i) Unusual size of upper modifiers

Figure 6. Unusual size of upper modifier

Due to large size of upper modifier as shown in Figure 6, the determination of position of header line in a word is very difficult. It results in non segmentation of upper modifier from the consonant.

ii) Touching of upper modifier with another upper modifier

In some words upper modifier merges with another upper modifier as shown in Figure 7. It is very difficult to segment these types of modifiers from the word.

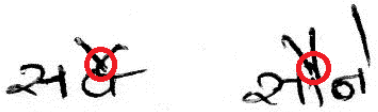

Figure 7. Upper modifier touching with each other

iii) Touching of upper modifier of previous character with next character.

In Figure 8, the modifier of previous character in upper region touches with next character in middle part. Such cases are very frequent and are very difficult to segment.

Figure 8. Upper modifier touching with next character

4.3.2 Problems in lower region. The problems in lower region can be further divided into following categories:

i) Determination of presence of lower modifier in a word

Due to variation in heights of different characters in a word it is very difficult to determine the presence of lower modifier in the word.

ii) Unusual size of lower modifiers

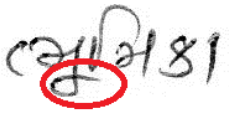

Figure 9. Unusual size of lower modifier

Due to large size of lower modifier as shown in Figure 9, the two vowel modifiers overlap.

iii) Merging of lower modifier with consonant in middle region

Figure 10. Merging of lower modifier with consonant

In Figure 10, the lower modifier merges with character '  '. Due to merging of lower modifier with the character it is very difficult to determine the presence of lower modifier in a word.

iv) Presence of lower modifier like features in some characters

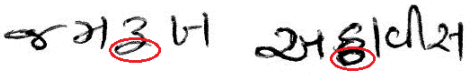

Figure 11. Lower modifier like feature

In Figure 11, the character '*ra*' and the character '*tha*' have lower modifier like features. They have loop in lower part which is similar to lower modifier.

4.3.3 Problems in middle region. The problems in middle region can be divided into following categories:

i) The problem of touching characters can be further divided into three parts:

a) Touching of modifier with consonants in middle region.

Figure 12. Modifier touching with consonant

The problem of touching the left modifier with the consonant generally occurs in many of the handwritten documents. In Figure 12, left modifier ' matra' touches with character '  ' and right modifier also touches with character '  '.

b) Touching of two or more consonants in middle region.

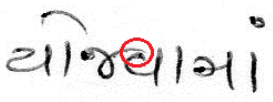

Figure 13. Consonant touching with other consonant

In Figure 13, two consonants touch each other ie. Character '  ' touches with character '  '. But it is very difficult to determine the presence of two or more touching consonants in a word.

c) Touching of half character with full character (conjuncts).

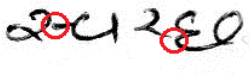

Figure 14. Conjunct character

The presence of half character touching full character makes the problem of segmentation of handwritten Gujarati text very complex. In Figure 14, half character '  ' touches the full character '  ' and half character '  ' touches the full character '  '. The above problem can be solved easily if we are able to determine the presence of conjunct in a word. The determination of presence of conjunct in a word is very challenging task

ii) Overlapping of characters in middle region

Figure 15. Overlapping character

In Figure 15, character '  ' overlaps with half character '  ' and character '  ' also overlaps with character '  '. These types of characters are difficult to segment by vertical projection. This type of problem mostly occurs with no vertical bar characters.

iii) Broken Characters

Some characters are difficult to write completely without lifting the hand at least once.

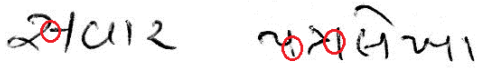

Figure 16. Broken character

In such cases sometimes space left with in a character i.e. Some pixels are missing which divides the character into two or more parts In Figure 16 (left), character '  ' has some missing pixels which breaks the character into two parts. This is very common problem in handwritten documents and it is very difficult to solve. It is an over segmentation problem. It can be solved during recognition. Broken character problem may arise due to improper writing of element

e.g. some times while writing, the pen stops working properly in between the words or words do not scanned properly. This leads to the formation of broken character Image is as shown in Figure 16 (right).

iv) Skewed Character

In this problem, as shown in Figure 17, characters in a word are not written straight but the word inclined either left-skewed or right-skewed which causes difficulty during segmentation.

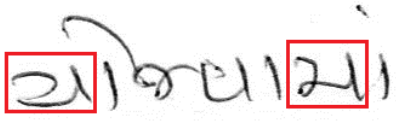

Figure 17. Skewed character

5. Concluding Remark

The difficulty of performing accurate segmentation is determined by the nature of the material to be read and by its writer. Generally, missegmentation rates for handwritten text increase progressively from machine print to cursive writing. Thus, simple techniques based on white separations between characters are adequate for machine printed texts. For handwritten text from many writers and a large vocabulary, sophisticated methods are being followed.

From the problems explained above, we conclude that complete segmentation of handwritten Gujarati text will increase the recognition rate. Some problems can be removed if writer uses the better material and write patiently. To solve the problems related with writer's natural handwriting efficient algorithms are to be designed to segment the handwritten text and we are working on it.

References:

[1] S. Mori, C.Y. Suen, and K. Yamamoto, "Historical review of OCR Research and development", In Proceedings of the IEEE, 1992, Vol. 80, No. 7, pp. 1029-1058.

[2] S. Antani, L. Agnihotri, "Gujarati Character Recognition", Fifth International Conference on Document Analysis and Recognition (ICDAR'99), 1999, pp. 418-421.

[3] [3] J. Dholakia, A. Negi, S. Ram Mohan, "Zone Identification in the Printed Gujarati Text", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR05), 2005.

[4] A. Desai, "Handwritten Gujarati Numeral Optical Character Recognition using Hybrid Feature Extraction Technique", Proceeding of International Conf. on IPVC2010, 2010.

[5] C. Patel, A. Desai, "Zone Identification for Gujarati Handwritten Word", Proceedings of the 2011 Second International Conference on Emerging Applications of Information Technology, 2005

[6] N. Tripathy, and U. Pal, "Handwriting Segmentation of unconstrained Oriya Text", In International Workshop on Frontiers in Handwriting Recognition, 2004, pp. 306–311.

[7] G. Louloudis, B. Gatos, I. Pratikakis and K. Halatsis, "A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 2006, pp.515-520.

[8] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "A new algorithm for detecting text line in handwritten documents", In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 35–40.

[9] M. K. Jindal, R. K. Sharma, and G.S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts", In International Journal of Computational Intelligence Research (IJCIR), Research India Publications, 2007, Vol. 3, No. 4, pp. 277-286.

[10] M. K. Jindal, R. K. Sharma, and G.S. Lehal, "Segmentation of Touching Characters in Upper Zone in printed Gurmukhi Script", In Proceedings of the 2nd Bangalore Annual Compute Conference, Banglore, ACM, No. 9, 2009.

[11] B.B. Chaudhuri, U. Pal, and M. Mitra," Automatic recognition of printed oriya Script", In International Conference on Document Analysis and Recognition,2009, pp. 795–799.

[12] N. Garg, L. Kaur, and M.K. Jindal, "Segmentation of Handwritten Hindi Text", In International Journal of Computer Applications (IJCA), 2010, Vol. 1, No. 4, pp. 22-26.

[13] N. Garg, L. Kaur, and M.K. Jindal, "A new method for line segmentation of Handwritten Hindi Text", In Proceedings of the IEEE 7th International Conference on Information Technology: New Generations (ITNG 2010), 2010, pp.392-397.

[14] N. Garg, L. Kaur, and M.K. Jindal, "Half character segmentation of Handwritten Hindi Text", In Proceedings of ICISIL2011, 2011, pp.48-53.