A Review on Segmentation Problems on Gujarati Handwritten Text Document

Mr. Rakesh R. Savant
Lecturer
Department of Computer Science & Technology Engineering
Babu Madhav Institute of Information Technology, India

Ms. Preeti P. Bhatt
Assistant Professor
Department of Computer Science & Technology Engineering
Babu Madhav Institute of Information Technology, India

Abstract

Optical Character Recognition (OCR) on Gujarati handwritten text document is important area of research. OCR on handwritten text document is very difficult. The reason is different people having different writing styles. For Gujarati handwritten text document OCR a very less research is done till date and there are many challenges too in Gujarati handwritten text document OCR. The whole process of OCR having many steps to be perform. In this paper the focus on the segmentation phase of Gujarati handwritten text document OCR and segmentation problems in it. In this paper we discover the difficulties of segmentation on Gujarati hand written text document. Also discover the comparison of the various segmentation methods which is works effectively for other languages and may be suit for Gujarati handwritten text document OCR too. This review paper illustrate the overview of OCR, various techniques of segmentation phase on Gujarati hand written text document with comparative study and problems in segmentation.

Keywords: OCR, Gujarati, Handwritten Text Document, Segmentation, Methods

I. INTRODUCTION

Gujarati language is popular language in Gujarat the state of India. The writing style in Gujarati is left to right side. In Gujarati language there are more curves and lower part of character or extension present. Gujarati language has sub-categories that is Surti Gujarati, Parsi Gujarati, kathiyawadi Gujarati etc. Languages are different so its writing style is also different for each language the people who speaks. And while you apply OCR on such handwritten language documents to recognize lines, word or characters the major problem is arise when you apply segmentation on that hand written document. For other Indian languages up to the good level of research work is done but for Gujarati language it is in preliminary phase. After study many research papers we conclude that they work on printed Gujarati text documents. The main problem they face is free handwriting style of people. Because of every human have different writing style, it is not an easy to identify the words or characters from such hand written text document. They also describe some methods for other languages.

II. OCR

In Optical Character Recognition (OCR), a process takes images as an input in a system and gives output as a text documents. OCR is an area of research in digital image processing. OCR takes image of handwritten or printed text document as input and after process on the image it will convert the image into text form. Basic applications of OCR is data entry for business documents, optical mask reader, electronic images of printed documents searchable, automatic number plate recognition etc.

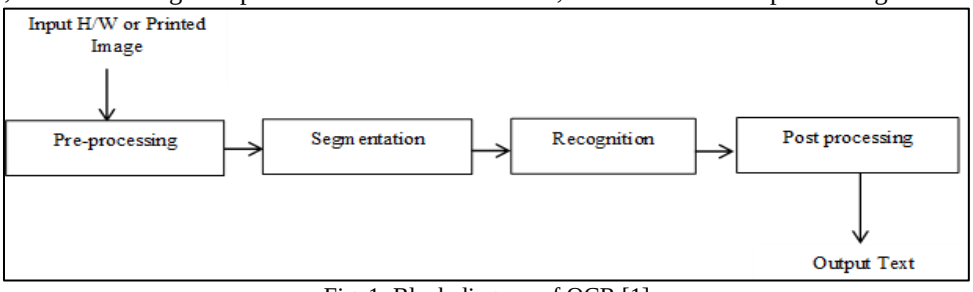

Fig. 1: Block diagram of OCR [1]

A. Importance of Segmentation Phase in OCR

The segmentation step is second step in OCR process. To recognize the line, word or character we first need to segment the image document in line, word or character. The fundamental steps in segmentation is to partition or divide text image in some regions and grab meaningful regions from that text image. The accuracy of OCR is mostly depends on how effectively the segmentation done. The more accuracy in segmentation will lead to more accurate recognition in OCR.

B. Process of Segmentation

1) *Line Segmentation*

Line segmentation apply on handwritten text documents to extract the lines.

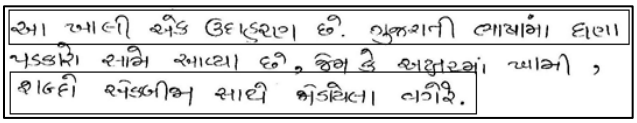

Fig. 2: Line segmentation

2) *Word Segmentation*

Word segmentation apply on handwritten text documents to extract the words.

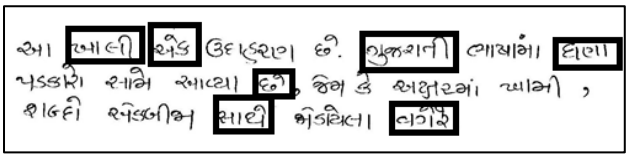

Fig. 3: Word segmentation

3) *Character Segmentation*

Character segmentation apply on handwritten text documents to extract the characters.

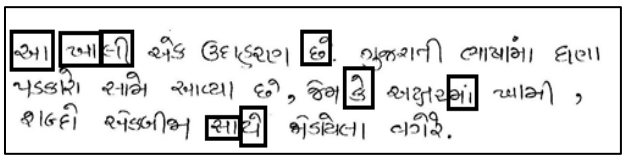

Fig. 4: Character segmentation

4) *Gujarati Script*

As every languages having its own character sets Gujarati language have the character set which consist of 35 consonants, 13 vowels, 13 dependent vowel signs, 6 signs, 10 digits and 1 sign for Indian currency. Consonants may connected with any of the vowel extensions.

C. Consonants

ક	ખ	ગ	ઘ	ઙ
ચ	છ	જ	ઝ	ઞ
ટ	ઠ	ડ	ઢ	ણ
ત	થ	દ	ધ	ન
પ	ફ	બ	ભ	મ
ય	ર	લ	વ	શ
ષ	સ	હ	ળ	ક્ષ
જ્ઞ				

Fig. 5: Constants

1) *Vowels*

અ	આ	ઇ	ઇ	ઉ	ઊ	ઋ	ૠ	એ	ઓ	ઔ	અઃ

Fig. 6: Vowels

2) *Vowel extensions*

ા	િ	ી	ુ	ૂ	ૃ	ે	ે	ૃ	ો	ો	ૌ	ૈ	ઁ	ઃ

Fig. 7: Vowel extensions

3) *Digits*

૦	૧	૨	૩	૪	૫	૬	૭	૮	૯

Fig. 8: Digits

D. Indian currency sign

₹

Fig. 9: Indian Currency Sign

III. SEGMENTATION PROBLEMS IN HANDWRITTEN GUJARATI TEXT DOCUMENT [8]

A. Line Segmentation

1) *Modifier Overlapping*

The lower modifier of one line overlaps with the upper modifiers of very next line.

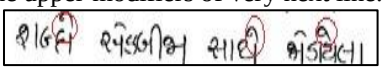

Fig. 10: Modifier Overlapping

2) *Unusual Line Spacing*

Spacing is not proper between two or more than two lines

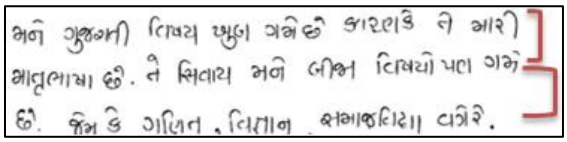

Fig. 11: Unusual Line Spacing

3) *Zigzag Line/Word/Character*

It creates curvature in the lines. Text is not in proper line.

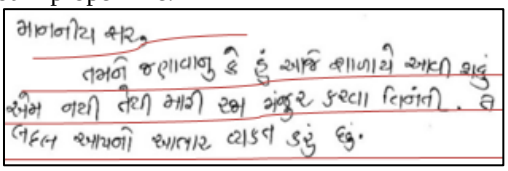

Fig. 12: Zigzag Line/Word/Character

B. Word Segmentation

1) *Unusual spacing in inter-word and intra-word*

Spacing between two words are not proper because of that spacing problem occurs.

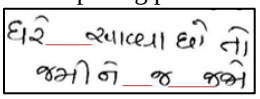

Fig. 13: Unusual spacing in inter-word and intra-word

C. Character Segmentation

1) *Lower Region Problems*

– Lower modifier problem

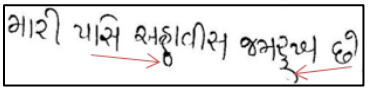

Fig. 14: Lower modifier problem

– Touching of half character with full character

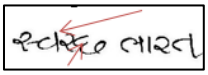

Fig. 15: Touching of half character with full character

– Unusual size of lower modifier

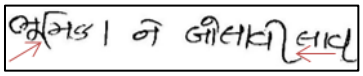

Fig. 16: Unusual size of lower modifier

– Skewed character

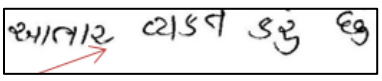

Fig. 17: Skewed character

2) *Middle region problems*

– Modifier touching with consonant

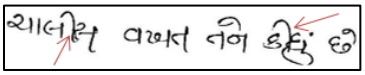

Fig. 18: Modifier touching with consonant

– Consonant touching with other consonant

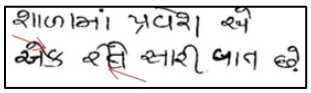

Fig. 19: Consonant touching with other consonant

– Overlapping of characters in middle region

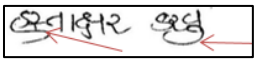

Fig. 20: Overlapping of characters in middle region

– Broken character

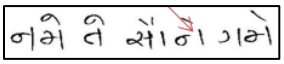

Fig. 21: Broken character

3) *Upper region problems*

– Variation in the size of upper modifier

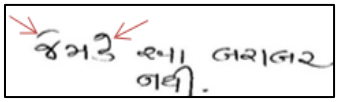

Fig. 22: Variation in the size of upper modifier

– Touching of upper modifier with another Upper modifier

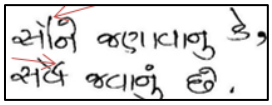

Fig. 23: Touching of upper modifier with another Upper modifier

– Merging of lower modifier with consonant

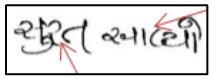

Fig. 24: Merging of lower modifier with consonant

IV. LITERATURE REVIEW

Plenty amount of research work is done in printed as well as hand written text document in languages like English, Arabic, Assamese, Oriya and Hindi. [1][3][5] Many researchers preferred projection based methods (vertical and horizontal projection) used in many languages for various types of segmentation.

The paper [8] presents many problems in handwritten text document segmentation process in Gujarati language. Major focus on segmentation like line, word and character then it increases recognition accuracy.

The paper [5] presents the segmentation process on Hindi handwritten document is described. Also applying various methods and algorithm to segment the upper and lower modifiers. The experiments and results are conclude based on the dataset consist of 15 different writer's handwritten documents.

The paper [9] presents the OCR on Tamil script. Tamil script is non-heading based script and by applying horizontal projection on document to segment the line. Horizontal projection is most commonly used technique to segment the lines from document. Also conclude that the proposed algorithm is used for segment the non-heading based scripts like Telugu, Malayalam, Kanada, Gujarati, etc.

V. COMPARATIVE SUMMARY FOR SEGMENTATION METHODS

Table – 1
Comparative Summary for Segmentation Methods

Sr. No	*Ref.*	*Methods*	*Languages*	*Useful*	*May not useful*
1	*[2]*	*Kalman filter*	*English*	*This method is able to deal with overlapping text lines and various kinds of writing or noisy or damaged character.*	-
2	*[3][4]*	*Projection based (horizontal projection)*	*Bangla, Arabic, Gurmukhi, hindi, Tamil*	*This method suitable for straight lines.*	*Segmentation of overlapped or connected lines of text.*
3	*[2][5]*	*Smearing method*	*Hindi, English*	*It is used when gap between two word are proper in printed or handwritten text document.*	*In handwritten text gap between two or more than two words are proper.*
4	*[4][5]*	*Hough transform based methods*	*Bangla, Arabic, Gurmukhi*	*The text line detection method for unconstrained handwritten documents based.*	*This method is not suitable for variable skewed lines.*
5	*[5]*	*Grouping method*	*Gujarati, Hindi*	*It is used for text line detection.*	-
6	*[6][5]*	*Graph based method*	-	*The graph based method is useful for line segmentation.*	-
7	*[7]*	*Based line dictation method*	*Hindi, Devanagari*	*It is useful for line segmentation of various skew of handwritten text*	*In this method two consecutive lines touch or overlap each other due to modifiers.*

VI. CONCLUSION

Almost negligible work has been done up till now in segmentation for hand written Gujarati text document. Many challenges are there to segmentation on handwritten Guajarati document. From above literature review we conclude that there are many problems in OCR of handwritten Guajarati document to segment lines, words and characters. Mainly problems come in character segmentation phase. There is work done in segmentation of Guajarati language but that is for printed text document. There are many challenges and problems exist for handwritten Gujarati text document. Good accuracy in segmentation will lead to increase the good recognition rate.

REFERENCES

[1] prof s k shah, (2006) , "design and implementation of optical character recognition system to recognize gujarati script using template matching" ,ie (i) journal-et , vol 86.
[2] aurelie lemaitrea, (2011) , a perceptive method for handwritten text segmentation , document recognition and retrieval xvii – electronic imaging.
[3] a. Zahour, b. Taconet, p. Mercy, and s. Ramdane, (2001), "arabic hand-written text-line extraction", proceedings of the sixth international. Conference on document analysis and recognition, icdar, pp. 281–285.
[4] n. Tripathy and u. Pal. (2004), "handwriting segmentation of unconstrained oriya text", international workshop on frontiers in handwriting recognition, pp. 306–311.
[5] naresh kumar garg alt. (2010), "a new method for line segmentation of handwritten hindi text" ,seventh international conference on information technology.
[6] Likforman-sulem and c. Faure, (1994), "extracting text lines in handwritten documents by perceptual grouping", advances in handwriting and drawing : a multidisciplinary approach, pp. 21-38.
[7] i.s.i. abuhaiba, s. Datta and m.j.j. holt, (1995), "line extraction and stroke ordering of text pages", proceedings of the third international conference on document analysis and recognition, canada, pp. 390-393.
[8] shailesh chaudhari and dr. Ravi Gulati, (2014, January), "segmentation problems in handwritten gujarati text" international journal of engineering research & technology (ijert) vol. 3 issue 1.
[9] r. Indra gandhi alt. (2010), "A technique for segmentation over overlapping line of uniform sized text on non-headline based distorted tamil scripts" int. J. Of advanced networking and applications volume: 02, issue: 02, pages: 491-495.