ISSN: 2277-9655
Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW ON SEGMENTATION TECHNIQUES OF LINES, WORDS AND CHARACTERS ON GUJARATI HANDWRITTEN DOCUMENT USING OCR

Nilam Mistry*, Sameer Vashi, Vidhi Patel, Kunal Shah, Denish Rixawapla, Foram Rakholiya, Rakesh Savant

[*] Babu Madhav Institute of Information Technology, Uka Tarsadia University
Maliba Campus, Gopal Vidhyanagar, Bardoli, Gujarat, India

DOI: 10.5281/zenodo.54779

ABSTRACT

OCR is technique to convert the handwritten or printed document into the digital format by scanning it which can be understandable by a computer. OCR is important and challenging task in many computer vision applications. Segmentation is generally the first stage in any attempt to analyse or interpret an image automatically. Segmentation is separate the document into lines, lines to words and words to characters which has been one of the major laboriousness in handwritten text recognition. The role of segmentation is a crucial in most tasks requiring image analysis. The success or failure of a task is often a direct consequence of the success or failure of segmentation. Handwritten text documents contain text in free flow manner, also writing style of users may different even sometimes same user's handwriting are different in different time. That is why segmentation is difficult in case of handwritten text document. As this paper focuses on Gujarati language, it contains more curves, overlapping character & slopes. So, it is very difficult to do segmentation on it. In this paper we have applied some of the segmentation techniques to segment the handwritten Guajarati documents & reached to some conclusion.

KEYWORDS: OCR, Connected Components, Gujarati Script, Segmentation.

INTRODUCTION

OCR stands for optical character recognition. It is the popular technique in digital image processing. In document processing, Image processing and pattern recognition, OCR is the most challenging research field. In computerization of any language, one of the vital tasks is to develop an efficient and effective OCR system for the respected language.

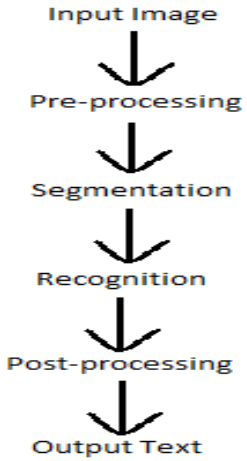

Figure 1. Block diagram of OCR

Gujarati is one of the popular language in India. There are more than 65 million speakers of Gujarati language present in India. Gujarati is the official regional language of Gujarat state in India. Gujarati is a language from the Indo-Aryan family of languages [3]. It follows left to right writing style. Gujarati script is used to write the Gujarati language. Like other Indian languages, Gujarati is also a multilevel script [3].Gujarati OCR has been currently in development phase and it is more challenging. In some kind of books, there are characters having irregular thickness. Hence, problem occurs due to Broken and merged characters in document images pose serious challenges for recognition accuracy. If these problems can be reduced by any means, then overall efficiency can be increase by up to 10-15% now.

Segmentation

Segmentation is process of partitioning a digital image into multiple regions and extracting meaningful regions known as Regions of Interest (ROI) for further image analysis. Segmentation is needed since handwritten characters frequently interfere with one another. In case of many Indian script characters may have modifiers called *Kanas & Matras*. So, identifying *kanas & matras* is crucial step in segmentation [2].

Normal Processes of Segmentation:

- **Line Segmentation**

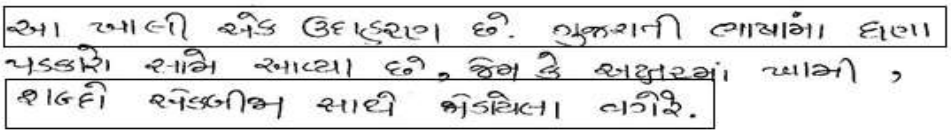

Figure 2. Line Segmentation

- **Word Segmentation**

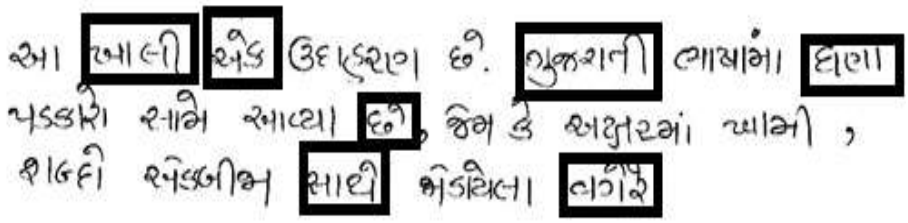

Figure 3. Word Segmentation

- **Character Segmentation**

Figure 4. Character Segmentation

Characteristics of Gujarati Script

The Gujarati Character set is made of 35 consonants, 13 vowels and 6 signs, 13 dependent vowel signs, 4 additional vowels for Sanskrit, 9 digits and 1 currency sign. The consonants can be combined with the vowels, and consonants to form conjunct consonants [3].

ISSN: 2277-9655
Impact Factor: 3.785

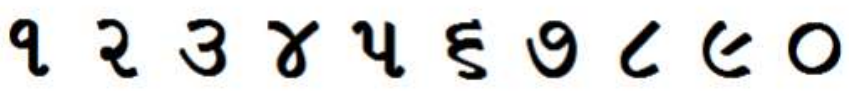

Figure 5. Digits in Gujarati

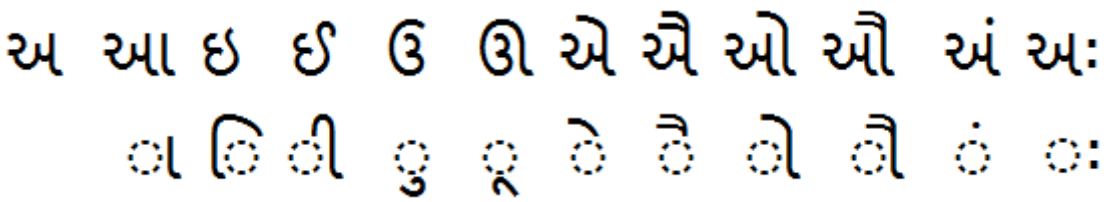

Figure 6. Special symbol of Gujarati Script

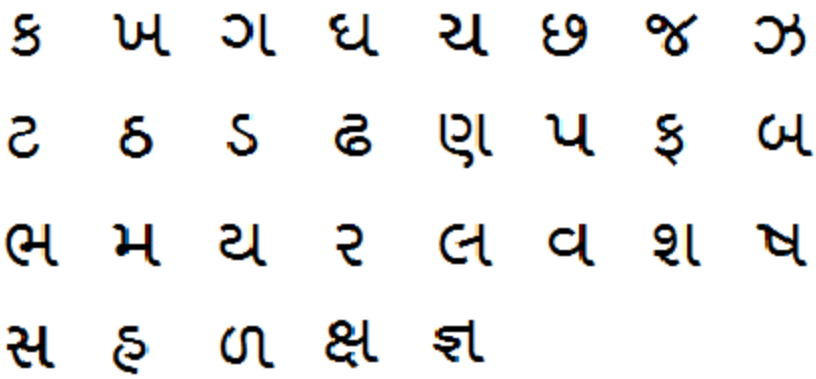

Figure 7. Alphabets in Gujarati Language

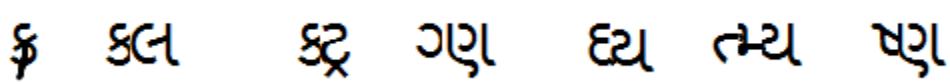

Figure 8. Symbols for consonants without the vowels sounds

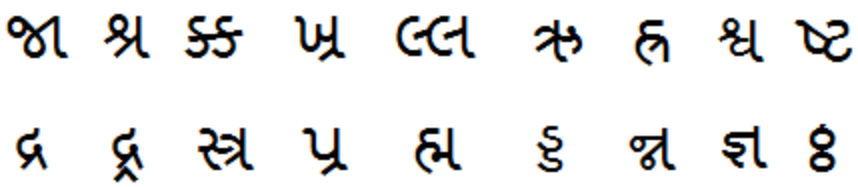

Figure 9. Some Conjunct Characters in Gujarati

As far as handwritten OCR is concerned negligible work is done. For handwritten Gujarati text document very few work is done in OCR as compare to other languages like English language because of diverse human handwriting style & customs, curves present in that document.

LITERATURE REVIEW

Sr No.	Author	Segmentation Method / Algorithm	Advantage	Disadvantage
1	[2]	Projection(Devanagari, Bangla, Gurumukhi, Gujarati)	➢ The presence of header line or shirolekha in case of scripts like Devanagari, Bangla and Gurumukhi helps to detect the header line as it generates a prominent peak in the horizontal projection profile of a word	➢ Horizontal ➢ Projection of Gujarati word is not useful to detect the upper zone boundary especially in the cases where the number of modifiers is significantly large
2	[5]	Delaunay graph(English)	➢ An advantage of Delaunay graph is independent of the order the points are processed	➢ TIN models are less widely available than raster surface models and are more time consuming to construct and process –it is a highly complex data structure.
3	[6], [12]	Smearing methods (English, Hindi)	➢ It is more efficient for printed document	➢ If gape between two words are increase not segment properly
4	[5]	Kalman Filter (English)	➢ The main advantage of the Kalman filter is its ability to provide the quality of the estimate (i.e., the variance), and its relatively low complexity.	➢ It provides accurate results only for Gaussian and linear models. For Gaussian models with limited nonlinearity, extended Kalman filter (EKF) is appropriate. For non-Gaussian and non-linear models, particle filtering (PF) is the most appropriate approach, since it is able to provide arbitrarily posterior probability distribution.
5	[5]	DMOS-P (Description & Modification of The Segmentation with Perception Vision) (English)	➢ It uses EPF (enhanced position formalism) language which enables logical description of the structure of document. ➢ This method is generic and can be applied on any kind of document.	
6	[13]	MLP (multi-layered perceptron)	➢ Neural networks are capable of generalisation, that is, they classify an unknown pattern with other known patterns that share the same distinguishing features. This means noisy or incomplete inputs will be classified because of their	➢ Occasionally, the multilayer perceptron fails to settle into the global minimum of the energy surface and instead find itself in one of the local minima. This is due to the gradient descent strategy followed. A number of alternative approaches can be taken to reduce this possibility

ISSN: 2277-9655
Impact Factor: 3.785

			similarity with pure and complete inputs.	

Table: Advantages & Disadvantages of Segmentation Techniques

IMPLEMENTATION

Steps for doing Segmentation:

Following are the steps we have applied for doing segmentation of handwritten Gujarati documents:

1. Pre-processing
2. Line Segmentation
3. Word Segmentation
4. Character Segmentation

Pre-processing

Pre-processing includes the operations like conversion of rgb image into grayscale image, binarization, smoothing, cropping and skew correction on a document image so that final classification can be made simple & more accurate. As we are using scanned document images for the OCR system which are not good candidate for the segmentation so for achieving accuracy pre-processing step is require.

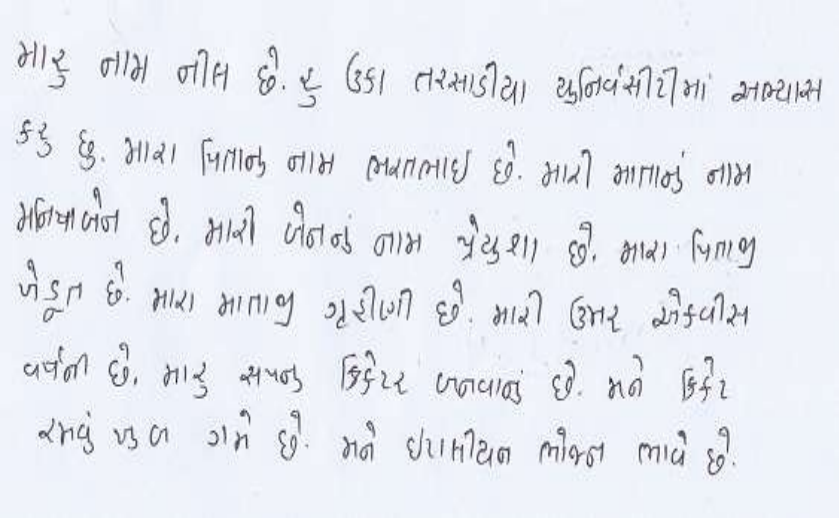

Figure 10. Original Image

conversion of rgb image into grayscale Image:

It is the process of converting an rgb image into a grayscale image (i.e. in 8 bit format) so that we have to deal with 256 number of colours (i.e. contains all the shades of grey colour) & because of that it is easy to perform the different operation on a document image which contains small number of colours.

ISSN: 2277-9655
Impact Factor: 3.785

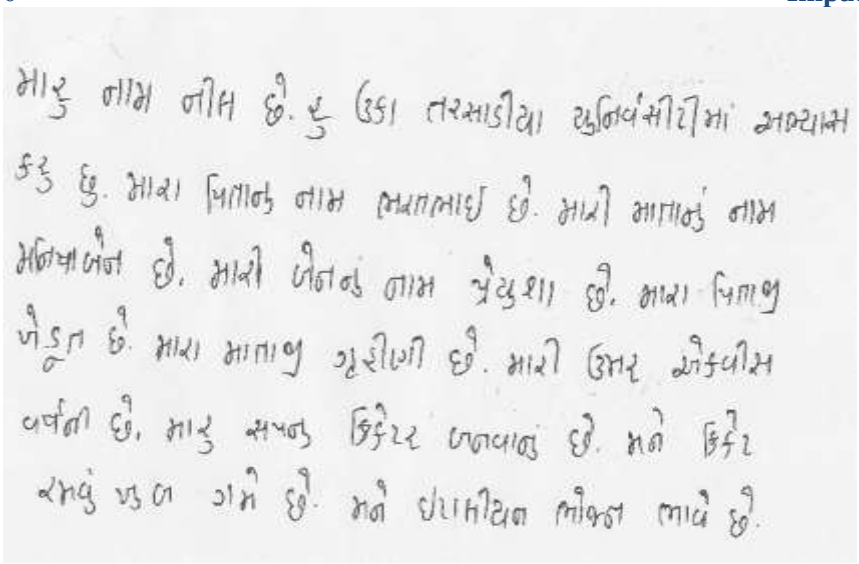

Figure 11. Grayscale Image

Binarization:
It is the process of converting a grayscale image into a binary image (i.e. in 2 bit format) so that we have to deal with only two colours (i.e. white & black) & because of that it is easy to perform different operation on a document image which contains small number of colours. We have applied first binarization on an image and then negation of it.

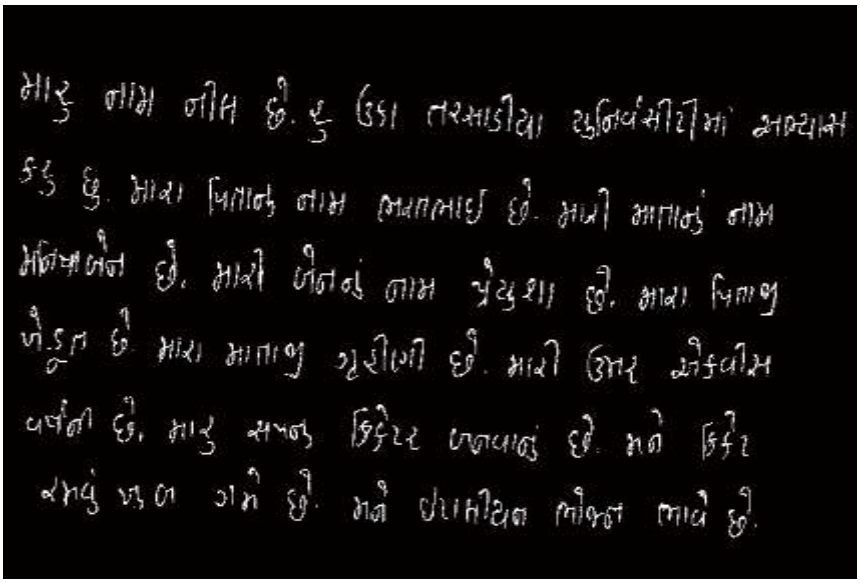

Figure 12. Image Binarization

Smoothing:
If your data are noisy, you might need to apply a smoothing algorithm to expose its features, and to provide a reasonable starting approach for parametric fitting. The smoothing process attempts to estimate the average of the distribution of each response value. The estimation is based on a specified number of neighbouring response values. For doing smoothing in document image we have used the dilation process.

ISSN: 2277-9655
Impact Factor: 3.785

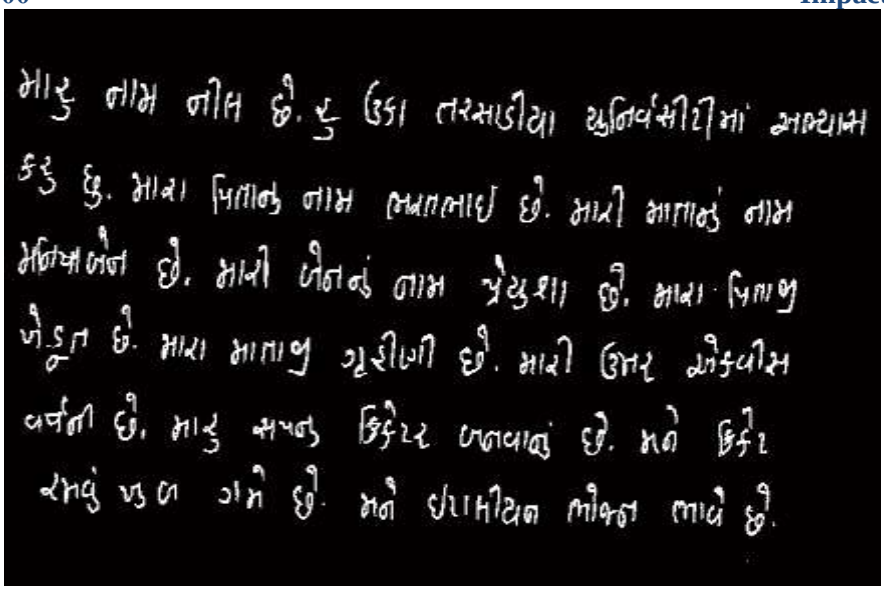

Figure 13. Image Smoothing

Cropping:

Cropping an image means creating a new image from a part of an original image**.** It is require for removing extra space from a document image. For doing the cropping of a document image we have used horizontal & vertical projection of a document image.

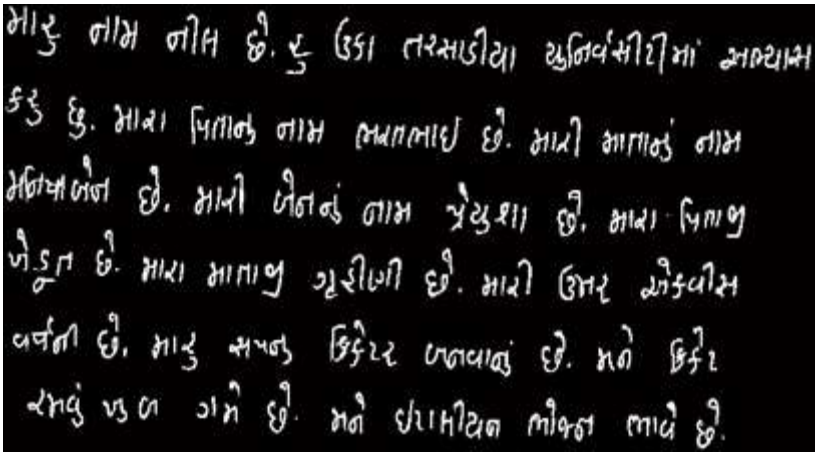

Figure 14. Image Cropping

Skew Correction

Skew Correction is the process of correcting skews present in a document image. Most of the handwritten documents contain skews because of the diverse human writing style & because of that there will be problem occur while segmenting those handwritten document. So, it is require for correcting the skews present in a document image. For correcting skew in a document image we have rotate an image with a specified angle.

ISSN: 2277-9655
Impact Factor: 3.785

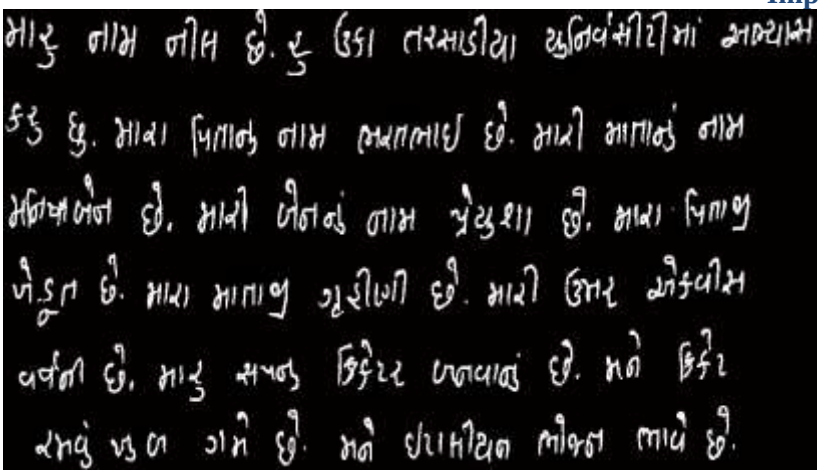

Figure 15. Skew Correction

Line Segmentation:

It is the process of segmenting a scanned document in number of lines present in that document. We have used horizontal projection technique for doing line segmentation in a handwritten Gujarati document.

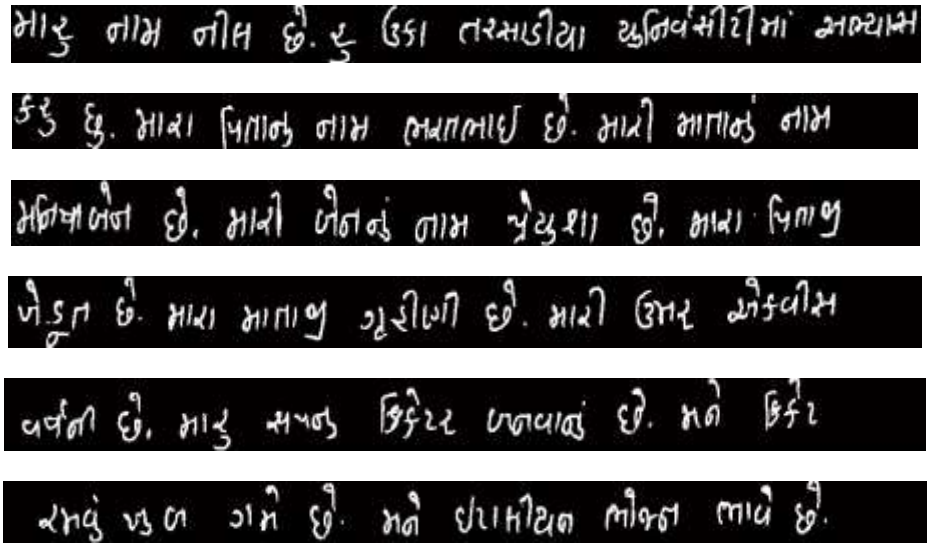

Figure 16. Line Segments

Word Segmentation:

It is the process of segmenting a scanned document in number of words present in that document. We have used vertical projection technique for doing word segmentation in a handwritten Gujarati document.

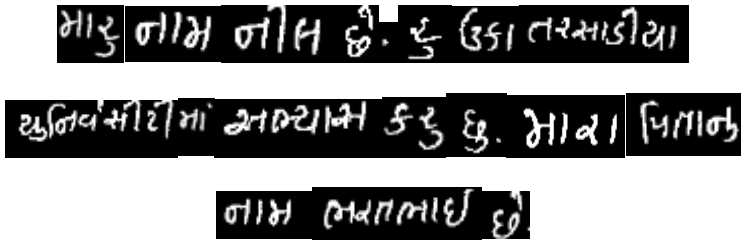

Figure 17. Word Segments

ISSN: 2277-9655
Impact Factor: 3.785

Character Segmentation:

It is the process of segmenting a scanned document in number of characters present in that document. We have used bounding box technique for doing character segmentation in a handwritten Gujarati document.

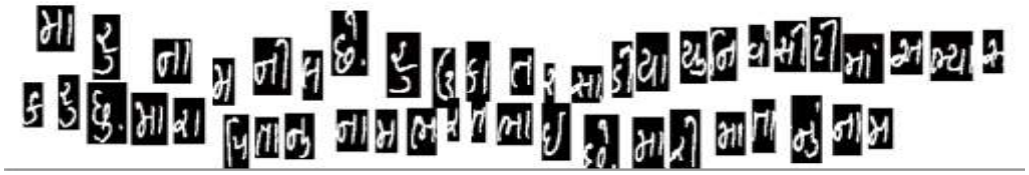

Figure 18. Character Segments

Result Set:

Line Segmentation:

Types of Documents	Number of Documents	Number of lines in Documents	Correct detected lines	Incorrect detected lines	Accuracy
Without modifier	15	127	110	17	86.61%
With modifier	48	451	420	31	93.13%

Word Segmentation:

Types of documents	Number of correct lines	Number of words	Correct detected words	Incorrect detected words	Accuracy
Without modifier	110	1100	907	193	82.45%
With modifier	420	4200	3798	402	90.42%

Character Segmentation:

Types of documents	Number of correct words	Number of character	Correct detected character	Incorrect detected character	Accuracy
Without modifier	907	20861	19250	1611	92.27%
With modifier	3798	87354	83834	3520	95.97%

Table 3.1 Result Set

CONCLUSION

Using “Horizontal Projection” technique, the line segmentation has done with the “93.13%” of accuracy after correcting the document skew. And using “Vertical Projection” technique, the word segmentation has done with “82.45%” of accuracy & using “Connected Component” technique, character segmentation has done with “95.97%” of accuracy. The result will improve if a modifier connected with a character can be segmented separately rather than segmented with modifiers.

ACKNOWLEGDEMENT

This paper would not have been possible without the support of several thoughtful and generous individuals. Foremost among those are our guide **Mr. Jitendra Nasriwala & Mr. Rakesh Savant**, who have been a tremendous mentor for us. We would like to thank them for encouraging our research and for allowing us to grow as a research scientist. Their advice on the research has been priceless throughout the stages of dissertation. Discussion with them always leads us in the direction of innovation. Their knowledge in this domain has always helped us to solving technical

obstacles and finishes this dissertation.

REFERENCES

[1] Www.wikipedia.com.
[2] Chhaya Patel and Apurva Desai, Extraction of characters and Modifiers from Handwritten Gujarati Words, International Journal of Computer Application, vol 73, issue 3, 2013.
[3] Prof S K Shah, Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching, IE (I) Journal- ET , vol 86 , 2006.
[4] Hetal R Thaker and C K Kumbharana , Study of Different Off-line Handwritten Character Recognition Algorithms for Various Indian Scripts , International journal of Computer Applications , vol 65 , 2013.
[5] Aurelie Lemaitre, "A perceptive method for handwritten text segmentation", Document recognition and retrieval XVII – Electronic Imaging, 2011.
[6] A.Nicolaoul and B. Gatos, "Handwritten Text line segmentation by shredding text into its line", department of informatics, technological educational institute of Athens, 2009.
[7] Apurva Desai, "Extraction Of Characters and Modifiers from Handwritten Gujarati Words", International Journal Of Computer Application, vol.73, issue 7, 2013.
[8] Safwan W shah al, International Conference on Document Analysis and Recognition, 2009.
[9] prof s k shah , "design and implementation of optical character recognition system to recognize Gujarati script using template matching" ,ie (i) journal- et , vol 86 , 2006.
[10] aurelie lemaitrea , a perceptive method for handwritten text segmentation , document recognition and retrieval xvii – electronic imaging , 2011.
[11] a. Zahour, b. Taconet, p. Mercy, and s. Ramdane,"arabic hand-written text-line extraction", proceedings of the sixth international. Conference on documentanalysis and recognition, icdar, pp. 281–285, 2001.
[12] n. Tripathy and u. Pal., "handwriting segmentationof unconstrained oriya text", international workshop on frontiers in handwriting recognition, pp. 306–311, 2004.
[13] naresh kumar garg alt."a new method for line segmentation of handwritten hindi text" ,seventh international conference on information technology,2010.
[14] l. Likforman-sulem and c. Faure, "extracting text lines in handwritten documents by perceptual grouping", advances in handwriting and drawing : a multidisciplinary approach, pp. 21-38, 1994.
[15] i.s.i. abuhaiba, s. Datta and m.j.j. holt, "line extraction and stroke ordering of text pages", proceedings of the third international conference on document analysis and recognition, canada, pp. 390-393, 1995.
[16] shailesh chaudhari and dr. Ravi gulati "segmentation problems in handwritten Gujarati text" international journal of engineering research & technology (ijert) vol. 3 issue 1, january – 2014.
[17] r. Indra gandhi alt." 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, 2010.
[18] B.Gatos et al., "Efficient off-Line Cursive Handwriting Word Recognition", Institute of Informatics and Telecommunications National Centre, vol. 3, issue 4, pp.104-150, 2015.
[19] Kaustubh Bhattacharyya and kandarpa kumar sarma, "ANN-based innovative segmentation method for handwritten text in assesse", IJCSI International journal of Computer Science, vol 5, 2009.
[20] Shailesh Chaudhari and Dr.Ravi Gulati, "Segmentation Problems in Handwritten Gujarati Text", International Journal of Engineering Research & Technology, vol 3 issue 1, 2014.
[21] Naresh Kumar Garg ate, "A New Method for Line Segmentation of Handwritten Hindi Text", Seventh International Conference on Information Technology, 2010.
[22] Nisha Sharma ate, "Recognition for Handwritten English Letters: A Review", International Journal of Engineering and Innovative Technology (IJEIT) Vol 2, Issue 7, January 2013.
[23] Mamatha H R and Srikantamurty K, "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document", International Journal of Applied Information Systems (IJAIS), vol 4, 2012.
[24] Vassilis Katsouros and Vassilis Papavassiliou, "Segmentation of Handwritten Document Images into Text Line", Institute for Language and Speech Processing,2009.