iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** X **Month of publication:** October 2020

DOI: https://doi.org/10.22214/ijraset.2020.31939

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

Segmentation Methods: A Review

Nikita Mehta[1], Dr. Jyotika Doshi[2]
[1]*PhD Scholar, GLS University, Ahmedabad, Gujarat*
[2]*Retd. Associate Prof., GLS (SRP) ICT, Ahmedabad, Gujarat*

Abstract: ***This paper presents a review of different segmentation techniques. Here, an overview of some important and widely used segmentation techniques are presented. There are two types of documents: machine printed and handwritten. Segmentation of a handwritten document is more difficult than printed one. Choosing and applying right segmentation technique highly affects the character recognition rate. A document is segmented on different levels to extract the smallest individual unit of the text – an individual character. Page segmentation, line segmentation, word segmentation and character segmentation are different levels of segmentation which are discussed in this paper.***

General Terms: Optical character recognition (OCR), Segmentation, Page segmentation, line segmentation, word segmentation, character segmentation et. al.
Keywords: Optical Character Recognition (OCR), Segmentation, Page segmentation, line segmentation, word segmentation, character segmentation, Projection profile, Hough transform

I. INTRODUCTION

The aim behind Optical Character Recognition (OCR) is to create human like perception and character identification by artificial systems. A lot of research has been done and still being done for character recognition of different languages.
The optical character recognition work is divided in different phases. Figure 1 shows the different phases of OCR.

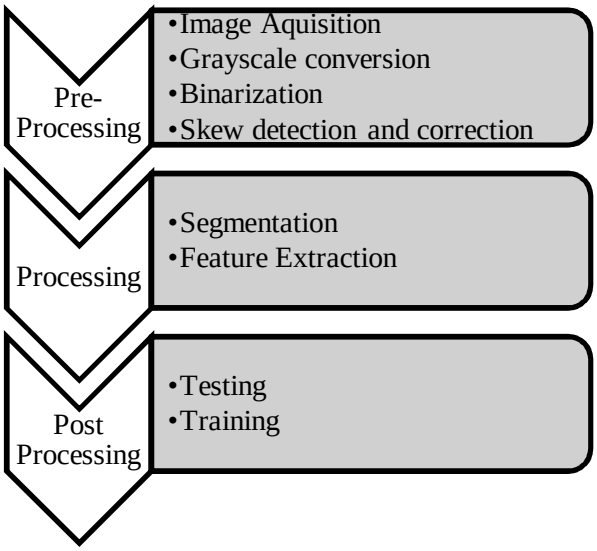


Figure 1: Phases of OCR

Pre-processing phase includes the processes that makes the input image clean and noise free. It includes the steps like image acquisition, grayscale conversion, binarization, skew detection and correction etc. Aim of this phase is to produce an image that is ready for the processing.
Next is processing phase where actual character recognition takes place. To recognize any character, first we need to extract it from the whole document text. This is the work of segmentation process. It extracts the smallest individual character from the document text image. Then feature extraction is done for the individual character and finally character is identified.

In post processing the accuracy of correctly identified characters is tested and system trained against the false identification.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

A. Segmentation

Segmentation is one of the crucial step of OCR system. It highly affects the accuracy of character recognition. Segmentation decomposes a document image into many sub-images or individual symbols [1]. These individual sub-images contain some identifiable unit for which the system is designed to recognize.

Segmentation can be classified according to the level on which it is applied like:

1) Page Segmentation
2) Line Segmentation
3) Word Segmentation
4) Character Segmentation

a) *Page Segmentation:* Page Segmentation is process which divides the scanned document image into homogeneous blocks like text blocks, figures, tables or images [7]. For OCR system, page segmentation is not always required process. This type of segmentation is applied to the document images which contain some graphical representation or images or tables etc. For OCR project, page segmentation separates text area and non-text area [8].
b) *Line Segmentation:* Line segmentation divides the text area in multiple horizontal lines. Line segmentation in handwritten documents have multiple challenges like line gaps are not uniform, skewed lines etc. many Indian scripts use upper and lower modifiers (matras) which is a big challenge as it can touch upper and lower lines.
c) *Word and Character Segmentation:* Both word and character segmentation can be achieved with similar techniques. They are used to segregate words and characters from lines. Words segmentation is important for the applications that builds a dictionary or where spell check or translation is needed to be done. A character is the smallest individual unit on the document image. OCR applications work on characters to identify it.

B. Factors that affects the Segmentation

There are some factors of document image which will affect the segmentation. Segmentation method for any document image should be chosen with considering following characteristics:

1) Document pattern

Document pattern includes all characteristics of a document like:

a) Layout of document – portrait or landscape
b) If the document contains images, pictures, tables, text etc.
c) Size of the document
d) Age of the document
e) Color of the document
f) Noise (any dots or patches other than data) contained by the document
g) Number of columns in a page

2) Language characteristics

All languages have different characteristics while writing the text. These could be:

a) Modifiers used in language (at upper, lower, left and right of the character)
b) Joint characters (joining pattern like: up – down join or side by side join)
c) Spacing between words and characters

3) Writing style

Each writer writes text with different styles which makes his/her writings different than others. This style includes:

a) Is document printed of handwritten?
b) Is it written with cursive letters?
c) If writer has written text with any skew angle
d) Pressure given on pen
e) Size of the font
f) Uniformity of fonts

II. RELATED WORK

Some very popular and widely used segmentation techniques are listed below. There are many other methods proposed by different researchers but most of them uses following techniques as their base and then modify it somewhat.

1) Projection Based Methods
a) X – Y cut
b) White space analysis
c) Smearing

2) Grouping Methods
a) Connected component
b) Bounding boxes

3) Hough Based Methods
4) Graph based methods
a) Docstrum

The x-y cut or recursive x-y cut algorithm [9] [10] is a top down approach decomposes a document image into a set of rectangular blocks. For this, horizontal and vertical projection profile is calculated.
Then a zone division is performed based on the most prominent valley in either of the projection profile. This process is recursively done until no prominent valley found in either direction. Same projection profile techniques can be used for line, word and character segmentation.
The classical run – length smearing algorithm (RLSA) works on binary images [8] [12]. Black pixel on the image represented as 1 and white pixel is represented as 0.
This algorithm converts binary sequence b1 to b2. Any 0 in b1 is converted to 1 in b2 if it has laser adjacent 0s than a predefined threshold value [2].
This algorithm applied on image row wise and column wise with different threshold value and generates two different bitmap images.
A logical AND operation is performed on two resultant bitmap images to combine it. Then, connected component analysis performed on the image to identify different zones.
There are several methods for document layout analysis. It performs global segmentation of document. It divides the document into distinct geometric regions corresponding to entities like columns, paragraphs, tables, headings using features like proximity, texture, or whitespace [13]. Then each individual region is processed separately.
White space and pitch is a common technique which observes the white spaces between two words or characters [1]. In machine printed documents these spacing between words and characters are quite uniform. So, it can
Grouping is a bottom – up strategy for segmentation. In this technique first the bounding box of the connected component (the smallest rectangle which circumscribe the connected component) [9], or other units like blocks or salient points are identified. These units are then joined together to form alignments. The joining scheme depends on local and global criteria [6].
Docstrum [11] is a graph based method where the relationship of the objects on document are expressed with polar coordinates (distance and angle).
KNN (k – nearest neighbor) pairs between objects are considered for the image segmentation. Text orientation and the spacing parameters are estimated in this method. This algorithm works for most document layouts.
The Hough transform is used for many purposes like skew detection and line segmentation [15]. There are two ways of Hough transform – pixel based and block based.
It can be used to determine skew and slope of the elements. This method can be used for documents with variations in the skew angle between the text lines.

III. EXPERIMENT RESULTS COMPARISON

As discussed in section 2, various methods are proposed by many researchers for different languages. Table 1 shows experiment results of some segmentation methods.

Table 1: Experiment results comparison

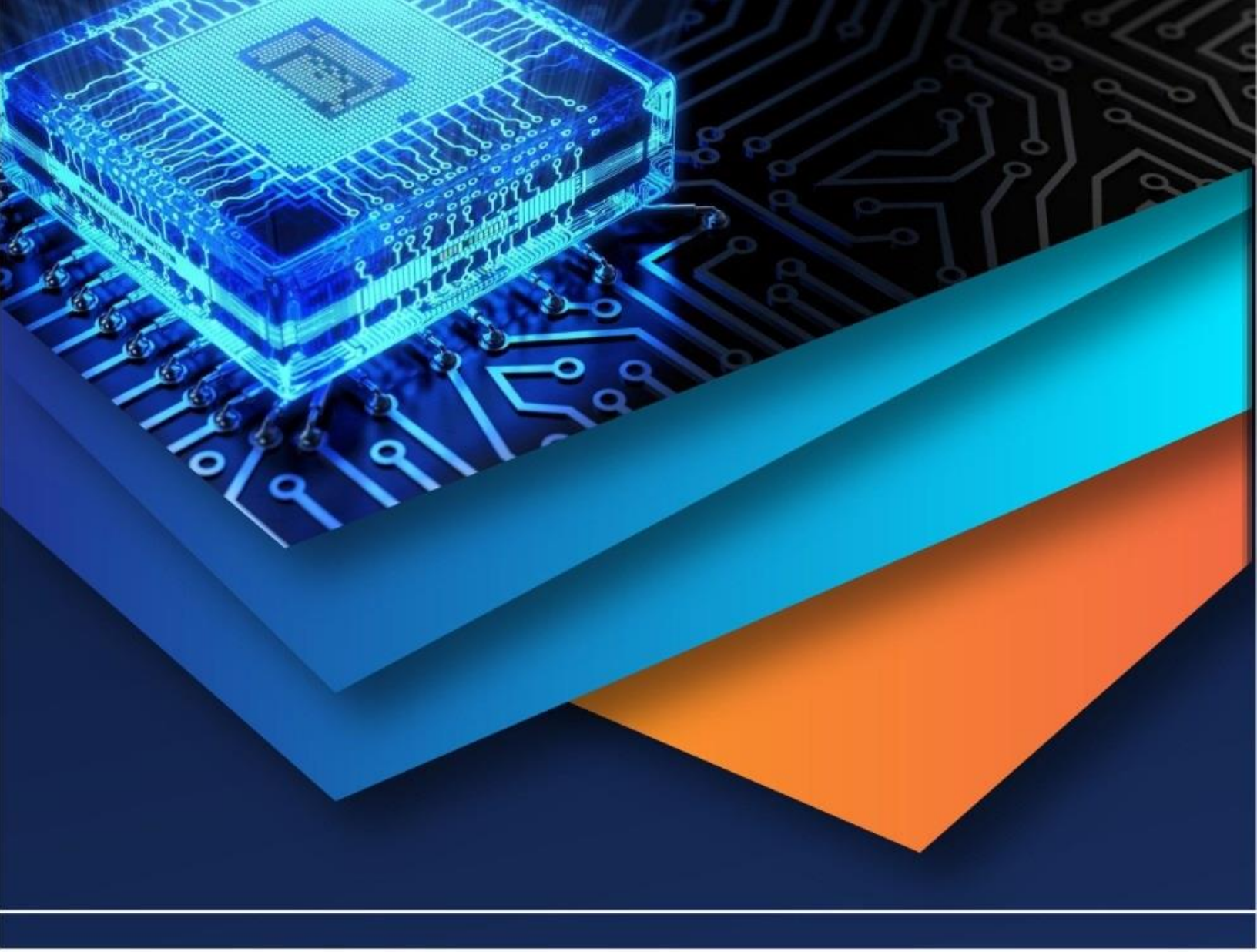
Sr. No.	Reference	Language	Segmentation type	method used	Dataset	Correct Segmentation	Accuracy achieved
1	Gupta et. al. [16]	Handwritten Hindi	Character Segmentation	Polygonal approximation	3426 characters	3279	95.7
2	Gupta et. al. [17]	Handwritten Hindi,	Character Segmentation	Polygonal approximation	9034 characters	8208	90.86
		Handwritten Marathi,			14437 characters	12986	89.95
		Handwritten Punjabi,			9160 characters	8307	90.69
		Handwritten Bangla			7305 characters	6485	88.77
3	Arefin et. al. [18]	Handwritten Bangla	Line Segmentation	Distance based segmentation and histogram based gradients	Not mentioned	98	91.5
			Word Segmentation		500 words	372	84.73
			Character Segmentation		Not mentioned	1124	86.06
4	Ramteke et. al. [19]	Handwritten Marathi,	Line Segmentation	Projection based method	1200 lines	1000	83
			Word Segmentation		400 words	362	90
			Character Segmentation		1990 characters	1720	86
5	Din et. al .[20]	Printed Urdu	Line Segmentation	projection profile and heuristics based segmentation	310 lines	306	98.7
			Character Segmentation		7364 characters	6811	92.5
6	Garg et. al. [3]	Handwritten Hindi	Line Segmentation	Projection profile based method	200 lines	183	91.5
			Word Segmentation		1380	1354	98.1
			Character Segmentation		3870	3862	79.12

IV. CONCLUSION

Segmentation method for any OCR system should be chosen based on the document type and language used in document image. Projection based approach works well with machine printed text or a clean handwritten document with minimum skew angle and where line gaps are significant. But this method is not appropriate when width of character is variable or characters are connected or text is slanted [1] [3]. For connected characters, some different approach is needed. In [4], they have proposed fuzzy multifactorial analysis implementation for touching characters. Smearing techniques gives consistent result under reasonable variation in skew angle and character size. Connected-component or grouping technique is useful for handwritten documents where characters are of different sizes and uneven spacing. Hough transform is a good choice when document have skewed lines with variant angles.

REFERENCES

[1] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 18, pp. 690-706, 1996.
[2] Z. Razak, K. Zulkiflee, M. Y. I. Idris, E. M. Tamil, M. N. M. Noor, R. Salleh, M. Yaakob, Z. M. Yusof and M. Yaacob, "Off-line handwriting text line segmentation: A review," International journal of computer science and network security, vol. 8, pp. 12-20, 2008.
[3] N. K. Garg, L. Kaur and M. Jindal, "Segmentation of handwritten hindi text," International Journal of Computer Applications (IJCA), pp. 22-26, 2010.
[4] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 32, pp. 449-459, 2002.
[5] Y. Li, Y. Zheng, D. Doermann and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, pp. 1313-1329, 2008.
[6] L. Likforman-Sulem, A. Zahour and B. Taconet, "Text line segmentation of historical documents: a survey," International journal on document analysis and recognition, vol. 9, no. 2, pp. 123-138, 2007.
[7] Kaur, S., Mann, P. S., & Khurana, S. (2013). Page segmentation in OCR system-a review. Inetrnational Journal of Computer Science and Information Technologies, 4(3), 420-422.
[8] Shafait, F., Keysers, D., & Breuel, T. (2008). Performance evaluation and benchmarking of six-page segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(6), 941-954.
[9] Ha, J., Haralick, R. M., & Phillips, I. T. (1995, August). Recursive XY cut using bounding boxes of connected components. In Proceedings of 3rd International Conference on Document Analysis and Recognition (Vol. 2, pp. 952-955). IEEE.
[10] Nagy, G., Seth, S., & Viswanathan, M. (1992). A prototype document image analysis system for technical journals. Computer, 25(7), 10-22.
[11] A. Simon, J. -. Pret and A. P. Johnson, "A fast algorithm for bottom-up document layout analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 3, pp. 273-277, March 1997, doi: 10.1109/34.584106.
[12] K.Y. Wong, R.G. Casey, and F.M. Wahl, “Document Analysis System,” IBM J. Research and Development, vol. 26, no. 6, pp. 647- 656, 1982
[13] Breuel, T. M. (2002, August). Two geometric algorithms for layout analysis. In International workshop on document analysis systems (pp. 188-199). Springer, Berlin, Heidelberg.
[14] Dave, N. (2015). Segmentation methods for hand written character recognition. International journal of signal processing, image processing and pattern recognition, 8(4), 155-164.
[15] Ptak, R., Żygadło, B., & Unold, O. (2017). Projection–based text line segmentation with a variable threshold. International Journal of Applied Mathematics and Computer Science, 27(1), 195-206.
[16] Gupta, D., & Bag, S. (2018, February). An Efficient Character Segmentation Approach for Handwritten Hindi Text. In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 730-734). IEEE.
[17] Gupta, D., & Bag, S. (2019). Handwritten multilingual word segmentation using polygonal approximation of digital curves for Indian languages. Multimedia Tools and Applications, 78(14), 19361-19386.
[18] Arefin, N., Hassan, M., Khaliluzzaman, M., & Chowdhury, S. A. (2017, December). Bangla handwritten characters recognition by using distance-based segmentation and histogram oriented gradients. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 678-681). IEEE.
[19] Ramteke, S., Gurjar, A. A., & Deshmukh, D. S. (2016, December). Automatic segmentation of content and noncontent based handwritten Marathi text document. In 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC) (pp. 404-408). IEEE
[20] Din, I. U., Malik, Z., Siddiqi, I., & Khalid, S. (2016). Line and ligature segmentation in printed Urdu document images. J. Appl. Environ. Biol. Sci, 6(3), 114-120.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY