Developments in the Field of Text Segmentation of Handwritten Script in OCR

Soureen Das[1], Monit Kr. Sharma[2], Chandan Jyoti Kumar[3]

[1, 2, 3] *Dept. of CSE and IT, Don Bosco College of Engineering and Technology,*
Assam Don Bosco University, Azara, Guwahati, INDIA

ABSTRACT

OCR (Optical Character Recognition) deals with converting scanned images of either handwritten or printed text into machine encoded texts. It is a type of offline character recognition where the texts are recognized after they have been written. The text segmentation stage plays a very vital role in OCR. Text segmentation can be further sub-divided into three phase line segmentation, word segmentation and character segmentation. The accuracy with which an OCR works depends on how well the lines, words and characters are separated. A good OCR can be used in various fields like form reader, check reader, bill processing, number plate recognition, preservation of historical documents etc.. This paper presents a detailed report of different technologies and techniques used in segmenting text. It also provides information about various researches that have been carried out in the field of text segmentation in OCR. Some of the challenges faced during segmenting texts and future scope of this field has also been discussed.

Keywords: OCR, Text Segmentation, Line Segmentation, Word Segmentation, Character Segmentation

1. INTRODUCTION

OCR takes scanned images of texts (handwritten or printed) and converts it into machine encoded text, which is then used for character recognition. The digitized text which is obtained from OCR can be stored more easily and it also requires less space. Efficient electronic searches can be performed on the these digitized texts and they can be used by various machine processes.

OCR is used in check readers, address readers, form readers, bill processing systems, number plate recognition systems, historical document readers, CAPTCHA antibot systems etc.

Generally OCR consists of 7 steps:

1. Input
2. Pre-processing

3. Segmentation
4. Feature Extraction
5. Classification
6. Post-processing
7. Output

In input stage input is obtained by loading an image, parsing previously recorded data, or by loading scanned data or captured data.

In pre-processing stage various noise removal techniques are applied to make the segmentation of texts easier. Processes like enhancement, smoothing, filtering and cleaning up of the input digital image take place. Using preprocessing, implementation for subsequent stages can be made more easy and accurate. There are various methods for pre-processing the data like binarization, skeletonization, noise removal, smoothing, contour smoothing etc.

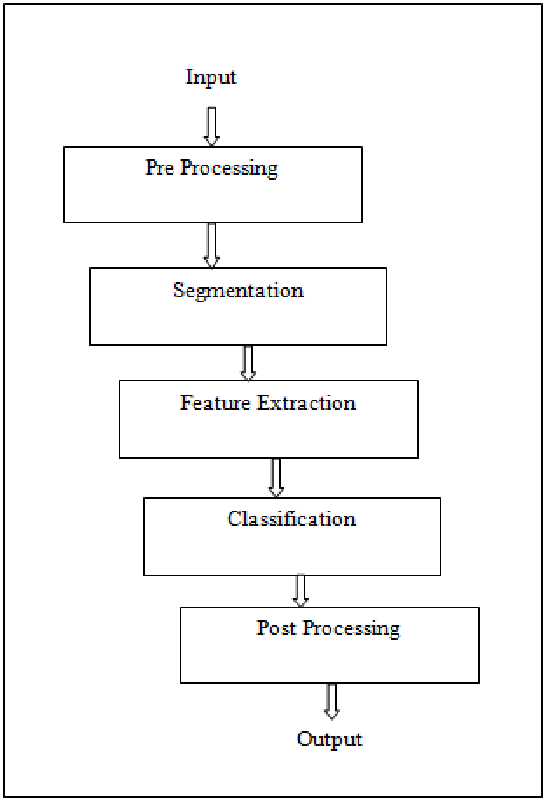

Fig. 1. Stages of OCR

In segmentation stage the whole text body is decomposed into individual characters. It includes line, word and character segmentation. Character segmentation plays a vital role in character recognition systems.

In feature extraction the information about the pattern of the segmented text image is extracted. It helps in recognizing the characters in the text. It collects the information about the shape of a pattern, which is provided as input for the classification stage. The selection of a proper set of features is very important for pattern recognition system design.

In classification stage the features extracted in the previous stage is used to identify the text segment according to predetermined rules.

Post-processing uses linguistic knowledge and contextual information to reduce errors in the system. The two most common post processing techniques for error correction are dictionary lookup and statistical approach.

In output stage the final output is obtained and is used according to the need.

The paper is divided into five sections. The first section provides a basic introduction on optical character recognition, uses of optical character recognition and text segmentation. The second section explains the text segmentation process and the various challenges faced during this process. The third section deals with related work in the field of text segmentation and the various techniques and technologies used. The fourth section contains conclusion and overview of the status of various works done in text segmentation and the fifth and final section contains the references.

2. TEXT SEGMENTATION

The segmentation phase of OCR includes line segmentation, word segmentation and character segmentation. In line segmentation the entire text region is segmented into text lines. Word segmentation deals with segmenting each individual word from a text line. Character segmentation separates each individual character from word. The block diagram of segmentation process is given below in figure 2.

There are various issues faced while segmenting handwritten texts. For handwritten script, the skewness of a line varies from writer to writer, which makes line segmentation difficult. The presence and absence of header lines, close proximity of two nearby words with each other makes word segmentation difficult. Compound characters and various vowel modifiers are difficult to segment. Overlapped characters causes problem in character segmentation.

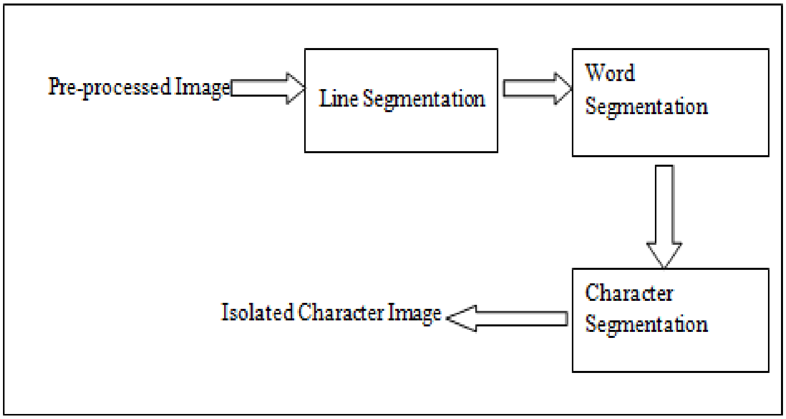

Figure 2. Block diagram of segmentation process.

3. RELATED WORK AND OBSERVATIONS

S. Basu et al. [14] presented a technique where line segmentation and word segmentation is done using row and column histogram respectively. The isolated word is subdivided into four horizontal sections to find segmentation points. And then character segmentation is done using connected component labelling algorithm. The overall segmentation success was 97.7%.

U. Pal et al. [7] proposed a line segmentation method which is based on horizontal projection and the segmentation of lines into words by vertical projection. For character segmentation, isolated and touching characters in a word are detected. Water reservoir concept is then used to segment touching characters in a word. The water reservoir concept states that if water is poured from top of a component, the water will be settle in the bottom cavity regions of the component reservoirs Accuracy of text line segmentation was about 99%., word segmentation 97.8%, character segmentation 97.69% and for the connected character accuracy was 95.97%.

T. K. Bhowmik et al [8] segmented Bengali words based on detection and correction of the skewness of lines. Analysis of directional chaincode and its positional information is used to find the features for segmentation. Multilayer Perceptron (MLP) Neural Networks are used for recognition of the segmenting points. These segmentation points are used for segmentation, which is carried out with 88% of accuracy.

Li et al. [16] proposed a text line detection technique for handwritten documents. Text line structures are enhanced using Gaussian filter. Gaussian filter converts binary image into grey scale. Level set method is used to estimate text line boundaries and segment merging technique is used to extract final result. Text lines were detected with a great accuracy.

A.R. Khan et al. [5] presented an approach for unconstrained handwritten words using neural network. The original grey scale image is converted to a binary image and then to a skeleton format. The sum of foreground pixels for each columns are stored as candidate segment columns (CSC). A threshold value is then determined from CSC for carrying out character segmentation. A neural network is used to properly segment m, n, u, v and w.

J.W. AlKhateeb et al. [6] used distances among different connected components to determine a threshold for word segmentation. A modified and improved projection based method is implemented for baseline detection. Using Bayesian minimum classification optimal distance d_0 is obtained.

$$d_0 = argmin(err(d))$$

$$err(d) = \int_d^\infty p_{s-w}\,(x)dx + \int_0^d p_w\,(x)dx \int_d^\infty p_{s-w}\,(x)$$

Here, $p_w(d)$ and $p_{s-w}(d)$ are probability representing separation of words and sub words. Words with distance, d > d_0 are then segmented. The system was tested on IFN/ENIT database and out of 200 images 85% were correctly segmented, 4% over segmented, 9% under segmented and 2% misplaced segmented.

N.K. Garg et al. presented [9] used a line segmentation method which is based on header line and base line detection. The word segmentation is done by vertical projection. For each column of the line the number of black pixels is counted and columns with zero black pixels are used for word separation. After headerline detection of each word vertical projection is used for character segmentation. Accuracy of text line segmentation was 91.5 % out of 200 total lines. Accuracy of word segmentation was 98.1% out of 1380 words. Accuracy of character segmentation was found out to be 85.74%.

A. Lemaitre et al. [1] proposed a method that deals with finding the text lines in a low resolution image, and then at a higher level of resolution position of the baselines are determined. Cooperation among digital data and symbolic knowledge is applied for word segmentation. The database is made of handwritten pages by different writers in four in English, French, German and Greek. 99.25% of text lines have more than 95% correct pixels while 94.20% of words have more than 90% correct pixels.

H. Adiguzel et al. [12] used a hybrid approach for line segmentation in handwritten documents. Text line segmentation is done by combining connected component based methods and projection based methods. Detection of line is done by grouping baselines of connected components of each line using projection techniques. A new method 'Fourier curve fitting' is proposed for detection of peaks in a projection profile. A pixel based matching score (MS) is used for evaluation of result. The algorithm detected different printed and handwritten Ottoman datasets with at least 92% accuracy.

I.B. Messaoud [2] et al. used three different methods for text line segmentation. The first method is bottom-up method where vertical projection technique is used.

$$\mathrm{H(y)}^{\mathrm{r}} = \sum_{x=1}^{Nx} \mathrm{Ib(x, y)}, \forall\ \mathrm{y} \in \{1, \dots, \mathrm{Ny}\},$$

Where H is the histogram of the binary image for each position y in Ib. Top-down method groups the connected components and detects the nearest neighbour for estimation of text lines.

$$\exists\ i', \text{such as } ||pos_{i'},o_k|| = \min(||pos_{i'},o_k||)\forall i,\ 1\leq i\leq Nr$$

Where a_k, o_k is the area and gravity centre of each connected component respectively, pos_i is the local means of the positions of o_k gravity centres. The third method is the nearest neighbour method which is used for correction of overlapping and touching problems.

$$||N^{o}(p), p||=\min(p^{'}, p),$$

Where $N^{o}(p)$ is the nearest neighbour in pr_b to p and $N^{o}(p) \notin o_k$. 60 images of handwritten Latin documents from the 9th century are used as datasets. This framework was tested on various historical documents and the result obtained was satisfactory.

A. Kumar et al. [3] used modified horizontal projection to segment Gurumukhi text. Piece-wise separating lines (PSL) are calculated for performing line segmentation. The algorithm successfully segmented Gurumukhi texts lines when lines were not very close to each other and no overlapping of words was present.

C.J. Kumar et al. [17] used seam curve and morphological based segmentation method for segmenting machine printed Indian texts. line extraction is done by first making the image binarized. Then energy map is computed using Signed Distance Transform (SDT). The seam is then calculated using the following equation.

$$S = \{x(i), i\}, \forall i, | x(i) - x(i-1)| \leq K$$

$$e(S) = e(\{x(i), i\}) = \sum E(x(i))$$

S is the seam and e(S) is the energy cost. Word segmentation is carried out using region and hole filling, connected component detection and half matra checking. Techniques like header line removal and thinning operations are performed for character segmentation. The algorithm is tested on multi script documents and the accuracy of text, word and character segmentation obtained is 99.48, 99.60 and 97.31 respectively.

4. CONCLUSION

An OCR for recognizing handwritten or printed texts can be used for many purposes like in reading checks, addresses, forms, historical documents, bills etc. From the various studies conducted on the text segmentation methods of OCR system it has been observed that the segmentation is the most critical stage of optical character recognition. Any wrong segmentation will lead to a faulty OCR. Segmentation of handwritten text is not easy. Handwritten texts vary from writer to writer. There is variation in skewness of lines, presence and absence of header lines in words, close proximity of two nearby words, compound characters, various vowel modifiers and overlapped characters. All these cause problem in text segmentation process. As segmentation plays a very important part in optical character recognition systems, a good segmentation process will result in a good OCR which can help in reducing man hours wasted in doing the above mentioned activities and can also be used for various machine understanding and learning processes.

REFERENCES

[1] A. Lemaitre et al. "A Perceptive Method for Handwritten Text Segmentation". In Proceeding of Document recognition and retrieval XVIII - Electronic Imaging, San Francisco, United States, 2011. Vol.7874, pp 9.

[2] I.B. Messaoud et al. "A Multilevel Text line Segmentation Framework for Handwritten Historical Documents". In Proceeding of International Conference on Frontiers in Handwriting Recognition, 2012, pp 515-520.

[3] A. Kumar et al. "Segmentation of Handwritten Gurmukhi Text into Lines". In Proceeding of International Conference on Recent Advances and Future Trends in Information Technology (iRAFIT2012), 2012, pp 353-356.

[4] M.K. Jindal1 et al. “Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts” International Journal of Computational Intelligence Research. Vol.3, No.4 (2007), pp. 277–286

[5] A.R. Khan et al. "A Simple Segmentation Approach for Unconstrained Cursive Handwritten Words in Conjunction with the Neural Network". In Proceeding of International Journal of Image Processing, volume 2 issue 3(2008), pp 29-35.

[6] J.W. AlKhateeb et al. "Component-based Segmentation of Words from Handwritten Arabic Text" in "International Journal of Computer Systems Science and Engineering "2009, Vol. 5, No. 1, pp 54-59.

[7] U. Pal et al. “Segmentation of Bangla Unconstrained Handwritten Text” in proceeding of the "Seventh International Conference on Document Analysis and Recognition". 2003, Vol. 2, pp 1128-1132.

[8] T. K. Bhowmik et al “Character Segmentation For Handwritten Bangla Words Using Artificial Neural Network”.

[9] N.K. Garg et al. "Segmentation of Handwritten Hindi Text". In Proceeding of "International Journal of Computer Applications", 2010, Vol. 1, No. 4, pp 22-26.

[10] C.J. Kumar et al. "Recognition of Handwritten Numerals of Manipuri Script". In Proceeding of "International Journal of Computer Applications (IJCA)" December, 2013, Vol. 84, No-17, pp 1-5.

[11] B. B. Chaudhuri et al. “A complete printed Bangla OCR system”. In Proceeding of Pattern Recognition, vol. 31, no. 5, pp 531-549, 1998.

[12] H. Adiguzel et al. "A Hybrid Approach for Line Segmentation in Handwritten Documents". In Proceeding of International Conference on Frontiers in Handwriting Recognition, 2012, pp 503-508.

[13] G. S. Lehal et al. “Text segmentation of machine printed Gurmukhi script”. In Proceeding of SPIE, vol. 4307, pp. 223-231, 2001,.

[14] S. Basu et al. "Segmentation of Offline Handwritten Bengali Script". In Proceeding of 28th IEEE ACE, pp. 171-174, Dec 2002, Science City, Kolkata.

[15] E. Bruzzone et al. “An algorithm for extracting cursive text lines”. In Proceeding of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 749–752.

[16] Y. Li et al. "Detecting text lines in handwritten documents". In Proceeding of 18th International Conference on Pattern Recognition, 2006, vol. 2, pp. 1030–10.

[17] C.J. Kumar et al. "Seam Carving and Morphological Based Segmentation of Machine Printed Indian Scripts". In Proceedings of 7th International Conference on Advanced Computing and Communication Technology. November, 2013.