Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

A Survey on Character Segmentation in Historical Document Images

Anjana Ramachandran
M.Tech Image Processing
College of Engineering Chengannur
Kerala, India

Jyothi R L
Assistant Professor
College of Engineering Chengannur
Kerala, India

***Abstract*—Character segmentation is one of the main key element of an OCR system. This fact is clearly illustrated while comparing the high recognition rates of single isolated characters with connected string characters and words. Without proper segmentation, a recognition system remains inefficient. When considering about characters, there is a need to talk on ancient documents which comprises thousands of ancient characters. An immeasurable amount of historical documents with priceless cultural legacy was left behind by our progenitor to tell foregone stories, especially in written document format. Years of storage, makes these historical documents to suffer from degradation. The difficulty of how to safeguard this priceless culture legacy to the next peer group received extreme curiosity from countless experimenters. One way to protect such precious particulars on historical knowledge and artistic work is Historical document digitization. Historical documents are digitized through photographing, followed by document segmentation, recognition, preservation, management, and research. Among all the above-mentioned stages, document segmentation is conducted as first step and the overall digitization performance of the system heavily depends on the segmentation quality. This paper comes up with a review on several segmentation techniques and the aim is to give an appreciation for the collection of techniques that have been developed.**

Keywords*— *Optical character recognition, Character segmentation, Character Recognition, Recognition-based segmentation.

I. INTRODUCTION

Historical Documents are indigenous documents that consist of predominant historical information about an event a person, place, or about various traditions. Different forms of writings exist which depends on the region and culture where it originates. Both characters and pictures were used by ancient people in their writings to express their ideas, for communication and to pass their cultural system to the next generations. To preserve these documents, digitization techniques were implemented, of which Document segmentation play the main active role. Character segmentation is the process of decomposing a character sequence image into sub-images of individual symbols. It plays foremost role in Character Recognition.

Recognition procedure can be explained in three steps:

1. Given a document image, find a starting point in the image.
2. Then go on looking for the next character image.
3. After getting the next character image, extract distinguishing features of the character image.
4. From the dataset of symbols present, find a character which best matches to the input character image and output its identity.

In the execution of step 1, the segmentation step, needed to answer a simple question: "What makes up a character?" Whoever whether it's experimenters or analyst who tried to find solution to this question in an algorithmic way found themselves in a Catch-22 state. A character is a specimen that look alike one of the symbols recognized by the system. The specimen should be segmented from the document image to decide such similarity. Here one stage depends on the other one, moreover in complex situation it is paradoxical to seek a specimen that will compeer a member of the system's recognition set of alphabet symbols without including definite understanding of the structure of those symbols into process. The segmentation choice is not a local choice, it is unconstrained of previous and following decisions. Generating a good match to a library of symbol is mandatory, but not adequate for authentic identification. The accuracy of the current recognition or segmentation result can cast doubt if a poor match to the library exists. Say for example sequence a "d" closely resemble to the letter sequence "cl", but such an alternative will not add up to a contextually justifiable result. Therefore, the segmentation choice is interconnected with local choices while considering shape similarity, and with global choices while considering contextual acceptability.

Later, well known OCR techniques came which was a faster and much more powerful electronic circuitry in segmenting complex documents. The problems of segmentation persist today. A high percentage of errors to segmentation was consistently assigned by the well famed tests of commercial printed text OCR system from Las Vegas [1] and University in Nevada. The drawbacks of current machine print system for recognition when segmentation hassle increases is clearly illustrated in article [2]. Many authors formerly surveyed segmentation as part of a more extensive work, e.g., cursive recognition [3] [4], or document analysis [5] [6].

II. METHODS

Below figure shows Document segmentation levels:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

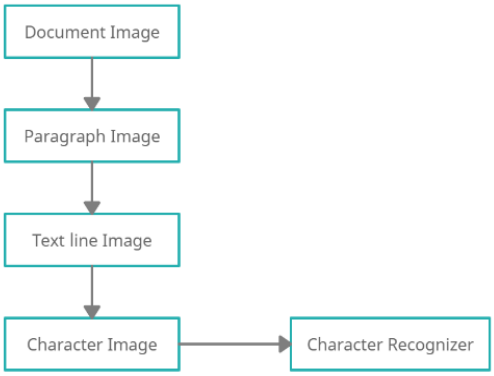

Fig. 1 Document Segmentation Levels.

First a document image is segmented to find the region of interest. After segmenting out the ROI, the region is segmented down to text line images. Then text line images are taken for character segmentation, where each character is separated out. Later these separate character images are given to a character recognizer for recognition purpose.

Mainly there are there fundamental approach for segmentation and countless Hybrid Approaches.

1. First one is classical approach, where segments are recognized based on "character-like" features. This kind of process where cutting up the image into meaningful part is given by a specific name, called "dissection,".

2. Second one is Recognition-based segmentation, where the system looks into the image for parts that match classes in its alphabet.

3. Third one is Holistic methods, where the system looks to identify words as a whole, thereby, avoiding the requirement of segmenting into characters.

Few segmentation techniques are discussed below. Decomposition of a picture into a succession of sub-images applying universal features is called as Dissection. In machine printing, successive characters are often separated by vertical white space. The action of detecting vertical white space between successive characters has an important space in dissecting hand print or machine print images. Reader IBM established in 1965 at the U. S. Social Security Administration [7] is one of the well set-down prompt commercial machines which engaged with pitch and white space. The device was capable of scanning alphanumeric data typed by the employers in the forms given quarterly to SSA. This process was made common by McCullough and Hoffman [8] to add more unique framework to it. In the articulation made by them the segmentation phase comprises with three steps: 1. Finding the starting of a character. 2. A judgement to test whether it is end of a character (sectioning). 3. Revelation of end-of-character.

Projection analysis talks about Vertical Projection or Vertical Histogram, which consist of simply the number of counts of black pixels in each column. It can be used to detect white space between successive letters. Locations of vertical strokes in hand and machine print docs can be specified by it. For the segmentation in non-cursive writings examination of the projection of a line of print was utilized as a bedrock. In paper [9] first the projection was taken, then the ratio of height to the second derivative of the curve was taken as a benchmark for deciding separating columns. The ratio given to peak at minimal of the projection, and circumvent the issue of division at points across horizontal thin lines. To ameliorate this method a peak to valley function was formulated in [10]. Then a minimal of projection is detected and projection value is taken. After that a sum of the differences between the obtained peaks on each side and the minimal value is calculated. The differentiator used to choose the segmentation boundary is this ratio of the sum to minimal value. The ratio shows fondness for small valleys which have large peaks on both sides.

In [11], a comparison between implicit and explicit based segmentation techniques is represented for off-line cursive handwriting recognition. Segmentation and Recognition are achieved at the same time in Implicit segmentation, where class overlapping problem is simply removed. The computational complexity of explicit segmentation-based approach is more the implicit based one, yet former achieved better results.

Xiu et al. proposed a new model for probabilistic segmentation system [12] where a contour primarily based on over segmentation approach is put forth. Slicing is applied to words in the document to obtain graphemes. Three queues are used to store graphemes; first one is for character, second for major components and last one for sub-components which contain diacritics. Taking geometric features, logical constraint and recognizer output into attention, assurance is calculated for each letter by using probabilistic model. To discover the optimal cutting direction, taking the letter weighted average as objective feature, the general optimization is performed.

Over segmentation was discussed as the most common and initial segmentation method by Brodowska [13]. On the kind of alphabet, the character strings were built over, a particular technique should be selected for task. Mixed approach, Holistic approach and Classical approach were also discussed.

A novel binary segmentation with neural validation was described by Lee et al. [14]. This approach contains over segmentation ruled segmentation point generator with neural validation modules.

A technique on features fusion approach was proposed by Ali and Suresha [15] to enhance the correctness of Arabic handwritten characters for recognition and segmentation purpose. To reach better identification of characters automatic selection of relevant features issue has been controlled predominantly for multi-font in Arabic script. To implement fusion strategy several features such as transition features, feature set of directional chain code frequencies and Gabor filter response are used. Fusion result is passed as input into three classifiers, first sequential minimal optimization classifier, second linear classifier and linear discriminant analysis.

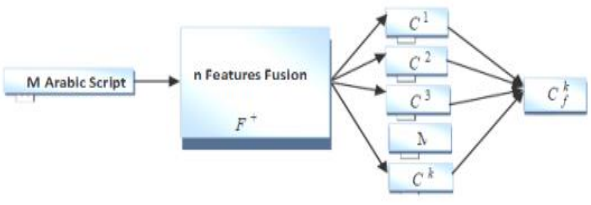

Fig. 2 Classifier and Feature level fusion.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Sari proposed an algorithm [16], where first the document image is smoothened to eliminate false pixels on the contours. Secondly word baseline is detected. Contour following does word decomposition, pseudo-words extraction, in secondary and primary parts of character. Picture decomposition is done on the higher, lower and median zones concerning to the baseline. All the existing processing one after one is performed on every pseudo-word. On the outer contour of the median zone, Local Minima (LM) locating is carried on. LMs filtering topological rules has been identified by Sari and Sellami [16].

Motawa et al. [17] adapted the Simon' algorithm [18] for handwritten Arabic character segmentation. First a preprocessing phase is executed, then text slant correction and binarization is done, after that extraction of secondary parts and pseudo-words identification is done.

The character recognition system and the various stages included in the process were discussed by Phukan and Borah [19]. Mainly the internal and external segmentation were also discussed. There are four broad categories for character recognition structural technique, template matching, neural networks and statistical technique.

The segmentation scheme for automated recognition of off line depraved cursive handwriting from static surfaces [20] was described by Choudhary. Here various research results of segmentation techniques for handwritten word were compared and also reviewed segmentation levels like implicit, explicit and holistic approaches.

In [21] Shi et al. discussed about an adaptive thresholding technique which is capable of exposing the text lines location. At first the grey level image generated by ALCM approach is binarized. After that each and every connected component in a text line are grouped into position masks. Finally, the lines of text are segmented by constructing the components with text line pattern mask. The contour is separated into pieces for touching characters and then segmentation is done in accordance to the distance and then reallocated based on the nearest line of text from center of mass.

Gaussian filter method has been used in Stamatopoulos et al. [22]. At first the document image is segmented into blocks then these blocks are binarized by an adaptive threshold where skew angle of each block is evaluated. Then, the blocks are linked to fetch the text lines path. At last thinning is utilized for the image path of background for text lines segmentation.

Hough transform approach has been utilized by Louloudis et al. in [23] for spotting straight lines in text images. Applying the rule of Hough transform to the gravity centre, the skew line orientation of text in Arabic handwritten is obtained for every document image. This technique incorporates exposed lines in handwritten images and it may include annotations among lines oriented in various directions, main lines and erasures.

In [24] Kumar et al. presents a novel graph depended segmenting lines approach of handwritten text document images of monochromatic Arabic script. Technique consist of coarse text line rating, using the main components that realize the task and line of diacritic part. To outline a graph with sparse similarity, every main component calculates an estimated local orientation. To calculate alikeness among non-neighbouring components, a shortest path algorithm was worn.

Images with various interference can be better handled by Deep-learning-based character segmentation methods. FCN was firstly proposed to for semantic segmentation to split the input image into different semantically interpretable categories. FCN has efficient and easy to be expand architecture of encoder-decoder. Therefore, more FCN based approaches evolved, such as RefineNet [25], U-Net, Deeplab [26] etc. All these can segment the characters using subtle spacing in the images combining slight interference.

A weakly supervised precise segmentation system for historical document images is proposed in [27]. The system mainly consists of four stages, including Pre-processing, Boundary Box Segmentation (BBS), Incremental Weakly Supervised Learning and Recognition-guided Attention Boundary Box Segmentation (Rg-ABBS). Rg-ABBS algorithm significantly reduces time consumption by performing recognition-guided segmentation only on 'attention' area and achieves comparable performance in comparison to other traditional methods.

Author	Method	Quantitative measures(%)
A. Rehman[11]	Implicit Based Segmentation Explicit Based Segmentation	Acc=79.23 Acc=80.91
H. Lee[14]	Binary Segmentation with Neural Validation	Acc=61.4
Zecheng Xie[27]	Recognition-guided ABBS	Rec=78.80 Prec=77.04 F-score=77.91
Liang-Chieh Chen[26]	Semantic Image Segmentation With Deep Learning(Deeplab)	Acc=79.7%
Xiu P[12]	Probabilistic Segmentation Model	Acc=59.2
Ali AAA[15]	Feature Fusion Approach	Multi-font Dataset – AHDB-Acc=97.30 AHCD-Acc=98.95
Louloudis G[23]	Block- Based Hough Transform Mapping	Acc=96.87
A. Choudhary[20]	Vertical Segmentation Approach	Acc=83.5

TABLE 1. SUMMARY OF DISCUSSED MODELS

III. CONCLUSION

Segmentation is an important stage in document digitization. The complete process of digitizing depends mainly on the output from segmentation. Character segmentation is the primary step in Recognition system. By the detailed analysis of the literature, it is observed Deep Learning based methods produces better results compared to other traditional methods.

REFERENCES

[1] T. Nartker, "ISRI 1992 annual report," Univ. of Nevada, Las Vegas, 1992.
[2] M. Bokser, "Omnidocument Technologies," Proc. IEEE, vol. 80, Issue. 7, pp. 1,066-1,078, July 1992.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

[3] G. Lorette and Y. Lecourtier, "Is Recognition and Interpretation of Handwritten Text: A Scene Analysis Problem?" Pre-Proc. IWHR III,p. 184, Buffalo, N.Y., May 1993.
[4] G. Dimauro, S. Impedovo, and G. Pirlo, "From Character to Cur sive Script RecogAition: Future Trends in Scientific Research," Proc. 11th Int'l Conf. Pattern Recognition, vol. 2, p. 516, Aug. 1992.
[5] D.G. Elliman and I.T. Lancaster," A Review of Segmentation and Contextual Analysis Techniques for Text Recognition", Pattern Recognition, vol. 23, no. 3/4, pp. 337-346, 1990.
[6] H. Fujisawa, Y. Nakano and K. Kurino," Segmentation methods for character recognition: from segmentation to document structure analysis", Proceedings of the IEEE, vol. 80, no. 7 pp. 1079- 1092, July 1992.
[7] R.B. Hennis ,"The IBM 1975 Optical Page Reader: system design ",IBM Journ. of Res. & Dev., pp. 346-353, Sept. 1968.
[8] R.L. Hoffman and J.W. McCullough," Segmentation methods for recognition of machine-printed characters", IBM Journ. of Res. & Dev., pp. 153-65, March 1971.
[9] H.S. Baird, S. Kahan and T. Pavlidis," Components of an omnifont page reader", In the Proceedings of 8th International Conference on Pattern Recognition, Paris, pp. 344-348, 1986.
[10] Y. Lu, "On the segmentation of touching characters", International Conference on Document Analysis and Recognition, Tsukuba, Japan, pp.440-443, Oct. 1993.
[11] A. Rehman, D. Mohamad, and G. Sulong "Implicit Vs Explicit based script segmentation and recognition: A performance comparison on benchmark database," International Journal Open Problems Compt. Math., vol. 2, pp. 352-364, 2009.
[12] Xiu P, Peng L, Ding X, Wang H,"Offline handwritten Arabic character segmentation with probabilistic model", In: Document analysis systems, Springer; 2006. p. 402–12.
[13] M. Brodowska, "Oversegmentation methods for character segmentation in off-line cursive handwritten word recognition," Schedae Informaticae, vol. 20, pp. 44-65, 2012.
[14] H. Lee, and B. Verma, "Binary segmentation with neural validation for cursive handwriting recognition," In the Proceedings of International Joint Conference on Neural Networks, pp.1730-1735, 2009.
[15] Ali AAA, Suresha M," A new design based-fusion of features to recognize Arabic handwritten characters",Int J Eng Adv Technol (IJEAT), 8(5):2570–4, 2019.
[16] Sari T, Souici L, Sellami M, "Handwritten Arabic character segmentation and recognition system: ACSA-RECAM", In the Proceedings of IWFHR'02, Canada,pp. 452–457, 2002.
[17] Motawa D, Amin A, Sabourin R" Segmentation of Arabic cursive script", In the International conference on document analysis and recognition proceedings of ICDAR'97, vol. 2; pp. 525–628,1997.
[18] Simon JC, "Off-line cursive word recognition", In Proceedings of the 1992 IEEE Conference, 80(7):1150–61, 1992.
[19] A. Phukan, M. Borah, "A survey paper on character recognition focusing on offline character recognition," International Journalof Computer Engineering and Applications, vol. 6, pp. 51-60, 2014.
[20] A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition," International journal of Information and Computation Technology, vol. 4, pp. 559-564, 2014.
[21] Shi Z, Setlur S, Govindaraju V, "A steerable directional local profile technique for extraction of handwritten Arabic text lines". In the Proceedings of 10th International conference on document analysis and recognition ICDAR'09, IEEE, pp. 176–180, 2009.
[22] Stamatopoulos N, Gatos B, Louloudis G, Pal U, Alaei A, "Handwriting segmentation contest", In the Proceedings of 12th International conference on document analysis and recognition (ICDAR), IEEE, pp.1402–1406, 2013.
[23] Louloudis G, Gatos B, Pratikakis I, Halatsis K, "A block-based hough transform mapping for text line detection in handwritten documents", In the Proceeding of International workshop on frontiers in handwriting recognition, 2006.
[24] Kumar J, Abd-Almageed W, Kang L, Doermann D, "Handwritten Arabic text line segmentation using affinity propagation", In the Proceedings of the 9th IAPR international workshop on document analysis systems, ACM; pp 135–142, 2010.
[25] G. Lin, A. Milan, C. Shen, and I. Reid, "Re_neNet: Multi-path re_nement networks for high-resolution semantic segmentation," In the Proceedings of IEEE Conference on Computer Vision and Pattern Recognit. (CVPR), Honolulu, HI, USA, pp. 5168-5177, July 2017.
[26] Liang-Chieh Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834-848, Apr. 2018.
[27] Zecheng Xie, Yaoxiong Huang, Lianwen Jin , Yuliang Liu , Yuanzhi Zhu, Liangcai Gao, Xiaode Zhang, "Weakly Supervised Precise Segmentation For Historical Document Images", Neurocomputing,350:271-281, July 2019.