The Algorithms For Segmentation Of Text-Lines In Handwriting Images

Huo Liulei, Kamil Moydin, Abdusalam Dawut, Askar Hamdulla*
Institute of Information Science and Engineering, Xinjiang University Urumqi 830046, China
*corresponding author's email: askarhamdulla@sina.com

ABSTRACT—Text line segmentation from handwriting image is the basis of handwriting text image processing, and the accuracy of line segmentation plays a decisive role in handwriting identification, handwriting recognition, handwriting retrieval and other research fields. The accuracy of line segmentation may directly lead to the accuracy and efficiency of handwriting identification, character recognition and text retrieval. Because offline handwriting has lost the order of writing and other information, which makes it more difficult to segment the offline handwriting image. This paper mainly aims at the complexity of the segmentation problem caused by the diversity of off-line handwriting styles, such as tilt, adhesion, overlap and so on, and compares the related solutions in recent years. In the end, some problems in line segmentation research are put forward or omitted, which is more convenient for readers to understand the field.

Keywords—Offline, Handwritten scripts, Text line, Segmentation

Ⅰ. INTRODUCTION

Text line segmentation is the first step in processing text information, and then subsequent research such as words recognition or retrieval and even information extraction of historical documents. Chirography can be divided different form.

Compared with the printed text line, the distribution is very neat, so the projection method can be used to segment the text image by the projection method. Handwritten text is not as simple as printing text lines, handwritten fonts are more random, and text layout is not regular. The following figure is divided into the renderings obtained by using the projection method when the threshold is 30 and 75. The experimental results in this paper are shown in figure 1

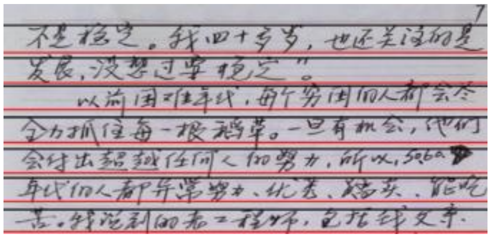

Figure 1: projection method with thresholds of 30 and 75

From the above image, it is possible to lose some of the smaller structures in the text even when the threshold it's set improperly, such as the point above the word '代'. This kind of loss will bring great obstacles to handwriting retrieval and have a small structure with many additional parts for Uyghur scripts. As a result, handwriting is generally not separated by projection alone.

This article mainly selects the articles from 2014 to 2018 to introduce them, and compares the advantages and disadvantages of their methods, which helps readers to more easily understand the advantages and disadvantages of each algorithm in recent years, and the progress of algorithm for row segmentation.

II. CLASSIFICATION OF TEXT LINE SEGMENTATION

In 1982, the first RLSA algorithm (RLSA)[1] was proposed by K.Y. Wong, R.G. Casey, F.M. Wahl et al. [1]. At present, text line segmentation or extraction methods for handwritten text images are mainly divided into the following three types: bottom-up, top-down, Hybrid.

A. Related bottom-up algorithms

The Bottom-up text image is segmented by pixels, pixel block (font), and text line. Such methods mainly include spectral clustering [2], feature corner aggregation [3], smearing effect [4], Mumford-Shah model [5], minimum spanning tree clustering [6], convolutional neural network [7], Markov decision process [8] and so on.

Ayman Al-Dmour and Fares Fraij use the already well-developed horizontal projection method to perform handwritten Arabic text segmentation [9], which is better for handwritten text images with well-written and large line spacing. The operation is simple and the running time is relatively short. Yi Xiaofang et al. proposed a Uyghur handwritten text image segmentation based on connected domains [10]. This method firstly divides the connected domains into three categories according to the size of the connected domain, and application an adaptive smear algorithm and deal with inflation. In this case, the text line skeleton has been basically formed. The area of the third type of connected domain is detected as a sticky character, and then processed.

Alireza Alaei et al. proposed an unconstrained handwritten text line segmentation method [11], which divides the text image into different vertical parts according to the line spacing after the line spacing obtained by the line spacing of the statistical text line. The text image is applied based on the average width smear. After the smear, the smaller black frame is removed. Then use the grayscale smear algorithm to repeatedly fill and remove the small boxes. Then the erosion operation is performed. The candidate segmentation line can be obtained by using the background-based thin algorithm, and then the overlapping and touching portions are processed. For the touching part, the line is divided at the smallest point of the pixel. For the overlapping part, the position of the overlapping part is first found according to the intersection point, and then divided into the belonging line according to the number of the contour point. Yi Xiaofang et al. proposed a text line segmentation method for foreground smearing and background thin [12]. The algorithm can have a good effect on the over-loss problem of Uyghur strokes containing a large number of discrete

978-1-7281-1366-1/18/$31.00 ©2018 IEEE
DOI 10.1109/ICSCSE.2018.00198

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on October 30,2020 at 16:29:26 UTC from IEEE Xplore. Restrictions apply.

stroke points and additional parts. Perform segmental smearing on the foreground part, and delete the expansion area where the aspect ratio does not satisfy the condition, obtain the positioning of the text area, and use the thin of the image background to obtain the text line dividing line, and use the center of gravity determination algorithm to solve the overlap problems.

L. Liu, Y. Lu, C.Y. Suen et al. proposed a near-repetitive document image matching method characterized by graphical perspective [13]. To deal with the instability of object segmentation, a multi-granularity object tree is constructed for a document image. Different levels are characterized by various object granularities. Two graphs with the maximum similarity are found from the multi-granularity object trees of the two near-duplicate document images which are to be matched. Jewoong Ryu et al. proposed a handwritten text line extraction algorithm in independent language [14], which can extract a series of different language and writing style text lines. By introducing the concept of stroke length， the different connected components of the super-pixel segmentation of the text picture and the text picture is first divided into different parts according to the super pixel and the connected component, Then line-partitioned based on the connected component. Finally, state estimation and cost function extraction of the line of text. Xi Zhang et al. propose text line segmentation for handwritten documents using constrained seam carving [15]. First, the larger part (there is adhesion between the two lines) detects the sticking portion and horizontally projects the portion and removes the black pixel smaller than the threshold portion by setting the threshold of the horizontal projection. After calculating the energy of each position of the whole graph using the energy function, the energy is counted, and normalized after statistics. It is normalized after statistics. Limits the flow direction of energy functions so that energy can be passed primarily to adjacent points in the same line of text and can avoid crossing different lines of text. Askar Hamdulla et al. proposed Uyghur handwritten text line segmentation based on shading [16]. The flow chart of the method is shown in figure 2.

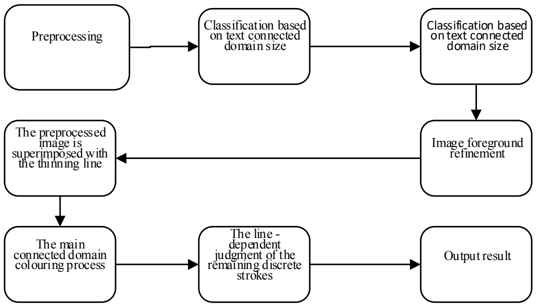
Preprocessing
Classification based on text connected domain size
Classification based on text connected domain size
Image foreground refinement
The preprocessed image is superimposed with the thinning line
The main connected domain colouring process
The line - dependent judgment of the remaining discrete strokes
Output result

Figure 2: [16] algorithm flow chart

Cao Wei and Xue Yuyang proposed an off-line handwritten Uyghur text line segmentation algorithm based on curve fitting and object selection [17] to solve the problem of inaccurate segmentation caused by oblique lines, alienated handwriting, etc. The method is based on the curve fitting and object selection of the handwritten Uygur text line segmentation. The projection segmentation method combined with the curve fitting and object selection processing methods. Zhu Zongxiao and Yang Bing proposed the completion of handwritten text line segmentation based on feature discrete point calculation [18]. A feedback method for segmentation line projection is proposed, which is divided into four parts: feature discrete point selection, feature discrete point sampling and optimization, feature discrete point grouping and feedback, and line edge optimization. Lei Xin et al. proposed a handwritten text line segmentation based on pooling operations [19]. Firstly, the pre-processed text image is pooled, and then the parallel search algorithm is used to obtain a connected region for each line. The attribution of the isolated region of above and below each row is adjusted, and finally the multi-line text image is divided into a single row.

B. Related top-down algorithms

Top-down: The text image is divided into text blocks, text paragraphs, and text lines in sequential order. Such methods mainly include piecewise projection analysis [20], run-length smoothing [21], adaptive local connection graph [22], Seam Carving [23], and so on.

Quang Nhat and GueeSang Lee proposed handwritten text line segmentation based on density prediction [24], predicting the structure of text lines through a trained convolutional network (FCN). By using the FCN network, you can estimate the line skeleton map and then extract the text lines. For the adhesion part, the adhesion part is decomposed into several small parts, and the length of the decomposition part to the other part is defined as [0.5, 2],The adhesion problem can be solved by distance clustering. Zhang xin etal. proposed the location and segmentation of the handwritten Uygur text line character adhesion zone[25]. Based on the features of the connected domain, the text image is fused with the positioning line. The text of the inter-line sticky characters is in the same connected domain, and the line of the attached text can be automatically extracted, Then the width and height of the attached characters are larger than the non-adhesive characters and automatically extracting the sticky characters. The extracted adhesion characters can be determined by the positioning line, and the site of the adhesion point is statistically analyzed, and then a thin line with the same color as the background is added to achieve the segmentation effect. Finally the segmented adhesion text line is passed. The coloring method is extracted line by line.

C. Related hybrid algorithms

The Hybrid method is a combination of top-down and bottom-up methods to achieve better results. For example, the literature [26] first roughly estimates the text line, followed by a series of corrected steps to ensure the correct segmentation of the text line. Literature [27] the character connected domain is decomposed, normalized, and the state of the sub-connected domain is estimated.

Liuan Wang et al. proposed a method based on the combination of connected regions and CNN to implement line extraction of text images [28]. First, the candidate

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on October 30,2020 at 16:29:26 UTC from IEEE Xplore. Restrictions apply.

connected components are extracted. Document images using the Maximum Stable Extremum Region (MSER) utilize Adaboost and convolutional neural filtering for noise networks (CNN). Then, generate a rough line of text. Localized linear layered edge reconstruction and cutting of text lines in the document spanning tree. Finally, for accurate text line extraction, cutting multiple components reconnects text line-based text line consistent minimized energy and fit errors. This method has a good effect on both horizontal and vertical lines. Bastien Moysset et al. proposed a new optical model based on text recognition [29], a recurrent neural network. Since these models are sequential, they are a column of text lines in the application. One of the main advantages of the method is that other data-driven methods are written for training. Datasets do not need to mark row boundaries, and only the number of rows required for each paragraph. Experimental results show that this method has similar or better results than traditional manual methods, and the method does not require too many parameters. But this method requires that the text distribution must be a horizontal straight line.Zhu Jianfei et al. proposed the handwritten text line extraction under the regression-cluster joint framework [30].

Samia Snoussi Maddouri et al. proposed a two-method mixed handwritten Arabic line segmentation method [31], which does not solve the problem of adhesion and overlap and is not very effective when the distance between two lines of text is relatively close. However, this method has a certain effect on the extraction of handwritten Arabic characters. The combination of the adapted morphological method and the peripheral equivalent coverage method and the combination of the Hough change and the morphological method are used to deal with the segmentation. Yin Yalin et al. proposed offline handwritten text line segmentation based on high-order correlation clustering [32]. This algorithm can well solve the step of handwritten text line sticking and bending. Firstly, a connected graph is used to form a document hypergraph. Then, under the constraint of the similarity measure obtained by the learning, the connected component pairs are marked as belonging to or not belonging to the same text line by a high-order correlation clustering algorithm; finally, union- find algorithm joins connected components into different lines of text.

D. Comparative analysis of these methods

This paper summarizes the algorithms described in this article are shown in table 1.

Table 1：Algorithm contrast

literature	conglutination	overlap	slant	label	train
【9】	No	No	No	No	No
【10】	Yes	Yes	Yes	No	No
【11】	Yes	Yes	Yes	No	No
【12】	No	No	Yes	No	No
【13】	No	No	Yes	No	No
【14】	No	No	No	No	No
【15】	Yes	No	No	No	No
【16】	No	No	Yes	No	No
【19】	No	No	Yes	No	No
【17】	No	No	Yes	No	No
【18】	Yes	Yes	No	No	No
【24】	Yes	Yes	Yes	No	Yes
【25】	Yes	Yes	No	No	No
【28】	No	No	No	No	Yes
【29】	No	No	No	Yes	Yes
【30】	Yes	Yes	Yes	Yes	No
【31】	No	No	Yes	No	No
【32】	Yes	Yes	Yes	No	Yes

III. CONCLUSIONS

In this paper we have discussed the text line segmentation algorithm from 2014 to 2018, and the article [33] studied the line extraction algorithms from 10 years only about the English, German, French, Arabic, Chinese. However, many other language’s texts have not been studied yet. There is some possibility to improve those algorithms in order to meet some special requirements, especially on three themes that the versatility, stability and cross-technical comparisons.

ACKNOWLEDGEMENTS

This work has been supported by the National Natural Science Foundation of China (under grant of 61462080 and 61662076).

REFERENCES

[1] K.Y. Wong, R.G. Casey, F.M. Wahl, Document analysis system, IBM J. Res. Dev.26 (6) (1982) 647

[2] Han X C，Yao H，Zhong G Q． Handwritten text line segmentation by spectral clustering［C］/ /Proceedings of the SPIE 10225，8th International Conference on Graphic and Image Processing．Tokyo，Japan: SPIE，2017: #102251A． ［DOI: 10． 1117/12． 2266982］

[3] Yadav V，Ragot N． Text extraction in document images: highlight on using corner points ［C］ / /Proceedings of the 12th IAPR Workshop on Document Analysis Systems． Santorini，Greece:IEEE，2016: 281-286.［DOI: 10. 1109/DAS．2016．67］

[4] Bukhari S S，Shafait F，Breuel T M． Text-line extraction using a convolution of isotropic Gaussian filter with a set of line filters ［C］ / /Proceedings of the 11th International Conference on Document Analysis and Recognition． Beijing，China: IEEE，2011:579-583． ［DOI: 10． 1109/ICDAR． 2011． 122］

[5] Du X J，Pan W M，Bui T D． Text line segmentation in handwritten documents using Mumford-Shah model［J］． Pattern Recognition，2009，42(12): 3136-3145． ［DOI: 10． 1016/j． patcog．2008． 12． 021］

[6] Yin F，Liu C L． Handwritten Chinese text line segmentation by clustering with distance metric learning［J］． Pattern Recognition，2009，42(12): 3146-3157． ［DOI: 10． 1016/j． patcog． 2008． 12． 013］

[7] Vo Q N，Lee G． Dense prediction for text line segmentation in handwritten document images ［C］ / /Proceedings of 2016 IEEE International Conference on Image Processing． Phoenix，Arizona，USA: IEEE，2016: 3264-3268． ［DOI: 10． 1109/ICIP． 2016．7532963］

[8] Boulid Y，Souhar A，Elkettani M Y． Detection of text lines of handwritten Arabic manuscripts using Markov decision processes ［J］． International Journal of Interactive Multimedia and Artificial Intelligence，2016，4 (1): 31-36． ［DOI:

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on October 30,2020 at 16:29:26 UTC from IEEE Xplore. Restrictions apply.

10. 9781/ijimai. 2016. 416］

[9] Al-Dmour A, Fraij F. Segmenting Arabic Handwritten Documents into Text lines and Words[J]. International Journal of Advancements in Computing Technology, 2014, 6(3):109-119.

[10] Yi Xiaofang,Ka Mi Li Mu Yiding, Escar Ai Mudu La. Uyghur Handwritten Text Line Segmentation Based on Connected Domain Features[J].Computer Engineering and Applications,2014,18:142-146.

[11] Alireza Alaei,Umapada Pal,P. Nagabhushan. A new scheme for unconstrained handwritten text-line segmentation[J]. Pattern Recognition,2011,44(4).

[12] Yi Xiaofang,Ka Mi Li Mu Yiding, Escar Ai Mudu La. Uyghur Handwritten Text Line Segmentation Based on Connected Domain Features[J].Computer Engineering and Applications,2014,18:142-146.

[13] L. Liu, Y. Lu, C.Y. Suen, Near-duplicate document image matching a graphical perspective, Pattern Recognit. 47 (4) (2014) 1653.

[14] J. Ryu, H.I. Koo, N.I. Cho, Language-independent text-line extraction algorithm for handwritten documents, Signal Process. Lett. 21 (9) (2014) 1115.

[15] Zhang X，Tan C L． Text line segmentation for handwritten documents using constrained seam carving［C］/ /Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition．Heraklion，Greece: IEEE，2014: 98-103．［DOI: 10．1109/ICFHR． 2014． 24］

[16] Eskar Amudu, Yi Xiaofang, Kamili Mu Yiding. Uyghur Handwritten Text Line Segmentation Based on Coloring Processing[J].Journal of Tsinghua University(Science and Technology),201302:259-264.

[17] Cao Wei,Xue Yuyang.Offline handwritten Uyghur text line segmentation algorithm based on curve fitting and object selection[J]. Computer and Digital Engineering,2015,43(08): 1375-1377+1439.

[18] Zhu Zongxiao,Yang Bing.Application of feature discrete point calculation in handwritten text line segmentation[J]. Computer Engineering and Applications, 2015, 51(08):148-152+204.

[19] Lei Xin,Li Junyang,Song Yu,Sai Linwei.Text segmentation method for handwritten Chinese character recognition[J]. Intelligent computers and applications,2018,8 (02):126-128.

[20] Arivazhagan M，Srinivasan H，Srihari S． A statistical approach to line segmentation in handwritten documents［C］/ /Proceedings of the SPIE 6500， Document Recognition and Retrieval XIV． San Jose，CA: SPIE，2007: # 65000T． ［DOI: 10． 1117/12． 704538］

[21] Nikolaou N，Makridis M，Gatos B，et al． Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths［J］． Image and Vision Computing，2010，28(4): 590-604． ［DOI: 10． 1016/j． imavis． 2009． 09． 013］

[22] Shi Z X，Setlur S，Govindaraju V，et al． A steerable directional local profile technique for extraction of handwritten Arabic text lines［C］/ /Proceedings of the 10th International Conference onDocument Analysis and Recognition．Barcelona，Spain: IEEE，2009: 176-180． ［DOI: 10． 1109/ICDAR． 2009． 79］

[23] Zhang X，Tan C L． Text line segmentation for handwritten documents using constrained seam carving［C］/ /Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition．Heraklion，Greece: IEEE，2014: 98-103．［DOI: 10．1109/ICFHR． 2014． 24］

[24] Vo Q N，Lee G． Dense prediction for text line segmentation in handwritten document images［C］/ /Proceedings of 2016 IEEE International Conference on Image Processing． Phoenix， Arizona， USA: IEEE，2016: 3264-3268． ［DOI: 10． 1109/ICIP． 2016． 7532963］

[25] Zhang Xin, Eskar Amudu, Camille · Mu Yiding. Offline handwritten Uyghur text line character adhesion area positioning and segmentation [J]. Laser Magazine,2014,35(11):4-10.

[26] Cohen R，Dinstein I，El-Sana J，et al． Using scale-space anisotropic smoothing for text line extraction in historical documents［C］/ /Proceedings of the 11th International Conference on Image Analysis and Recognition．Cham: Springer，2014: 349-358．［DOI: 10． 1007/978-3-319-11758-4_38］

[27] Ryu J，Koo H I，Cho N I． Language-independent text-line extraction algorithm for handwritten documents［J］． IEEE SignalProcessing Letters，2014，21 (9): 1115-1119． ［DOI: 10． 1109/LSP． 2014． 2325940］

[28] L. Wang, W. Fan, J. Sun, S. Naoi, Text line extraction in document images, in:Proceedings of 13th ICDAR, IEEE, 2015, p. 191.

[29] B. Moysset, C. Kermorvant, C. Wolf, Paragraph text segmentation into lines with recurrent neural networks, in: Proceedings of 13th ICDAR, IEEE, 2015, p. 456.

[30] Zhu Jianfei, Ying Zao, Chen Pengfei.Regression-handwritten text line extraction under the clustering framework[J].Journal of Image and Graphics,2018,23(08):
1207-1217.

[31] Maddouri S S, Ghazouani F, Samoud F B. Text lines and PAWs segmentation of handwritten Arabic document by two hybrid methods[C]// International Conference on Advanced Technologies for Signal and Image Processing. 2014:310-315.

[32] YinYalin,LiuAimin,Zhou Xiangdong.Offline handwritten text line segmentation based on high-order correlation clustering[J].Journal of Central China Normal University (Natural Science) ,2017,51(01):18-22+34.

[33] Sébastien Eskenazi,Petra Gomez-Krämer,Jean-Marc Ogier. A comprehensive survey of mostly textual document segmentation algorithms since 2008[J]. Pattern Recognition,2017,64.

Authorized licensed use limited to: Polytechnic University of Bucharest. Downloaded on October 30,2020 at 16:29:26 UTC from IEEE Xplore. Restrictions apply.