Text line segmentation based on local baselines and connected component centroids for Tibetan historical documents

Pengfei Hu¹, Yang Chen¹, Yusheng Hao¹,², Yiqun Wang³ and Weilan Wang³ *

¹ School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, China.
² Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Lanzhou, China.
³ Corresponding author’s e-mail: wangweilan@xbmu.edu.cn

Abstract. Text line segmentation is one of the key contents in document image analysis and recognition. Due to the complex situation of text images in Tibetan historical document, such as the coexistence of slanting and distortion of text lines, especially the adhesion between lines, text line segmentation has become a challenging task. In this paper, a new method of text segmentation for Tibetan historical document is proposed. This method first obtains the local baseline information of the text line, and then performs the detection and segmentation of the adhesion area. Finally, according to the barycentre coordinates of the connected component and the local baseline information, each connected component is assigned to the corresponding text line. The method avoids the effect of text line distortion on the segmentation of the stuck text line and greatly reduces the error rate of the adhesion text line segmentation and can effectively deal with the segmentation of slanted and distorted Tibetan sticky text lines.

1. Introduction
Ancient Tibetan books are rare and precious historical materials for the study of Tibetan politics, history, culture, and medicine, which have important academic value and research significance. Due to the decline in the quality of ancient Tibetan books, it is urgent to use digital technology to protect and reuse it. Currently, using OCR technology to convert historical documents into text is a better solution. However, the processing of ancient Tibetan books faces many challenges, such as binarization, layout analysis, text region extraction, text line segmentation, text segmentation and text recognition. Tibetan ancient books have a long history, and there are many degenerations in the process of preservation (e.g. document damage, broken stroke, and torn pages, etc.), in ancient books, there is not the only distortion, slanting and conglutination of text lines, but also broken of character regions. These phenomena make the line segmentation of ancient Tibetan books particularly difficult.

For text line segmentation of historical documents, there are mainly the following methods: projection-based methods[1-3], clustering approach[4-6], neural network[7-9], based on the center of gravity of connected components[10-11], based on baseline detection method[12] and based on connected components analysis[13].

Projection-based method projects the image horizontally and adds the foreground image projection value to the horizontal direction of each line, text lines are split by appropriate projection positions. This method has a good cutting effect on images with obvious cutting lines and is not suitable for Tibetan ancient text images with text line distortion and tilt.
The clustering approach first detects the text area and then performs cluster analysis based on the density or distance of the text area to aggregate the text areas. However, this method cannot effectively solve the problem of interline character adhesion in Tibetan ancient books.

In the method of neural network, the researchers used a neural network to detect text and connect the same text lines. This method has achieved good results for handwritten documents, but it does not consider the situation of text fracture. Moreover, such types of networks require large annotated datasets.

In the method based on the center of gravity of the connected components, the position of the center of gravity of the connected component is calculated and connected with the region in the vicinity. This method does not solve the problem of text line sticking.

In the method based on baseline detection and connected components analysis, common methods for processing Tibetan ancient books, the baseline is obtained by using module matching, projection, etc. The connectivity component analysis is first used for adhesion detection, and then the connectivity component assignment is completed based on the baseline information. However, broken strokes and text line distortions are not considered.

This paper focuses on the line segmentation of ancient Tibetan books, the proposed method combines local baselines, detection of adhesion areas and their segmentation, the judgment of connected components gravity center and its attribution, etc. The method avoids the influence of text line distortion on the segmentation of the stuck text line and greatly reduces the error rate of the adhesion line segmentation.

The paper is organized as follows: in Section 2, introduces the algorithm in detail, Section 3 presents the experimental results and discussions, Section 4 describes conclusions and prospection.

2. The material and methodology

The Tibetan script is a kind of spelling language, which consists of 30 consonants and 5 vowels. A single character is composed of the vertical superposition of its spelling rules. Uchen Script is a commonly used Tibetan font style, which is used in most extant Tibetan ancient books (including handwriting and engraving). In the writing process, the prominent feature is that the head of all characters is arranged along a horizontal line, as is shown in the red line in Figure 1. However, due to writing reasons, in ancient books of Uchen, there would be such phenomena as slanting of ancient texts (as shown in the blue line in Figure 1), touching characters (as shown by the green rectangle in Figure 1) and characters broken (as shown in the green ellipse in Figure 1), which make text line segmentation more difficult.

Based on the characteristics of Uchen Script baseline, line distortion of ancient texts, and character adhesion between upper and lower lines, we use a combination of local baseline detection, adhesion region detection and segmentation, connected components centroid calculation and attribution judgment to conduct text line segmentation.

Our method is as follows, mainly including four steps:

Step1: First, estimating the text line height through the detection information of the connected component, then further determine the area of the baseline through the straight line detection, horizontal projection and the number of the horizontal connected component, finally divide the image into 8 pieces in the horizontal direction and use the projection information to find the local baseline.

Step2: First, the connected components were detected, then the adhesion region was judged according to the local baseline information, and finally the adhesion region was segmented.
Step3: Calculate the barycentre coordinates of the connected component.
Step4: Assign the connected components according to the local baseline position, and then complete the text line segmentation.

2.1. Local baseline detection

2.1.1. Estimating the average height of text lines. The connected component information can more objectively reflect the text information, and all the connected components in the image can be obtained through the connected component detection, so as to estimate the average height of the text in the image. There is a lot of noise and component fracture in the document images of ancient Tibetan books, therefore, setting the connected component’s acreage threshold to filter these conditions reduces the influence on the connected component height calculation, take the average value of connected component’s acreage as connected component’s acreage threshold, at the same time, it is easy to calculate the total number of starting positions of the connected components in each horizontal direction and record the result as $\text{conNumList}$. The detection results of the connected components are shown in the red rectangle in Figure 2. Calculate the average height of the connected component according to the height of the connected component of the image as $\text{aveHeight}$, the $\text{aveHeight}$ can be calculated by equation (1):

$$\text{aveHeight} = \frac{\sum_{i=1}^{n} \text{conHeight}}{n}$$

(1)

Where $n$ represents the total number of connected component greater than the acreage threshold, and $\text{conHeight}$ represents the connected component’s height.

2.1.2. Horizontal projection. The horizontal foreground projection of the image refers to the accumulation of the number of pixels of the foreground in the same horizontal direction, project the image horizontally and record the projection result as $\text{projectionList}$. The blue histogram in Figure 2 represents the horizontal projection result.

2.1.3. Horizontal line detection. Horizontal line detection is usually used to find straight line segments in an image, and the baseline happens to consist of horizontal lines for each character. The horizontal straight line detection was carried out on the image, and the total length of the straight line segment in each horizontal direction was recorded as $\text{lineLengthList}$. The linear detection results are shown in Figure 3, the lines in red represent the lines that form the baseline, and the lines in green represent the noise line.
2.1.4. **Determining the text line baseline area.** So far, `conNumList`, `aveHeight`, `projectionList` and `lineLengthList` are obtained in 2.1.1, 2.1.2 and 2.1.3, respectively representing the total number of the starting coordinates of the connected components in each horizontal direction, the average row height, the projection list and the total length of the straight line segment in each horizontal direction. The baseline area has more text information, it should be in the area with `projectionList`, `conNumList` and `lineLengthList` biggish value, so it's a good idea to set a threshold to minimize the area of baseline.

Due to the distortion and inclination of the text line, the value of the horizontal projection in the text line will fluctuate greatly (In Figure 2, the fluctuation amplitude of the projection peak in the first line and the projection peak in the second to last line can be clearly observed). In this paper, the threshold of horizontal projection is simply selected as the median in `projectionList`, and the value greater than the median in `projectionList` is considered to be the area of the baseline.

All text line characters are horizontally aligned along the character baseline, so the length of the horizontal line in the baseline area should be greater than the length of the horizontal line between characters. Similarly, the median in the `lineLengthList` is selected as the threshold, and the position larger than the threshold is considered as the baseline area. The baseline area detection results using `projectionList` and `lineLengthList` are shown in Figure 4.

**Figure 3.** Horizontal line detection diagram.

**Figure 4.** The baseline area detection results using `projectionList` and `lineLengthList`

Based on this, the `conNumList` and `aveHeight` information are used to further determine the baseline area information. A non-baseline area is considered if there is only one connected component in the same horizontal direction and there are multiple connected components in one horizontal direction in the upper 1/2 `aveHeight` or lower 1/2 `aveHeight` range. The results of further determining the baseline area using `aveHeight` and `conNumList` are shown in Figure 5.

**Figure 5.** The results of further determining the baseline area using `aveHeight` and `conNumList`.
The filtered baseline area becomes more sparse, so the exact location of the baseline can be determined by grouping. If the distance between adjacent baselines is greater than 1/2 $aveHeight$, the adjacent baselines are divided into upper and lower groups from the middle position. After grouping, if intra-group distance is greater than 2.5 times the $aveHeight$, the calculated reorganization is divided into $N$ equal parts. $N$ can be calculated by equation (2):

$$N = \left\lceil \frac{D}{aveHeight} \right\rceil$$  

(2)

Where $D$ represents intra-group distance. The grouped results indicate the area of the baseline, the grouping result is shown in Figure 6.

2.1.5. Determining the local baseline location. Characters are lined up along the baseline, so the baseline is the line with the largest horizontal projection of the foreground color in the baseline area. Selecting the position of the maximum value of the text line projection as the overall baseline of the text line, the overall baseline is shown by the red line in the Figure 7 (a). Due to the distortion and tilt of the text line, part of the overall baseline is offset from the actual situation, in order to avoid the influence of these situations, the image is projected locally to obtain local baseline information. In this paper, the image is evenly divided into 8 equal parts from the horizontal direction, and the maximum value of the projection value is selected as the local baseline. The local baseline results are shown in Figure 7 (b), where different color on the same text line represents a complete local baseline.

2.2. Adhesion area detection and segmentation

The local baseline information represents the location of the text line, if a connected component passes two local baselines, it means that the connected component has passed two text lines, so this area is judged as the adhesion area. Adhesion detection results are shown in Figure 7(b) the four red rectangles are the smallest external box of the interline adhesion area.
A interline adherent region is a connected component passing through adjacent text lines. Figure 8 is an enlarged view of the rectangular frame of the adhesion area of Figure 7(b), where the red line represents the local baseline position. Although the sectioned area is the smallest circumscribed rectangle of the adhesive area, there may be other text parts of the non-adhesive area in the rectangle. Our idea is to temporarily delete these "other text parts", and then use the horizontal projection method to find the place where the projection value is the smallest to make the adhesion area segmentation. The "other text portion" is as shown in the area of the blue border in Figure 9. First, the blue area is deleted, and then the horizontal direction is projected. The position with the smallest horizontal projection value is as shown by the yellow line in Figure 10, and the segmentation point is also found. After the segmentation is completed, the blue area is restored.

2.3. The calculation of barycentre coordinates
In the image, each broken stroke, vowel, or character is a connected component. We can use the image moment to calculate the barycentre coordinates of the connected component. The \((p + q)\)-order moment of the contour can be obtained by equation (3):

\[
m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} I(x,y) x^p y^q
\]

Among them, \(p\) and \(q\) represent the \(X\) latitude and \(Y\) latitude of the moment. \(I(x,y)\) represent a connected component, and \(M\) and \(N\) represent the maximum value of \(X\) latitude and \(Y\) latitude.

When \(p = 1, q = 0\), that represents the sum of the horizontal coordinates of the foreground color in the connected component \(m_{01}\).

When \(p = 0, q = 1\), that represents the sum of the vertical coordinates of the foreground color in the connected component \(m_{01}\).

When \(p = 0, q = 0\), that represents the number of foreground colors in the connected component \(m_{00}\).

The abscissa \(x_c\) and ordinate of the barycentre coordinate \(y_c\) can be calculated by equation (4):

\[
\begin{align*}
x_c &= \frac{m_{01}}{m_{00}} \\
y_c &= \frac{m_{00}}{m_{00}}
\end{align*}
\]

Barycentre coordinates of the connected component are shown by the red dot in Figure 11.

Figure 8. Adhesion area external frame. Figure 9. Adhesive area frame and other text parts (blue box). Figure 10. Sectioning line of adhesive area (yellow line).

Figure 11. Local baseline and the barycentre coordinates.
2.4. Connected component assignment

The connected component that passes through the baseline area is directly assigned to the text line. Since the Tibetan vowel height is about 1/4 of the height of Tibetan characters, so if the connected component barycentre coordinates to the baseline distance is less than a quarter of the $aveHeight$, we also assign the connected component to the row of the local baseline. And we perform the following calculations for the area that have not been assigned.

Firstly, looking for the two nearest baselines above and below the barycentre coordinates, since the local baseline information represents the text line information, the connected region should belong to the text line represented by the two local baselines. Secondly, the Euclidean distance of the barycentre coordinates and the barycentre coordinates that has been allocated in the upper text line is calculated, and the calculation is also performed on the lower text line. In the calculation process, noise is prevented from affecting the calculation, so the connected component of small area is filtered and does not participate in the calculation of the distance. Finally, the sum of the values of the distance in each line is taken as the distance between the connected component and the text line, and the connected component is allocated according to the minimum distance of the barycentre coordinate.

The local baseline and the barycentre coordinates are shown in Figure 11, the colored line segments represent the local baseline position, and the red dots represent barycentre coordinates. Text line segmentation is performed according to the result of the connected component allocation. The segmentation result is shown in Figure 12.

3. Experimental results and discussions

The experimental dataset is a selection of 212 images from more than 60 thousand images from the Beijing woodcut edition of the Tibetan scripture “Kangjur”. The dataset contains 1696 lines of text and more than 148,400 characters.

For the data set, we define the number of all text lines is $N$, the number of detected lines is $G$, the total number of characters is $C$, and the number of improperly divided areas is $S$. The line detection rate ($LD$) can be calculated by equation (5):

$$LD = G / N$$  \hspace{1cm} (5)

Text line segmentation is prepared for recognition. If there is a connected component error assignment, it will affect at least two characters, so the recognition accuracy ($RA$) can be calculated by equation (6):

$$RA = 1 - 2S / C$$  \hspace{1cm} (6)

Table 1 shows the comparison between the method presented in this paper and the contour curve tracking method and the connected component analysis method we have previously proposed. Compared to the contour curve tracking method and the connected component analysis method, the effect of the method in this paper is superior under each evaluation value. And this method has reached 99.25% of the adhesion detection rate of adjacent text lines. However, in the process of assigning connected domains, there are still some broken strokes assigned incorrectly (As shown in the red rectangle in Figure 12).
### Table 1. The algorithm comparison results.

| Method                              | N    | LD  | RA  |
|-------------------------------------|------|-----|-----|
| Contour Curve Tracking[11]          | 1696 | 82.79% | 80.95% |
| Connected Component Analysis[13]    | 1696 | 91.17% | 90.79% |
| Method of This Paper                | 1696 | 98.76% | 96.52% |

### 4. Conclusion and prospection

In this article, we proposed a text line segmentation method for historical Tibetan documents that combine the local baseline and barycentre coordinates of the connected component. Local baselines are robust to text line warping and skewing, then perform adhesion detection and segmentation on the text image, finally, the barycentre coordinates of the connected component are calculated to allocate the connected component and complete the text line segmentation. This algorithm has a good effect in detecting the complete concatenation area of text lines and can guarantee the integrity of the text area with low time complexity, however, the effect of the allocation of the broken area does not achieve the desired. The focus of future work is on the allocation of broken stroke regions and the study of the adhesion character segmentation algorithm.

### Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61772430), the Innovative Research Team of State Ethnic Affairs Commission ([2018]98), the Gansu Provincial First-Class Discipline Program of Northwest Minzu University (No.11080304), the innovation ability improvement project of colleges and universities in Gansu Province in 2019(No.2019B-024), the Fundamental Research Funds for the Central University of Northwest Minzu University (No.31920180050), and also supported by the special fund to support graduate student project of the Fundamental Research Funds for the Central University of Northwest Minzu University(No. Yxm2020097).

### References

[1] Jindal, P., & Jindal, B. (2015) Line and Word Segmentation of handwritten text documents written in Gurmukhi Script using midpoint detection technique. In: 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS). Chandigarh. pp. 1-6.

[2] Ahmad, I., Wang, X., Li, R., Ahmed, M., & Ullah, R. (2017) Line and Ligature Segmentation of Urdu Nastaleeq Text. IEEE Access, 5: 10924-10940.

[3] Ptak, R., Zygadlo, B., & Unold, O. (2017) Projection–Based Text Line Segmentation with a Variable Threshold. International Journal of Applied Mathematics and Computer Science, 27: 195-206.

[4] Garz, A., Fischer, A., Sablatnig, R., & Bunke, H. (2012) Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering. In: 2012 10th IAPR International Workshop on Document Analysis Systems. Gold Coast, QLD. pp. 95-99.

[5] Garz, A., Fischer, A., Bunke, H., & Ingold, R. (2013) A Binarization-Free Clustering Approach to Segment Curved Text Lines in Historical Manuscripts. In: 2013 12th International Conference on Document Analysis and Recognition. Washington, DC. pp. 1290-1294.

[6] Calderon, M.Á., Garcia-Hernandez, R.A., & Ledeneva, Y. (2018) Unsupervised multi-language handwritten text line segmentation. Journal of Intelligent & Fuzzy Systems, 34: 1-11.

[7] Renton, G., Soullard, Y., Chatelain, C., Adam, S., Kermorvant, C., & Paquet, T. (2018) Fully convolutional network with dilated convolutions for handwritten text line segmentation. In:
International Journal on Document Analysis and Recognition (IJDAR). Berlin, Heidelberg. pp. 177-186.

[8] Vo, Q.N., Kim, S., Yang, H.J., & Lee, G. (2018) Text line segmentation using a fully convolutional network in handwritten document images. IET Image Processing, 12: 438-446.

[9] Mechi, O., Mehri, M., Ingold, R., & Amara, N.E. (2019) Text Line Segmentation in Historical Document Images Using an Adaptive U-Net Architecture. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). Sydney, Australia. pp. 369-374.

[10] Gupta, J.D., & Chanda, B. (2010) A Model Based Text Line Segmentation Method for Off-line Handwritten Documents. In: 2010 12th International Conference on Frontiers in Handwriting Recognition. Kolkata. pp. 125-129.

[11] Zhou, F., Wang, W., & Lin, Q. (2018) A Novel Text Line Segmentation Method Based on Contour Curve Tracking for Tibetan Historical Documents. International Journal of Pattern Recognition and Artificial Intelligence, 32 1854025:1-1854025:21.

[12] Li Y, Ma L, Duan L, et al. (2017) A Text-Line Segmentation Method for Historical Tibetan Documents Based on Baseline Detection. In: CCF Chinese Conference on Computer Vision. Singapore. pp. 356-367.

[13] Wang, Y., Wang, W., Li, Z., Han, Y., & Wang, X. (2018) Research on Text Line Segmentation of Historical Tibetan Documents Based on the Connected Component Analysis. In: PRCV 2018. Guangzhou, China. pp. 74-87.