Baseline detection based on local connection and evaluation for Tibetan historical document text line

Yiqun Wang[a], Weilan Wang[*a], Zhengqi Cai[b]
[a]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, Gansu, China; [b]School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, Gansu, China

ABSTRACT

The baseline position is an important reference information of the text line in the document written by Uchen Script. In order to overcome the distortion of text line and obtain accurate baseline position, this paper proposes a baseline detection method based on lines connection local and global evaluation. First, the slanted document image is corrected and the number of text lines is determined by projection. Then the discrete lines are connected locally; Finally, the baseline of text lines is obtained by evaluating the combination of baselines. Experimental results show that the proposed method can effectively overcome document skew and text line distortion, and the baseline position detected is accurate.

Keywords: Tibetan historical document, text line, baseline, local connection

1. INTRODUCTION

Text lines usually have obvious spacing between adjacent lines, characters of the same text line are arranged along the writing direction of the text line, and characters are close to each other. According to these characteristics, text line extraction algorithms can be divided into two categories[1]. The first category uses the feature of obvious space between adjacent text lines to extract text lines[2]. The other method takes advantage of the feature that characters belonging to the same text line are close to each other, and characters are extracted along the direction of the text line to achieve the purpose of text line extraction[3,4].

The Tibetan script is a kind of spelling language[5]. The spelling rules of the script are superimposed around the left and right sides of the base characters and in the vertical direction. Consonants and vowels form characters in the vertical direction (as shown in the vertical box in Figure 1). Syllables are composed of one or more characters, and are separated by syllabic points (as shown in the green part in Figure 1). The most notable feature of Uchen script is that the first stroke of each letter is horizontal, and all characters in the same text line are arranged along a horizontal baseline (as shown in the red line in Figure 1). However, due to the reasons of writing, manual printing and vertical superposition of characters, there are complex phenomena such as slanting, twisting and conglutination between adjacent text lines, which make the traditional method of text line segmentation unable to achieve satisfactory results. Therefore, baseline becomes an important information for text line extraction in Tibetan historical documents of Uchen script.

Researchers have proposed different methods for baseline detection of text lines in Tibetan historical documents. Li et al.[6] proposed a baseline detection method based on template matching, pruning algorithms and closing operation. Wang et al.[7] use the projection method to detect the baseline position of text lines, but projection-based method is not suitable for distorted text lines or tilted document. In order to overcome the distortion of text lines and detect the accurate baseline of text lines, Hu et al.[8,9] proposed local projection method to detect the baseline position, which uses multiple straight lines to approach the baseline of distorted text lines to reduce detection errors. Li et al.[10,11] proposed a baseline detection method based on local connection. The method extracted the upper edge of characters and connected the upper edge in the local range. The line with the longest length was selected as the baseline of text lines. This method is more accurate for wavy text line baseline detection, but less accurate for text line baseline detection with long distance interval.

This paper focuses on baseline detection of historical Tibetan text lines. In this paper, the baseline detection method proposed by Li et al. is improved. The baseline detection of text lines is carried out through tilt correction, determination

* wangweilan@xbmu.edu.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066E
 · 0277-786X · doi: 10.1117/12.2662891

of the number of text lines, local connection of upper edges, line combination and evaluation. This method can accurately detect the baseline of text lines and provide an accurate basis for text line extraction.

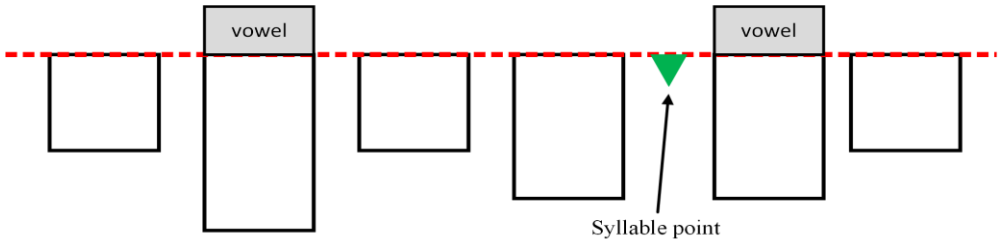


Figure 1. Position relationship between Uchen script and baseline.

2. METHODOLOGY

This paper proposes a baseline detection method for Tibetan historical document text line based on local connection and evaluation. The method consists of three steps. The first step is preprocessing, including document image skew correction and the number of text line determination. The second step is character upper edge detection and local connection. The upper edge of Uchen script is detected, and the upper edges is connected locally to obtain the long baseline. The third step is combination and evaluation. Baselines are grouped according to horizontal coordinate positions, and the baselines within each group are arranged. The combination of baselines with the highest score is selected through evaluation of the combination of baselines, and the baselines in the combination are successively connected to obtain the baseline of text lines. Figure 2 shows the framework of the proposed method.

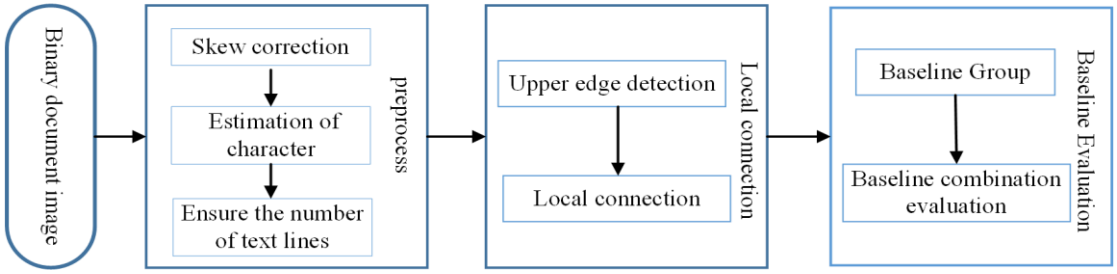


Figure 2. The framework of the proposed method.

2.1 Preprocessing

Preprocessing has two functions. The first function is to correct the tilted document, and the second function is to determine the number of text lines in the document image. Only knowing the exact number of text lines can ensure that all text lines are accurately extracted, otherwise, some text lines will be missed or segmentation errors will occur.

2.1.1 Skew Correction. The text lines of Tibetan historical documents are arranged horizontally and distributed in parallel, so the standard variance of the one-dimensional array obtained by horizontal projection is certain to be the largest. On the contrary, when the document image is tilted, the standard variance of the array obtained by horizontal projection will decrease. Based on the above analysis, the calculation of the document tilt Angle is equivalent to determining the maximum variance of the document projection at different rotation angles.

Generally, the tilt angle of Tibetan historical document images is small. Thus, the oblique angle of the document image is determined in the range of minus 5 degrees to plus 5 degrees by two decimal places. Rotate the document image by different angles, and then calculate the variance of the one-dimensional array obtained by horizontal projection. The angle corresponding to the maximum variance is the tilt correction angle of the document image.

2.1.2 Estimation of Character. Stroke width (SW) is an important parameter in document analysis. Calculation of stroke width is shown in equation (1):

$$SW = \sum_{x=1}^{W}\sum_{y=1}^{H} I(x,y) \Big/ \sum_{x=1}^{W}\sum_{y=1}^{H} I\_SK(x,y) \tag{1}$$

where, W and H represent the width and height of the image, I is the document image, and I_SK is the foreground skeleton of the document image.

The width of the Character also needs to be estimated. The calculation formula of Average Character width (ACW) is shown in equation (2):

$$ACW = \frac{1}{n}\sum_{i=1}^{n} CCW_i \quad if\ CCW_i > 2 \times SW \tag{2}$$

where, CCW_i is the width of the ith connected component in the document image, n is the number of connected components whose width is greater than 2 times SW.

2.1.3 The Number of Text Lines. The number of text lines in the document image is determined by foreground horizontal projection method. The peak number of foreground horizontal projection in the local area is equal to the number of text lines, denoted as *TLNum*, and the position of the peak point of horizontal projection is recorded as *Ppos*. The blue histogram in Figure 3 shows the horizontal projection result, and the red line shows the baseline location of the horizontal projection in Figure 3.

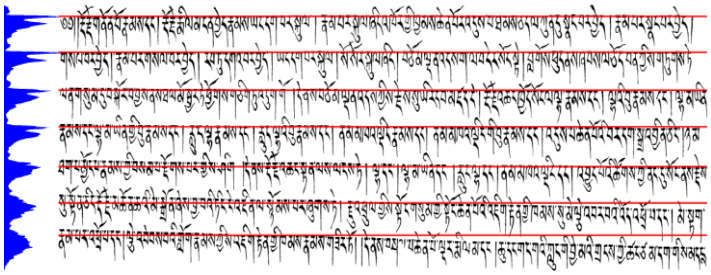


Figure 3. Document image overall projection.

2.2 Local connection

2.2.1 Upper Edge Detection. A morphological open operation is used to obtain the horizontal component of characters. The morphological operator is defined as a rectangle with SW as its height and SW as its width twice. The horizontal component is shown in Figure 4. A gradient operator is used to obtain the gradient of the horizontal component in the vertical direction, where the gradient of the upper edge is positive and the gradient of the lower edge is negative. The upper edge of the character is obtained by preserving all pixels with a positive gradient, as shown in red part in Figure 5.

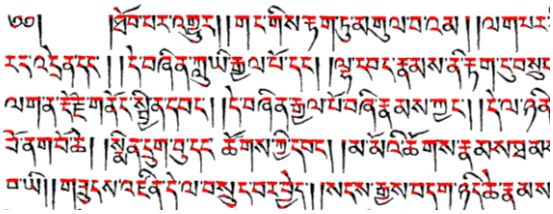


Figure 4. Horizontal component of characters.

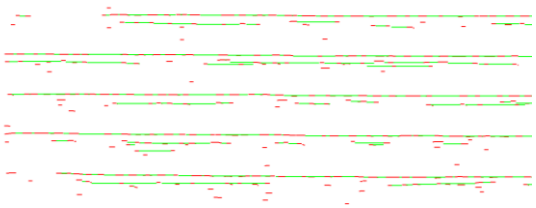


Figure 5. Upper edge local connection.

2.2.2 Local Connection. The upper edge is locally connected to form a longer baseline. The rules for connecting the upper edge lines are as follows: for each upper edge, subsequent edge lines are selected by calculating horizontal and vertical distances and then connected to get longer baselines. The horizontal distance of the upper edge is defined as equation (3):

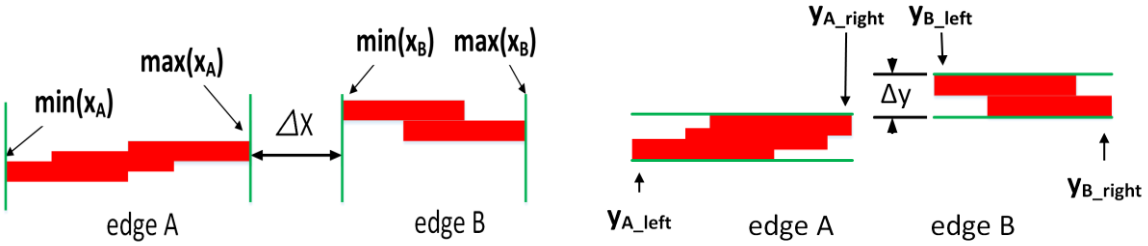
$$\Delta x = \min(x_B) - \max(x_A) \tag{3}$$

where, x_A and x_B represent the column coordinates of edges A and B, as shown in Figure 6a. Because Tibetan syllables usually consist of no more than 4 characters, the horizontal distance threshold is set to 4 times the ACW.

The vertical distance between upper edge line segments is shown in equation (4), the threshold value of vertical distance is 2 times of SW.

$$\Delta y = abs(y_{A\_right} - y_{B\_left}) \tag{4}$$

where, Y_{A_right} and y_{B_left} represent the row coordinates of the right endpoint of edge A and the left endpoint of edge B, as shown in Figure 6b.



(a) (b)

Figure 6. The horizontal and vertical distances between the upper edges.

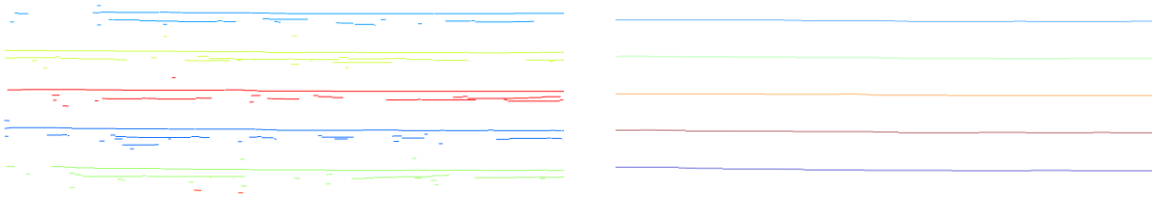
After the upper edge is connected locally, the short upper edges form a long baseline, as shown in Figure 5. The green line is the connecting line between the upper edges.

2.3 Baseline evaluation

2.3.1 Baseline Group. Baselines are clustered according to row coordinates, and seeds are the baseline positions obtained by horizontal projection. By clustering, baselines belonging to the same text line are marked. The number of document image text lines has been confirmed by horizontal projection, and the baseline is divided into TLNum sets, denoted as BaselineSets. Each set contains one or more elements, and each element is a baseline. Grouping results of baselines are shown in Figure 7a, different colors represent different sets.

The baselines in each set are analyzed to find the line pairs that are mutually exclusive. The line pairs that are mutually exclusive cannot be in the baseline combination at the same time. If the horizontal distance from line A to line B is negative, that is, the two lines overlap vertically, then line A and B are mutually exclusive.

Finally, the baselines in the same set are arranged. According to the mutual exclusion of baselines, the permutations with mutual exclusion are excluded and only reasonable line combinations are retained.



(a) (b)

Figure 7. Baseline groups and text line baselines.

2.3.2 Baseline Combination Evaluation. A reasonable baseline should have three characteristics: the longest length of the baseline, the highest number of pixels where the baseline intersects the foreground character, and the lowest horizontal fluctuation of the baseline. Therefore, the following three parameters are given to evaluate the baseline combination:

(1) Baseline Maximum Length (ML): the Maximum Length that can be reached after all lines are connected in a combination, as shown in equation (5):

$$ML\_k = \max(y_{BLgroup\_k}) - \min(y_{BLgroup\_k}) \tag{5}$$

where, $y_{BLgroup\_k}$ represents the column coordinates of all baseline segments in the Kth baseline combination.

(2) Crossed pixels between baseline and character: the number of Crossed pixels between baseline and character foreground in baseline combination is shown in equation (6):

$$CP = \sum_{i=1}^{n}\sum_{j=1}^{n} Image(x_{BLgroup\_k}, y_{BLgroup\_k}) \tag{6}$$

where, $x_{BLgroup\_k}$, $y_{BLgroup\_k}$ represent the row and column coordinates of all lines in the *K*th baseline combination.

(3) Horizontal Fluctuation Ratio (HFR) of baseline: the Fluctuation degree of baseline combination in the Horizontal direction, as shown in equation (7):

$$HFR = (\sum_{L=1}^{k-1} abs(\min(x_{L+1}) - \max(x_L))) \Big/ ML\_k \tag{7}$$

where, x_L represents the row coordinates of the L baseline in the baseline combination.

After obtaining the three parameters of each combination connection, the baseline combination is evaluated, and the evaluation function is shown in equation (8):

$$F(Lgroup\_k) = \frac{CP}{\left(\left[1 - \left(\frac{ML\_k}{LenMax + \varepsilon}\right)^2\right] \times ML\_k \times HFR\right)} \tag{8}$$

where, *LenMax* represents the maximum length of text line in the current document image, that is, the maximum horizontal distance of the document foreground character, and ε is a small positive number to prevent the occurrence of denominator 0.

The combination with the highest score is selected, and the baselines in the combination are connected from left to right to get the baseline of the text line. Baselines of text line are shown in Figure 7b, where different colors represent different baselines.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Dataset construction

In order to evaluate the baseline detection results of Tibetan historical documents, a data set for baseline detection evaluation is constructed. The data set contains 212 document images of Kangyur (Beijing version) with 1696 text lines. The baseline of text lines in 212 document images are manually marked, as shown in the red line in Figure 8.

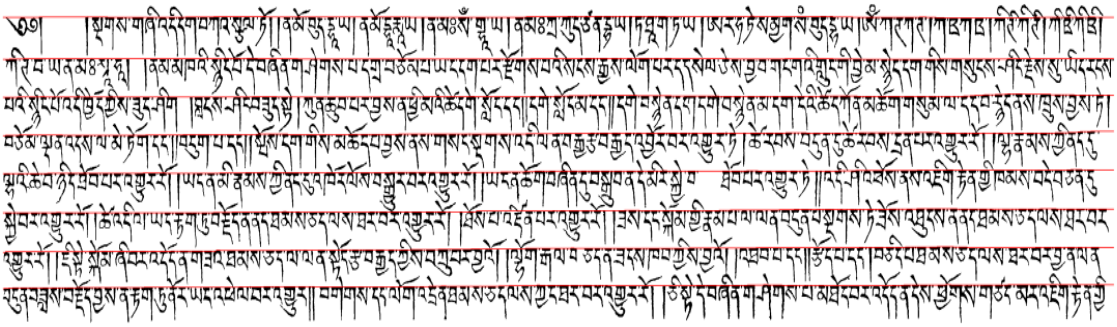


Figure 8. Tibetan historical document text line baseline ground truth.

3.2 Accuracy of baseline detection

3.2.1 Evaluation Indicators. In order to evaluate the results of baseline detection, the deviation distance of baseline (DDb) is proposed, as shown in equation (9):

$$DDb_{\mathrm{L}} = \frac{1}{n}\sum_{i}^{n} abs(y_{Li} - y_{MLi}) \tag{9}$$

where, DDb_{L} represents the deviation distance of baseline L, n represents the length of baseline, y_{Li} represents the abscissa of the ith point in baseline L detected by the algorithm, and y_{MLi} represents the abscissa of the corresponding pixel point in the data set.

If the deviation between the detected baseline and the labeled baseline is less than the Vertical Distance threshold (VDT), the current baseline detection result is considered correct.

Let N is the total number of baselines in the data set, and M is the number of correct Detection results. The definition of Base line Detection Accuracy (BDA) is shown in equation (10).

$$BDA = \frac{M}{N} \tag{10}$$

Table 1. The BDA of the method proposed.

VDT	M	BDA
1	1665	98.17%
2	1681	99.12%
3	1695	99.94%

Table 1 shows the accuracy of the baseline detection method proposed at different vertical distance thresholds.

3.2.2 Comparison with Other Methods. The horizontal projection[5], piecewise projection[7], local connection[8] methods and the proposed method were tested on data sets. DDb was taken as the evaluation index, and VDT was set to 3. Table 2 lists the indicators of different baseline detection algorithms.

Table 2. The algorithm comparison results.

Method	Horizontal projection	Piecewise projection	Local connection	Proposed
DDb	4.7	3.6	2.1	1.9
BDA	83.7%	94.1%	96.2%	99.9%

Table 2 shows the performance of the proposed method and others. Compared with other methods, the method proposed in this paper is superior to other methods in the evaluation index. It shows that the proposed method can well adapt to the distribution of text lines in Tibetan historical documents and can effectively overcome the distortion of text lines and extract the accurate baseline of text lines.

4. CONCLUSION

Tibetan historical documents are rectangular and have skew and distortion. The baseline position becomes important information for text line extraction. In this paper, the baseline of text line is detected by local connection and global evaluation. Firstly, the slanted document is corrected and the number of text lines is determined by horizontal projection. Then, the upper edge of characters is detected and connected. Finally, the baseline of text lines is obtained by grouping and evaluation. Compared with other text line baseline extraction methods, the method proposed in this paper has the best accuracy in line baseline detection and can effectively extract the baseline position of extra-long text lines with distortion.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61772430, No. 62166036), the Gansu Provincial first-class discipline program of Northwest Minzu University (No.11080305), the Program for Leading Talent of State Ethnic Affairs Commission, Natural Science Foundation of Gansu Province of China (No. 21JR1RA195), the Scientific Research Foundation of Northwest Minzu University (No. xbmuyjrc2021015).

REFERENCES

[1] Anusree, M., Dhanya, M. and Dhanalkshmy, "Text line segmentation of curved document images—A survey," International Journal of Engineering Research & Applications, 4(5), 332-343(2014).
[2] Ma, L., Long, C., Duan, L., Zhang, X., Li, Y. and Zhao, Q., "Segmentation and recognition for historical Tibetan document images," IEEE Access, 8, 52641-52651(2020).
[3] Zhou, F., Wang, W. and Lin, Q., "A novel text line segmentation method based on contour curve tracking for Tibetan historical documents," International Journal of Pattern Recognition and Artificial Intelligence, 32(10), 1854025(2018).
[4] Li, J., Wang, X., et al., "Text line segmentation combines the text core regions and expansion growth for tibetan historical document," Journal of Laser & Optoelectronics Progress, 58(02), 113-123(2021).
[5] Cai, Z. and Cai, R., "Research on the distribution of Tibetan character form," Journal of Chinese Information Processing, 30(4), 98-105(2016).
[6] Li, Y., Ma, L., Duan, L., et al., "A text-line segmentation method for historical Tibetan documents based on baseline detection," CCF Chinese Conf. on Computer Vision, 356-367(2017).
[7] Wang, Y., Wang, W., Li, Z., Han, Y. and Wang, X., "Research on text line segmentation of historical Tibetan documents based on the connected component analysis," PRCV, 74-87(2018).
[8] Hu, P., Chen, Y., Hao, Y., Wang, Y. and Wang, W., "Text line segmentation based on local baselines and connected component centroids for Tibetan historical documents," Journal of Physics: Conference Series, 1656(1), 012034(2020).
[9] Hu, P., Wang, W., Li, Q. and Wang, T., "Touching text line segmentation combined local baseline and connected component for Uchen Tibetan historical documents," Information Processing & Management, 58(6), 102689(2021).
[10] Li, Z. J., Wang, W. L. and Lin, Q., "Tibetan historical document recognition of Uchen script using baseline information," 10th Intern. Conf. on Graphics and Image Processing, (2018).
[11] Li, Z., Wang, W., Chen, Y., et al., "A novel method of text line segmentation for historical document image of the Uchen Tibetan," Journal of Visual Communication and Image Representation, 61, 23-32(2019).