Research on Text Line Segmentation of Historical Tibetan Documents Based on the Connected Component Analysis

Yiqun Wang[1], Weilan Wang[1(✉)], Zhenjiang Li[1], Yuehui Han[1,2], and Xiaojuan Wang[1]

[1] Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730000, Gansu, China
wangweilan@xbmu.edu.cn
[2] College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730000, Gansu, China

Abstract. Text line segmentation is one of the critical content in handwriting documents recognition especially in the historical documents' analysis and recognition. Because of the low quality and the complexity of these documents (background noise, scattered character, touching components between consecutive lines), automatic text line segmentation remains to be a hot spot for researching. In this paper we propose a new method to segment the text line from the historical Tibetan scripture "kangjur" of the Beijing version on the paper by means of woodcut. This method first performs document image skew detection and correction, using projection profiles to get the baseline of text line, then the connected component is allocated to text line according to the location relationship. For some connected components, analyzing their location and sharp to assign these connected components correctly. This method using connected component instead of pixels, avoiding the noise generated by splitting characters. Experiments show that this method is effective in copes with touching text lines and promising in text line segmentation from historical Tibetan document.

Keywords: Historical Tibetan document · Kangjur · Text line segmentation
Component analysis · Location · Sharp

1 Introduction

The Tibetans have a large number of historical documents; most of them are stored in temples. Those historical documents are exist in the form of scriptures for a very long time. It is urgent to protect and reuse them by using digital technology because of the deterioration of the quality of the historical documents. Using Optical Character Recognition (OCR) technology to converts the historical Tibetan documents into text files. The text files stored in the services is not only appropriate preserved but also convenient for reusing those precious historical documents. In document processing field, the segmentation is essential for document recognition which it needs several steps of binarization, layout analysis, text blocks extraction, text lines and words segmentation

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 74–87, 2018.
https://doi.org/10.1007/978-3-030-03338-5_7

and character recognition. The degraded historical documents (e.g. ink stains, torn pages, overlapped/touching character, broken stroke etc.) make a challenge for the text line segmentation task. The variation of the interline distance and the baselines undulation between lines or even along the same text line. The touching characters between adjacent text lines appear frequently in Tibetan documents. The whole characters may be divided into several parts because of broken stroke. All above greatly complicates the task of the text lines segmentation from historical handwritten document.

In this paper, we focus on the extraction of text lines from historical Tibetan documents and we propose a method based on the analysis of the location and shape of the connected component. This method cannot totally solve the problem of segmentation, but we try to reduce the error as much as possible to extract text line complete. For text line extraction of historical Tibetan documents, a few researche have been done such as: based on baseline detection method [1] and contour curve tracking method [2]. Other common text line extraction methods also include: projection-based method [3], Hough-transform [4], smearing method [5], clustering approach [6, 7].

In [1] the baseline is getting by template matching, pruning the salient strokes and closing operation, then touching characters is detecting and splitting, the text-line is extracted according to baseline and split position, this method can deal with the touching characters and fluctuating text lines. However, this method does not consider broken strokes, so it is inadequate for some historical document image with a large number of broken strokes.

In [2] the text line segmentation method based on contour tracking is proposed. The text line is extracted by the contour from the document image which comes from the constructed connected component. The method combine the barycentre coordinates of the connected component to form the curve line and the separated components are assigned to the corresponding text line by the barycentre gravity later. The text line is obtained by the contour curve of the text line. This method is innovative but the performance is not satisfactory when a document image with many touching characters is segmented.

Projection-based method [3] is most commonly used for the text line segmentation especially in printed or slightly document. The projection value is computed by summing the values of pixel in the foreground in horizontal axis of each line. The text lines is segmented by straight lines with suitable positions and directions, this method is not suitable for historical Tibetan document as there is no obvious line gap. According to the layout of the Tibetan Scripture “Kangjur”, the direction of the text lines is approximately horizontal parallel, so this method can be used to find the baseline of the text lines.

Hough-based method [4] is proper to detect text lines which are usually parallel in certain areas. Smearing method [5] enlarged area of black pixels, the white space between the black pixels is filled with black pixels if their distance is within a predefined threshold. But this method is not suitable for historical Tibetan document. Because some vertical stroke is overlong that smearing horizontal will produce more touching components.

Clustering method [6, 7] usually divides a picture into several connected components, blocks or other units according to some features, and then aggregates these units to form alignments according to some rules. Considering that there are a lot of touching

characters in historical Tibetan documents, it is very difficult to assign the characters to the correct text lines in this way. Thus, this method is not suitable for text line segmentation from historical Tibetan documents.

The paper is organized as follows: in Sect. 2, our method is described. In Sect. 3 the proposed method to segment text lines is detailed. Section 4 present the experimental results and discuss. Section 5 describes conclusions and future work.

2 Our Method

Tibetan character can be regarded as a kind of string composed of basic characters and characters in the vertical direction [8] (see Fig. 1). The authentic historical Tibetan document not only have lots of touching characters between adjacent lines as the height of the character is inconsistent but also have lots of broken strokes than other languages. The touching characters between adjacent line, the separated upper and lower vowels and the broken strokes make the text-line segmentation more complex (see Fig. 2). At present, there is no satisfactory segmentation method for the authentic historical Tibetan document of wooden printing.

Fig. 1. (a) Character with upper vowel (b) character with lower vowel.

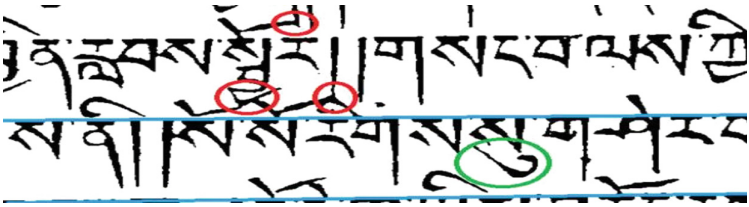


Fig. 2. Partial image with slanted baseline, separated character, overlong and touching characters.

We can see the characters in the historical Tibetan document are very close to each other because of the limited area of the document and there is no obvious gap between adjacent lines. The historical document images have large number of touching and overlapped characters and variety of broken strokes which are the main challenge to extract text lines accurately.

In order to extract the text-line completely from handwritten, degraded, historical Tibetan documents, we present a text-line segmentation method which combine the row projection location analysis and shape analysis of connection components. Our method stems from the idea that the text line is composed of a set of location related components. The task of text line segmentation is to find such a set of components and extract them from the document image to form a text line.

Our method detects the input document image whether the image is skew or not and perform skew correction if it is. Then the position of baseline is obtained using projection method as the text line is approximately horizontal after skew correction.

The connected component is allocated to text line according to the location relationship between the component and segmentation line by their location information. For some connected components, it is difficult to assign them to the corresponding text line only depends on location. Generally speaking, these components are broken strokes, separate vowels, symbols, touch characters, noise, and so on. Therefore, it is necessary to make a further analysis of the location and shape of these connected components in order to correctly determine their attributes. Combining location and shape information to determine which text line these connected components should belong to will be more accurate, especially for complex documents. At last, the components belong to the same alignment are merged to recover the text line.

Here is the architecture we extract text-lines from Tibetan historical documents shown in Fig. 3.

Fig. 3. The text line segmentation process.

Our method includes four stages:

1. Pre-processing: We detect whether the input image is skew. If the image is skewed, the skew correction is done to make the text lines in the image horizontally parallel. Then, the information about height of character is got which will be used to estimate the feature of characters in next stage. At last the position of baseline is detected using the projection method.
2. Location analysis: According to the baseline position we obtained before, the text line region is extracted from the input image as a rectangle, and divide the region into upper part and lower part according to the baseline position of the current baseline. The upper part is undoubtedly part of the current text line, but the lower part contains some components of the next text line. Next, the projection method is used to find the optimal segmentation line (SL) which is the row's location with minimum pixels in the lower part. Then the connected component in the lower part is divided into three classes according to whether it intersects with the SL. Some connected components are belongs to current text line or next text line certainly but

the others cannot easily determine which text line they belong to, so further analysis is needed.

3. Location and Shape analysis: By judging whether there is intersection point with SL, the connected components with uncertain attribution is divided into one class. By analyzing the location information and shape information of the connected component in this class, we classify it into the correct text lines, especially for the touching characters between the text lines, we use some features and rules to detect and separate them.
4. Image merging: Through the Location and Shape analysis (LSA) of the connected components, the connected components belonging to the current text line have been marked out. Combining these connected components to form the lower part of the current text line, and then splicing the upper and lower parts to form a complete image of the current text line.

3 Text Line Segmentation

The proposed text line segmentation method base on the projection, location and shape analysis of connected components for historical Tibetan handwritten document deals with the following challenges: (i) parts of neighboring text lines may be connected; (ii) overlong character and touching character in text line; (iii) the separated vowel may be appeared either above or below the text line and (iv) the broken strokes of characters in text line. The work flow of the text line segmentation is shown in Fig. 4.

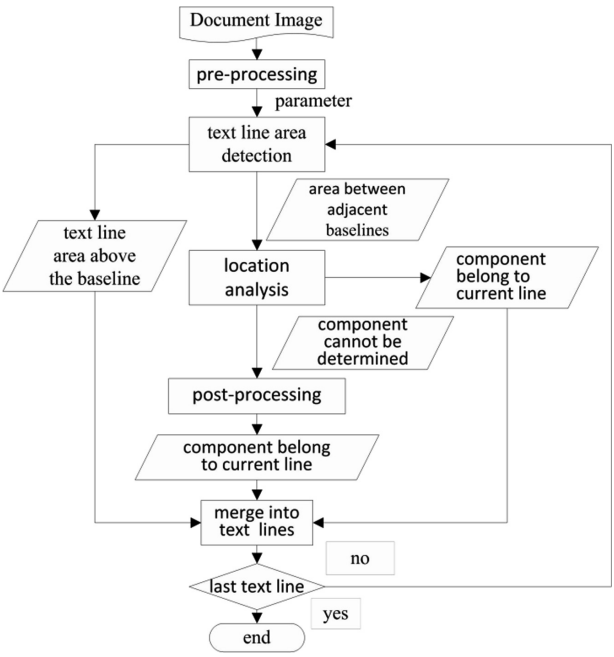
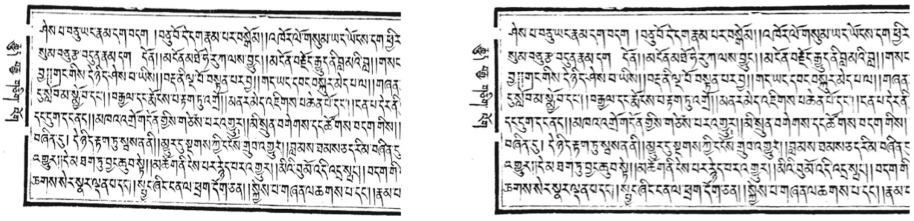


Fig. 4. Proposed method framework.

3.1 Pre-processing

The pre-processing stage consists of three steps. First, whether the input image is skew is detected, the document image is skew corrected if the image is skew. The angle of the skew correction is determined by the length of the border detection line, the method rotates the image from -2 angle to +2 angle at step 0.1, and detects the sum of the length of the edge lines of the four borders, the maximum sum corresponding angle is the correction angle. An example is shown in Fig. 5. Then, average character height (AH) and the average component height (ACH) for the whole document image are calculated and the bordering box is removed. Last, the baseline position of each text line is obtained by row projection profile method, and the number of locations equals the number of text lines. An example is shown in Fig. 6.



(a) the document image is skew (b) after skew correction

Fig. 5. The input document image is skew (a) and the document image after skew correction (b).

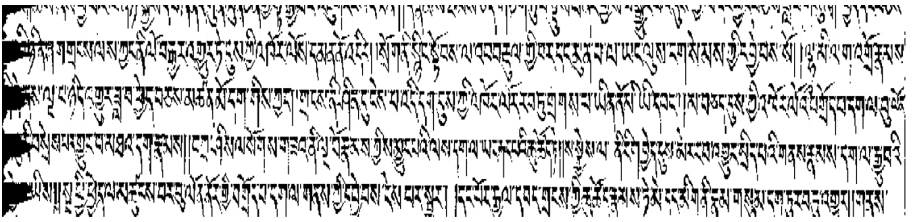


Fig. 6. Row projection diagram of binary document image.

3.2 Location Analysis

This stage includes two steps. At the first step, the projection method is used to get the initial row position of the text line that is the line of beginning (LB) then extract the area between LB and the baseline location of the current text line as the upper part of the current text line image, and this part is denoted as “upper image” (see Fig. 7. black part). The next step will analyze the image (“lower part”) between the baseline of current text line and the next baseline. Firstly, the statistical method is used to find the optimal segmentation line (SL) which is the row’s location with minimum pixels. Next,

by using relative location relations between components and SL, the connected components domain is divided into three sub-domains, which are denoted as "Subsetcur", "Subsetlow" and "Subsetcs" respectively.

"Subsetcur" contains all components which totally are located above the SL (see Fig. 7. green part) and "Subsetlow" contains all components which are located below the SL (see Fig. 7. blue part). "Subsetcs" contains all the components which have the intersected points with the SL, this subset have various components that need to be analyzed in different manners by the proposed method in the next stage (see Fig. 7. red part).

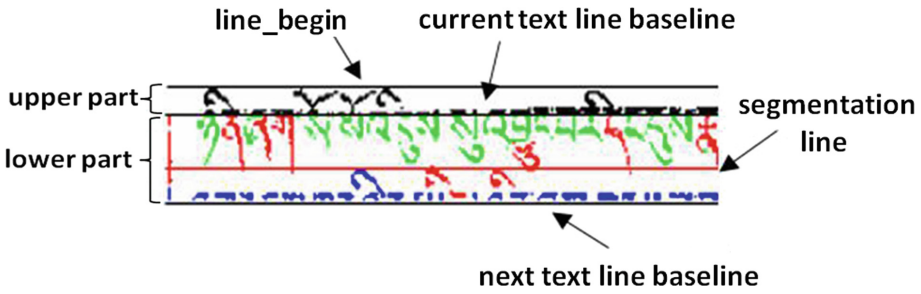


Fig. 7. An example of partitioning the connected components by the relationship between the component and the segmentation line. The black part is upper image, the lower part is the region between current baseline and next text line' baseline, the green part means "subsetcur", the red part means "subsetcs" and the blue part means "subset low". (Color figure online)

3.3 Sharp-Analysis

This stage analyzes the location and shape of the components which is in the "subsetcs" to determine whether it belongs to the current text line or not. The categories of these components in the "subsetcs" are separated into upper vowels and lower vowels, broken strokes ,overlong characters, touching and overlapped characters, and bar shaped connected components. All connected components in "subsetcs" have a common property that they intersect with SL, in other words, SL divides these connected components into upper and lower parts. In order to assign connected components to the corresponding text lines accurately, we need to extract some features of these connected components, such as the height of connected components (H), the height above the SL (HA), the pixel per row for the part above the SL (PPRA), the height below the SL (HB), the pixel per row for the part below the SL (PPRB), and the ratio of the foreground area to the minimum rectangular bounding area (RFB).

The PPRA is calculated as follows: (value 1 for foreground and 0 for background pixels)

$$PPRA = \sum_{x=1}^{width} \sum_{y=1}^{HA} I(x,y)/HA \; if \; I(x,y) = 1 \tag{1}$$

The PPRB is calculated as follows:

$$PPRB = \sum_{x=1}^{width} \sum_{y=1}^{HB} I(x,y)/HB \; if \; I(x,y) = 1 \tag{2}$$

The RFB is defined as follows:

$$RFB = \sum_{x=1}^{width} \sum_{y=1}^{height} I(x,y)/width * height \; if \; I(x,y) = 1 \tag{3}$$

The location and shape analysis (LSA) procedure consists of two steps. At the first step, the feature obtained above are used to determine whether the connected components lying in subsetcs are belong to the current text line or not according to the following conditions.

In the first step, the method take advantage of the feature we obtained above and the average character height (AH) and the average component height (ACH) which are got at first stage to classify them into three categories by rules. The first category have the connected components which are assigned to the next text line. One category consists of components that in this step cannot determine the attribution of text lines, and these components will be analyzed shapes in the next step. The last category includes the components of the current text line, usually consisting of overlong characters, symbols, and touching characters. The touching character will be segmented and retain the component belonging to the current text line.

The broken strokes and separated vowels were selected by conditions 4. The condition is described fellow:

$$H < ACH \tag{4}$$

The connected component is belongs to the current text line, if some features satisfy the condition below:

$$(H > AH) \; and \; (HA > HB) \tag{5}$$

Identify the connected component with height exceeds the height threshold which is defined as:

$$HT = 1.5 * AH \tag{6}$$

The connected components which satisfied the above conditions include the overlong characters (see Fig. 8, a b c), the touching characters(see Fig. 8, d e) and the bar-shaped connected components which generally are Tibetan character symbol(see Fig. 8, f).

The bar-shaped connected components usually are symbol which is belong to current text line. Such component will be selected if the following condition is satisfied:

Fig. 8. The image of overlong character, the overlong characters (a b c), the touching characters (d e) and the bar-shaped connected components (f).

$$RFB > 0.5 \tag{7}$$

The touching character are as long as the overlong character (see the Fig. 6a b c and d e). Choose the touching character according to the following constraint.

$$PPRB > 1.2 * PPRA \tag{8}$$

The LSA first step work flow is shown in Fig. 9

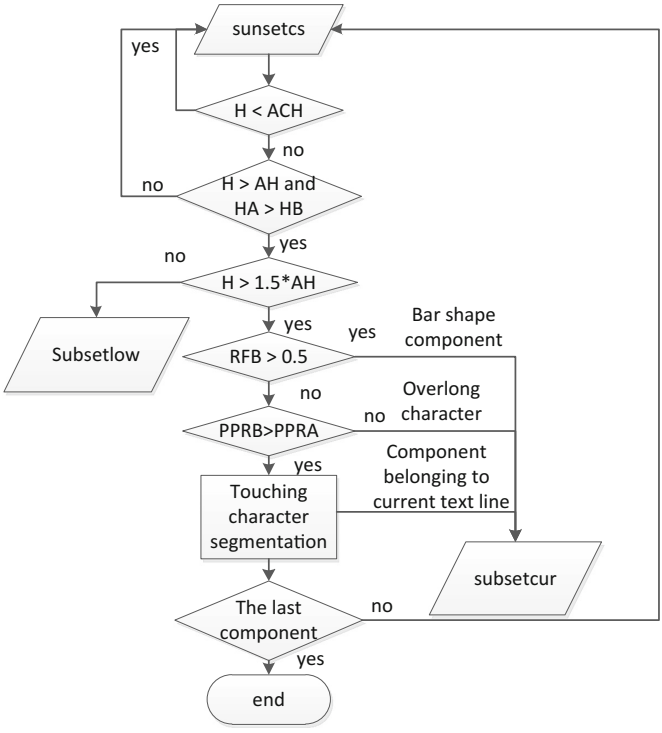


Fig. 9. The LSA first step work flow

The second step continues to deal with connected components still in subsetcs, which are broken strokes (see Fig. 10, a b c d), separated lower vowels belonging to the current text line (see Fig. 10, e f g), and upper vowels belonging to the next text line (see Fig. 10, h i j).

This step has three works to do:

1. Calculate the centroid and the skeleton of connect components, then detect the intersection between skeleton and the line located by the centroid, and calculated the numbers and the coordinate positions of the intersected points.
2. For the connected components with only one intersected point (see Fig. 11 a b c d), move it from subsetcs to the subsetcur if its centroid position is above the segmented line, or it belongs to subsetlow if its centroid position is below the segmented line.
3. For the connected components with two intersected points, the skeleton is segmented into the upper part and the lower part according to the line located by the centroid and the coordinate of two intersected points. The number of pixels in the two parts is counted respectively. Connected components are assigned to subsetcur if the pixels in the lower part is more than that in the upper part (see Fig. 11 e f g), otherwise, the connected components will belong to subsetlow(see Fig. 11 h i j).

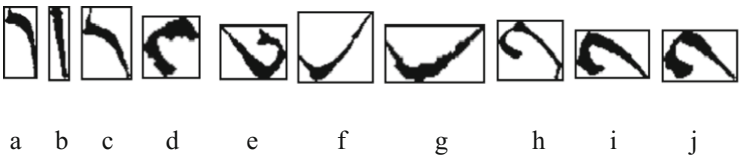


Fig. 10. The connected components of broken strokes, separated lower vowels and separated upper vowels

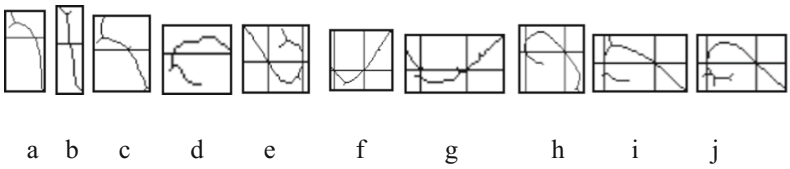


Fig. 11. The skeleton diagram with line located by the centroid

3.4 Merging Image

Since all the connected components that belong to the current text line have been marked in the subsetcur, so the lower part of the current text line is generated by the subsetcur. The complete image of the current text line image is got by merge the upper part and lower part. The input image subtracts the current text line image from the position of the LB to produce a image that is the input image for the next text line.

4 Experimental Results and Discussion

The experimental dataset are from the historical Tibetan scripture "kangjur" of the Beijing version on the paper by means of woodcut. The scripture "kangjur" of the Beijing version have more than 60 thousand images, the dataset just have 1696 text lines from 212 images which is selected at random. The method presented in this paper is implemented in matlab.

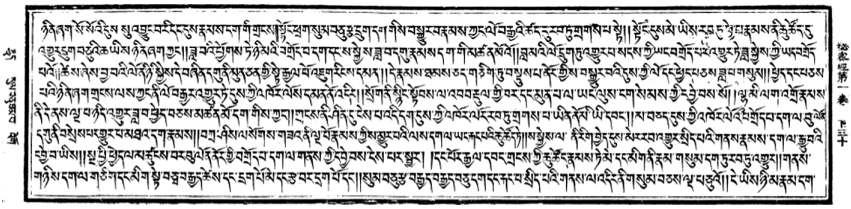


Fig. 12. The input image

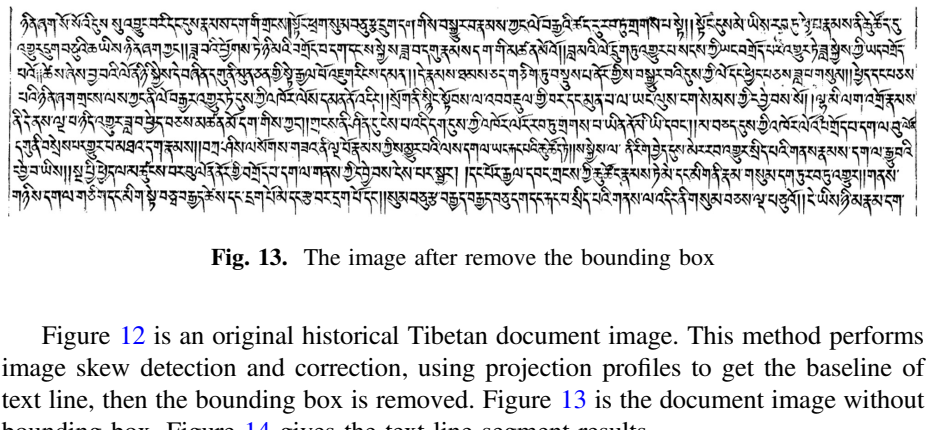


Fig. 13. The image after remove the bounding box

Figure 12 is an original historical Tibetan document image. This method performs image skew detection and correction, using projection profiles to get the baseline of text line, then the bounding box is removed. Figure 13 is the document image without bounding box. Figure 14 gives the text line segment results.

Let N be the number of all text lines, G_j the set of all points inside the ground truth region, R_i the set of all points inside the corresponding result region. The detection rate (DR) and the recognition accuracy rate(RA) are defined as follows:

$$DR = \frac{G \cap R}{G}, RA = \frac{G \cap R}{G} \tag{9}$$

Because text line segmentation is an important part of OCR recognition system, the ideal situation is that the text lines only contains all the components belonging to the text line, and it does not lose any component and does not have any component that do not belong to them. Therefore, we propose completeness rate to measure the segmentation effect. The definition of integrity is as follows:

$$CR = \frac{\sum N_i}{N} \quad N_i = 1, \ if \ G_i = R_i, otherwise \ N_i = 0 \tag{10}$$

Table 1 shows the performance of contour curve tracking method and our method. Comparing with the contour curve tracking method, our method has a considerable improvement in each evaluation value.

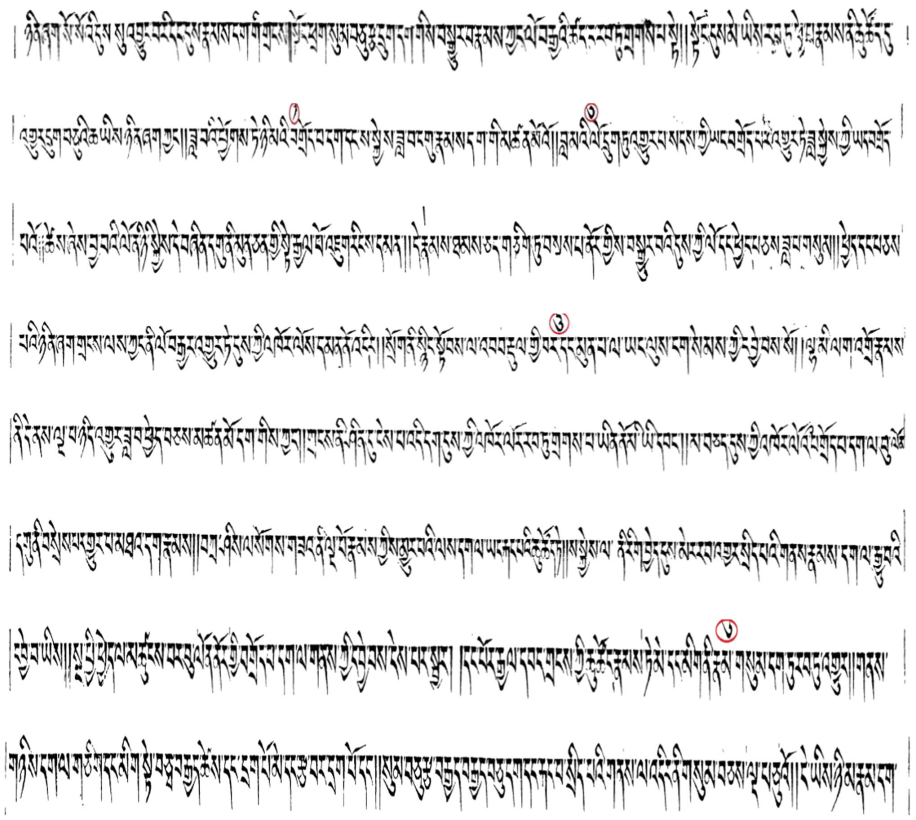


Fig. 14. Result of text line segmentation

Table 1. The performance of contour curve tracking method and our method

Method	N	DR	RA	CR
Contour curve tracking	2196	82.79%	80.09%	33.23%
Our method	2196	91.17%	90.23%	37.51%

Experimental results show that almost all the components belonging to the wrong text line are caused by broken strokes and separated lower vowels. And this method is very efficient to detect the touching characters in adjacent text lines. There are 874 touching characters in the dataset of the 212 pictures, and 840 of them are detected successfully, the touching character's detect ratio is 98.4%.

5 Conclusion and Further Work

Text line segmentation is still one of the most challenging topics in document image analysis. In this paper, we present a text line segmentation method for handwritten historical Tibetan documents based on connected components analysis. This method correct the skew document image, gets the reasonable baseline position by the contour projection, and obtains the text line region by the baseline position from the document image. The connected component's attribution is decided by analyzing the location and shape. The method is suitable for text segmentation from complex layout document image and can overcome the slightly fluctuation of text line. Although the algorithm is reasonably designed and many features about location and shape are analyzed, there are still many wrong parts in the extracted text line image.

Low completeness rate of text line segmentation is not only caused by strict standards, but also by the real historical handwritten documents that is more complicated because of the high frequency of separated vowel characters, broken strokes, and touching characters.

Through experiments, we get the following conclusions for the text line segmentation task for the degraded Tibetan historical document image of wooden printing: (i) the method based on the connected component analysis is feasible for text line segmentation. (ii) it is necessary to correct the skew document image for text line segmentation.(iii) the problem of touching and overlapped characters in text line segmentation of historical Tibetan documents can be solved effectively. (iv) it is not enough to make use of a few features to identified the shape of character.

The focus of future work is to study the shape recognition algorithm of similar vowels and broken strokes. Another issue is to research the better segmentation algorithm for touching and overlapped character.

Acknowledgments. This work was supported by the National Science Foundation (No.61772430), the Program for Leading Talent of State Ethnic Affairs Commission, the Fundamental Research Funds for the Central University of Northwest Minzu University (No. 31920170142), and also supported by the Gansu Provincial first-class discipline program of Northwest Minzu University.

References

1. Li, Y., Ma, L., Duan, L., Wu, J.: A text-line segmentation method for historical tibetan documents based on baseline detection. In: Yang, J., et al. (eds.) CCCV 2017. CCIS, vol. 771, pp. 356–367. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-7299-4_29
2. Zhou, F., Wang, W., Lin, Q.: A novel text line segmentation method based on contour curve tracking for tibetan historical documents. Recogn. Artif. Intell. **32**(10), 1854025 (2018). Image Processing
3. Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1212–1225 (2005)
4. Bar-Yosef, I., Hagbi, N., Kedem, K., Dinstei, I.: Line segmentation for degraded handwritten historical documents. In: 10th ICDAR, pp. 1161–1165 (2009)
5. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. Int. J. Doc. Anal. Recogn. **9**(2), 123–138 (2007)
6. Garz, A., Fischer, A., Bunke, H., Ingold, R.: A binarization-free clustering approach to segment curved text lines in historical manuscripts. In: International Conference on Document Analysis and Recognition, pp. 1290–1294 (2013)
7. Zahour, A., Likforman-Sulem, L., Boussalaa, W., Taconet, B.: Text line segmentation of historical arabic documents. In: 9th International Conference Document Analysis and Recognition, vol. 1, pp. 138–142 (2007)
8. Baima, Y.Z.: Research on feature extraction of tibetan characters. Comput. Knowl. Technol. **9**, 6362–6364 (2013)