A Text-Line Segmentation Method for Historical Tibetan Documents Based on Baseline Detection

Yanxing Li[1,2], Longlong Ma[3], Lijuan Duan[1,4](✉), and Jian Wu[1,3]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing, China
liyanxing15@outlook.com, ljduan@bjut.edu.cn, wujian@iscas.ac.cn
[2] Beijing Key Laboratory of Trusted Computing, Beijing, China
[3] Chinese Information Processing Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China
[4] Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data,
Beijing, China
longlong@iscas.ac.cn

Abstract. Text-line segmentation is an important task in the historical Tibetan document recognition. Historical Tibetan document images usually contain touching or overlapping characters between consecutive text-lines, making text-line segmentation a difficult task. In this paper, we present a text-line segmentation method based on baseline detection. The initial positions for the baseline of each line are obtained by template matching, pruning algorithms and closing operation. The baseline is estimated using dynamic tracing within pixel points of each line and the context information between pixel points. The overlapping or touching areas are cut by finding the minimum width stroke. Finally, text-lines are extracted based on the estimated baseline and the cut position of touching area. The proposed algorithm has been evaluated on the dataset of historical Tibetan document images. Experimental result shows the effectiveness of the proposed method.

Keywords: Historical Tibetan document · Text-line segmentation
Baseline detection

1 Introduction

As original documents that contain important information about person, place, or event, historical documents are usually stored in libraries or museums. Unlike others, most historical Tibetan documents are kept in temples. Besides, most historical Tibetan documents exist in the form of scriptures rather than books. As a result, those documents are extremely hard for accessing. At present, there is a growing trend towards the digitization to make these significant documents easier to be preserved and accessed. Considering that the number of historical Tibetan documents and the limitation of technology and human resources, manual transcription is not a reasonable solution. To recognize the historical Tibetan document automatically is a proper way to solve the problem.

J. Yang et al. (Eds.): CCCV 2017, Part I, CCIS 771, pp. 356–367, 2017.
https://doi.org/10.1007/978-981-10-7299-4_29

There are four steps in the document recognition. Firstly, the document images are preprocessed by noise removing, correction of image orientation, normalization and binarization. Secondly, the layout analysis module segments the preprocessed image into text regions, pictures and tables. Text areas are obtained by removing borders and margins that may interfere with later operations. Thirdly, the text areas are further segmented into text-lines by text-line segmentation methods. Finally, text-lines are sent into a character segmentation and recognition system, which converts digitized documents into text files, usually in ASCII format. After digitalization, the documents are not only convenient for accessing, but also are protected properly. During the process of document recognition, in particular, text-line segmentation is a significant stage to ensure better performance. What's more, text-line segmentation is also an important process for some document analysis tasks. However, historical Tibetan document images usually contain touching or overlapping characters between consecutive text-lines, making text-line segmentation a difficult task.

The main objective of our work is to extract text-lines in historical Tibetan documents. Unlike many documents written by other languages, the layout structure of historical Tibetan documents is more complex than printed document. Text-line extraction is more challenging on Tibetan historical document because of the curved lines and the touching components in the documents.

The rest of this paper is organized as follows. Section 2 gives a brief introduction about the basic methods in text-line extraction and some discussions about the touching components in Tibetan text-lines. Section 3 gives the main methodology. Section 4 shows the experimental results. The conclusion is presented in Sect. 5.

2 Related Works

Few researches have been done for text-line extraction of historical Tibetan documents. Considering that Tibetan is a spelling language, we refer some works which are used for text-line extraction of historical Latin script. Generally, text-line segmentation can be classified into five categories [1] projection-based method, smearing method, grouping method, Hough-based method, and repulsive attractive network (RA-network) method.

Projection-based method [2] is most commonly used for the printed document. This method regards the pixel in the foreground as 1 while other is 0. The projection value is computed by summing the values in horizontal axis of each line. Nevertheless, this method is very effective on printed or slightly sticky document. Considering the overlapping and touching components usually occurs in historical Tibetan documents and the text-line is curved, we cannot employ this method directly on text-line segmentation of historical Tibetan documents. However, according to the writing habit of humankind, the direction of the text-lines in the former columns is approximately parallel. Because of this, the method can be used to find the number of text-lines.

Smearing method [1] fills the white space between black pixels if the distance of consecutive white pixels is within a predefined threshold. This method is used to text-line segmentation of historical Tibetan documents, and the performance is far from satisfaction. The main reason is that some Tibetan vertical strokes can be so long that smearing the image horizontal could connect the stroke with vowels, which belong to the next text-line. As a consequence, it could produce much more touching components. Grouping method [1] regards connected components, blocks or other features as units, aggregating these units to form alignments. Taking into account lots of touching components in Tibetan document, to define some proper rules for obtaining the units and join the units together is very difficult. Thus, this method cannot be employed. Hough-based method [3] is proper to detect text-lines because text-lines are usually parallel in certain areas. RA-network [4] method constructs the base-line from the top of the image to bottom one by one. The extracted baseline acts as repulsive forces while pixels of the image act as attractive forces.

In historical Tibetan documents, touching or overlapping occurs more often than other languages. Touching and overlapping components between consecutive text-lines is the main challenges for the text-line extraction. According to the writing habits and the language features, Tibetan touching strokes can be generally grouped into three categories: SC1 (see Fig. 2) touching with the baseline of the next text-line, Fig. 1; FC (see Fig. 2) touching with the baseline of next text-line, Fig. 1(c) and the FC (see Fig. 2) touching with vowels in the next text-line Fig. 1(b).

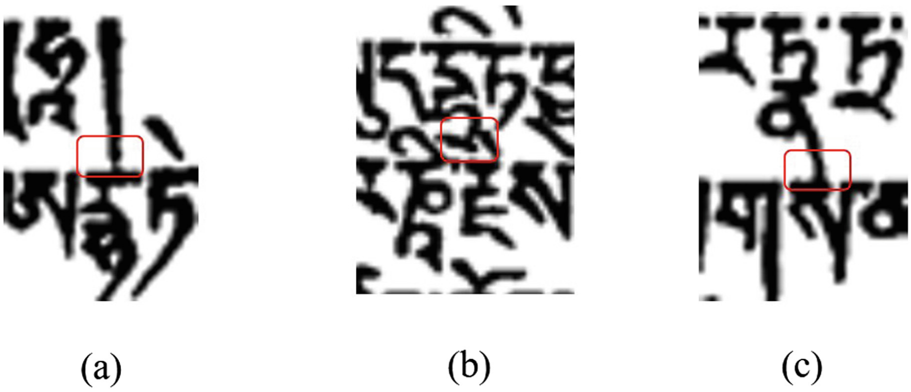

Fig. 1. Tibetan touching examples (a) SC1 touching with the baseline of the next text-line. (b) FC touching with vowels in the next text-line (c) FC touching with the baseline of the next text-line.

The detail of Tibetan scripts is shown in Fig. 2. Just like other spelling languages, Tibetan script consists of 30 consonants and 5 vowels, which are combined to produce Tibetan syllables according to the spelling rule. Each syllable has a base consonant (BC). According to the relative position to base consonant, other consonants are called prefix consonant (PC), foot consonant (FC), the first suffix consonant (SC1) and the second suffix consonant (SC2) [5] (see Fig. 2). In Fig. 2, the baseline is a fictitious line which follows and joins the consonants at the higher part of Tibetan character bodies. Most consonants lie under the baseline while only few of upper vowel letters lie above the baseline.

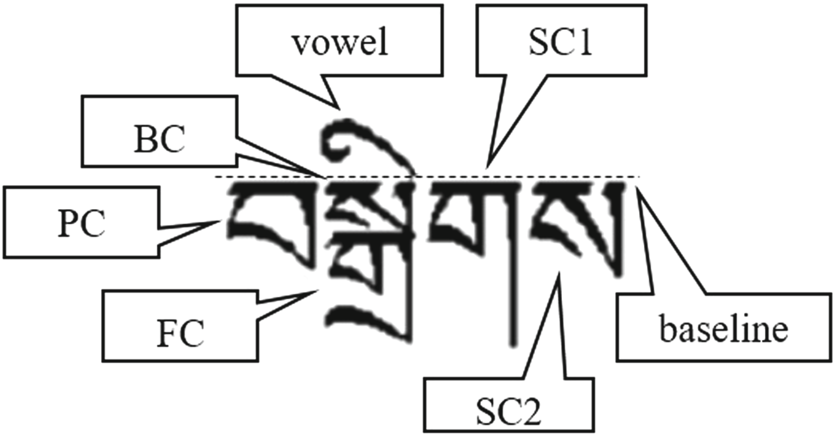

Fig. 2. The Tibetan syllable, which consist of base consonant (BC), vowel, prefix consonant (PC), foot consonant (FC), the first suffix consonant (SC1) and the second suffix consonant (SC2) [5].

3 Methodology

Here is the architecture we extract text-lines from Tibetan historical documents shown in Fig. 3. The input image is composed of the uniformed text areas of the Tibetan historical documents and is obtained by the layout segmentation method [6]. In this method, text areas are extracted based on connected component analysis and corner point detection. Firstly, document images are equally divided into $ESize * ESize$ grids. Based on the classification information of CCs and corner points, the grids are filtered by the predefined rules. By analysing vertical and horizontal projections of remaining grids, the approximate text area is located and further corrected. Based on the results of layout analysis, text-lines are extracted by our proposed text-line segmentation method.

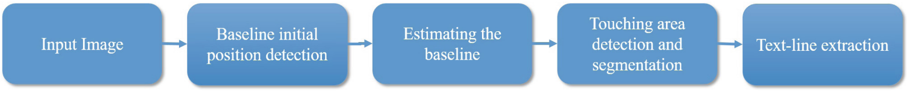
Input Image
Baseline initial position detection
Estimating the baseline
Touching area detection and segmentation
Text-line extraction

Fig. 3. The text-line segmentation process

Our method includes three stages:

1. Baseline initial position detection: Partial information is extracted from the left N pixel columns of input images. Then, the upper vowels and FC are removed based on this extracted information. Finally, the initial position is detected using the projection method. At the same time, the number of text-lines is obtained.
2. Estimating the baseline: Based on the above detected initial position, the baseline is estimated by dynamic tracing the pixel points from left to right direction. During dynamic tracing process, the context information between pixels is used to solve the baseline estimation of the curve text-lines.
3. Touching area detection and segmentation: For two consecutive baselines, the image between two baselines is divided into several patches. The patch-based touching area is detected based on the prior information and are further cut by finding the minimum width stroke.

3.1 Baseline Initial Position Detection

The objective of the baseline initial position detection is to find the estimated baseline initial position information and the text-line number of the document. The image for detecting initial position is the left N pixel columns of input image. Denote the image as *image A* (see Fig. 4(a)). Assuming the text-lines in *image A* are approximately parallel. Due to the touching strokes between consecutive text-lines (see Fig. 1), it will be useful to remove the vowels and FCs before finding the position. Figure 4 shows the steps of estimating the initial position.

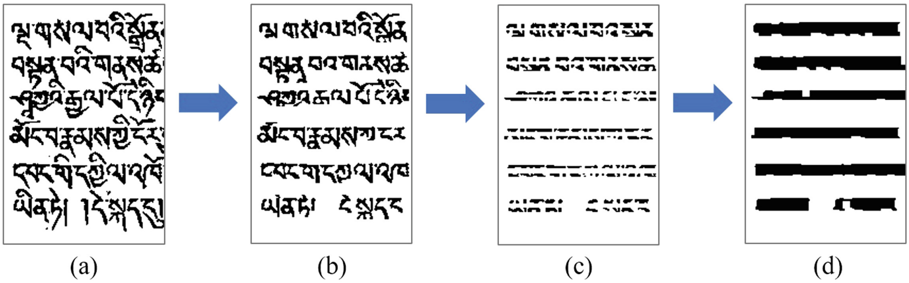

Fig. 4. The step for detecting the initial position of the input image. (a) The image for detecting the initial position. (*image A*) (b) The image that extracted by template matching. (*image B*) (c) The image that has been pruned salient strokes. (*image C*) (d) The image that has been performed closing operation. (*image D*)

The baseline initial position detection method includes three steps. The first step is to establish the detection template database and generate text blocks by template matching. Firstly, select randomly several input images and split them into patches by a sliding window. Secondly, the patches which contain the baseline of the syllable at the top are picked up manually as the detection templates (see Fig. 5). The detection template database with 80 patches is constructed. Thirdly, the text blocks are generating by the template matching method. The features of these patches are extracted using Principal Component Analysis (PCA [7]). *image A* with the same size of sliding window are obtained. Calculate the similarity between the patches in the sliding window and the templates using the computed PCA model. The similarity is calculated as follows:

$$similarity(x, y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{1}$$

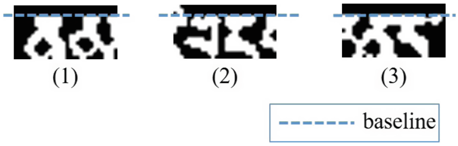

Fig. 5. Samples of the detection templates

Keep the patch if the similarity greater than a predefined threshold. After that, we can obtain a new image which has been removed FCs and vowels, denote it as *image B* (see Fig. 4(b)).

Secondly, pruning salient strokes. In *image B*, FC and most vowels in Tibetan syllables have been removed, although few vowels are remained incorrectly (see Fig. 6). Refer the block covering method [8], the *image B* is cut into text blocks based on horizontal projection profile. In each text block, recalculating the value of the projection profile by:

$$len(x) = \begin{cases} x \; x > \frac{\max(block)}{2} \\ 0 \; other \end{cases} \tag{2}$$

Max (block) is the max profile value in each text block. Set the pixels of the row to 0 whose profile value equals to 0 in *image B*, denote it as *image C* (see Fig. 4(c)).

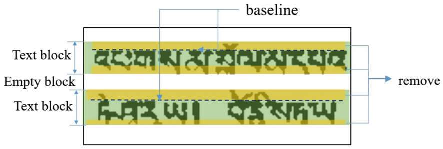

Fig. 6. Pruning salient strokes in text blocks of *image B*

Finally, we obtain *image D* (see Fig. 4(d)) by closing operation on *image C*. The initial position of the baseline is the upper boundary of each text block in *image D*. The number of text-lines equals to that of the text-blocks.

3.2 Estimating the Baseline

In this stage, we need to estimate the baseline. Because the baseline of document image is not strictly horizontal, the baselines are established from left to right by dynamic tracing within the pixel points of each line. We assume the value of foreground pixel as 1 and background pixel as 0. Based on the initial position of the baseline, in each line from left to right, we pick up the five key points every N pixels based on the previous key tracing point, which are P, U1, U2, D1 and D2 for picking up the key tracing point. U1 and U2 are the point that lie 1 and 2 pixels above P. D1 and D2 are the point that lie 1 and 2 pixels below P. Regarding the baseline initial position as the starting key tracing point. The next key tracing points are selected by the following rules:

- If the values of P, U1, U2, D1 and D2 are the same, P is considered as a back-ground point. (see Fig. 7(a)). Select P as the next key tracing point.
- If the value of P equals 0 and D1 or D2 equals 1, P is considered as a point that lies above the baseline. D1 is selected as the next key tracing point (see Fig. 7(b)).
- If the value of P equals 1 and U1 or U2 equals 1, P is considered as a point that lies under the baseline, U1 is selected as the next key tracing point (see Fig. 7(c)).
- Else, P is selected as the next key tracing point.

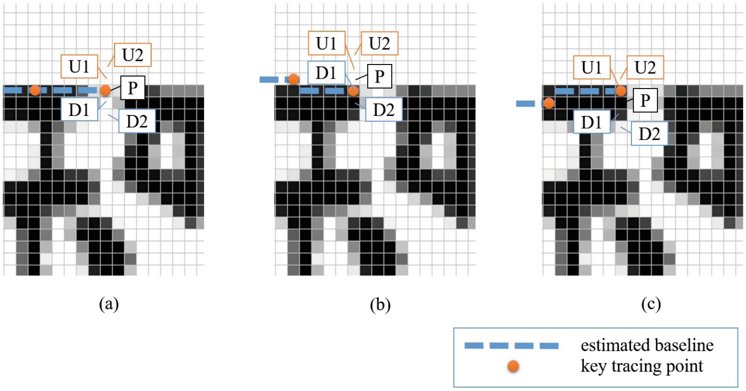

Fig. 7. Examples of selecting key tracing point. (a) Point P is the background point. Select P as the next key tracing point. (b) Point P lies above the baseline. Select D1 as the next key tracing point. (c) Point P lies under the baseline. Select U1 as the next key tracing point.

The baselines are estimated by joining the tracing key point together horizontally. Then, the input image is divided into several strips base on these baselines.

3.3 Touching Area Detection and Segmentation

The document image has been split into some horizontal strips based on the estimated baselines. However, the vowels in the syllable cannot be properly segment. In this stage, we present a method to locate the touching area and cut it by finding the minimum width stroke. Firstly, the strips are divided into a number of patches with a fixed step. In accordance with the information of connected components, these patches are classified into two classes by finding the bounding box (BB) of the connected component (CC). If there exists a BB has the same height with the patch, the patch is labelled as a touching patch (Fig. 8(b)), else label the patch as a normal patch (Fig. 8(a)). The segmentation for the normal patch can be easily found: if the CCs touch with the baseline of the next text-line, the CCs belong to the next text-line.

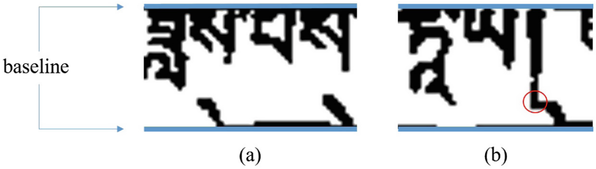

Fig. 8. Patch types: (a) normal patch (b) touching patch

For the touching patches, by observing the composition of Tibetan syllables, we can conclude that the touching strokes generally exist in the bottom 1/2 part of them and the stroke width of the touching area is short. With the method proposed by Epshtein et al. [9], we transform the patch by calculating the width of the stroke horizontally and vertically. The cut position of the patch is computed based on the projection profile of the transformed patch to Y-axis with the following criteria:

- If there is one minimum, select the corresponding row in the patch as the position for segment.
- If there are a few minima, select the corresponding row which is the nearest to the middle of bottom half of the patch as the position for segment.

After located the position for cutting, we need to verify if the position could segment the patch properly. If the value of cutting position in projection profile is greater than a predefined threshold, it shows that touching area is complex. A further research is needed to solve this problem. In this paper, these patches are cut with a simple method: if the text strokes lie under the cutting position, they belong to the next text-line.

4 Experiments and Results

The experimental dataset are provided by Qinghai Nationalities University. Limited by a variety of factors, we only acquired 120 high quality images for experiment. The method presented in this paper is implemented in python. The PCA model established base on the project of scikit-learn [10]. We use the library of opencv-python for image input and output and pre-processing, and scikit-image [11] for processing the connected components and bonding box.

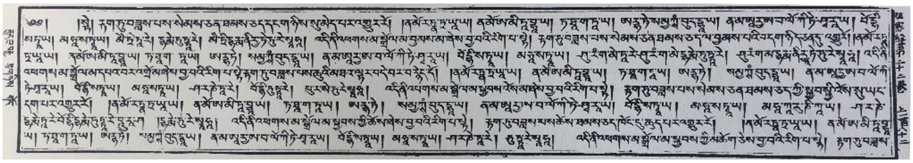

Fig. 9. The original Tibetan document image

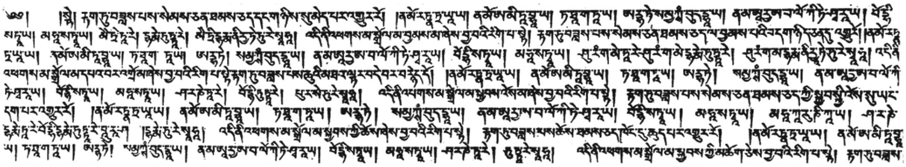

Fig. 10. The extracted image for text-line segmentation

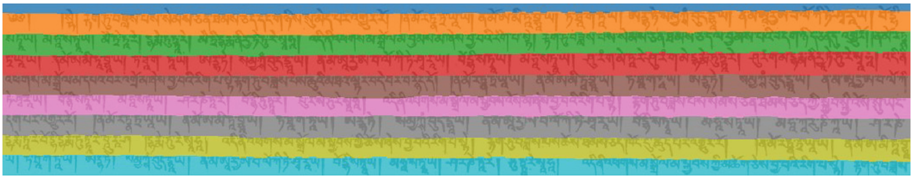

Fig. 11. The input image is divided into strips based on the baselines

Figure 9 is an original historical Tibetan document image. The method for layout analysis and text area extracting is proposed by Zhang et al. [6]. Figure 10 is the extracted text image that has been normalized to a width of 1300px with a scaled height and threshold for further processing. We extract 150 pixels columns from the input image for detecting baseline initial position. The width of the sliding window is 30px and the height is 15px. Two document samples are selected to establish the PCA model. They are split into patches with the same size of the sliding window. In the similarity calculation, 10-d is suitable. Figure 11 shows the text-line strips. The main part of a syllable has been segmented in each strip while the vowels are not properly labelled. It needs for further processing before extracting. Figure 12 gives the result, different with Fig. 11, the vowels are labelled correctly. To evaluate the performance of the presented method, we use the same evaluation method as that used in ICDAR2013 Handwritten Segmentation Contest [12]. It is based on counting the number of matches between

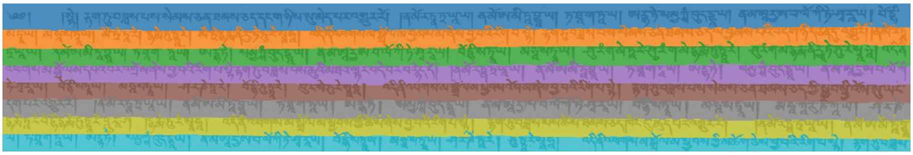

Fig. 12. Patch types: (a) normal patch (b) touching patch

the entities detected by the algorithm and the entities in the ground truth. The evaluation method uses a MatchScore (3) table to detect the matches whose values are computed by the intersection of the on pixel set of the result and the ground truth.

$$MatchScore(i,j) = \frac{T(G_j \cap R_i \cap I)}{T((G_j \cup R_i) \cap I)} \tag{3}$$

Let I be the set of all images points, G_j the set of all points inside the ground truth region, R_i the set of all points inside the i result region and $T(s)$ is a function that counts the points of set s. A region pair is considered as a one-to-one match only if the matching score is equal to or above a threshold. Let N be count of ground-truth elements, M be the count of result elements, and $o2o$ be the number of one-to-one matches, the detection rate (DR) and the recognition accuracy (RA) are defined as follows:

$$DR = \frac{o2o}{N}, RA = \frac{o2o}{M} \tag{4}$$

A performance metric FM can be extracted if we combine the values of detection rate (DR) and recognition accuracy (RA):

$$FM = \frac{2\, DR\, RA}{DR + RA} \tag{5}$$

Table 2 shows the performance of our method. Comparing with performance of project-based method in Table 1, our method has a considerable improvement in accuracy. As the stage of processing touching area is not very exactly, the

Table 1. The performance with project-based method

T(s)	M	N	o2o	DR	FM
0.90	769	770	552	67.79%	67.83%
0.91	769	770	446	57.92%	57.96%
0.92	769	770	360	46.75%	46.78%
0.93	769	770	269	34.94%	34.96%
0.94	769	770	197	25.58%	25.60%
0.95	769	770	135	17.53%	17.54%

Table 2. The performance with our method

T(s)	M	N	o2o	DR	FM
0.90	770	770	750	97.40%	97.40%
0.91	770	770	740	96.10%	96.10%
0.92	770	770	723	93.90%	93.90%
0.93	770	770	691	89.74%	89.74%
0.94	770	770	641	83.25%	83.25%
0.95	770	770	553	71.82%	71.82%

presented method does not perform so well when $T(s)$ increases. We will improve the accuracy in a further research.

5 Conclusion

This paper presents a text line segmentation method for historical Tibetan documents. The initial baseline position is obtained by template matching, pruning the salient strokes and closing operation. To solve the text-line segmentation of the curve text-lines, the baseline is estimated by joining together a number of key tracing points which are located by the context information. At the end, the touching area is detected and cut by analysing the patches between the two baselines. The patches are classified into touching patches and normal patches according to the decision if they have CCs that have the same height with the patches. The touching patches are cut by finding the minimum stroke width. However, at the stage of processing the touching area, the method presented by this paper is not suitable for complex areas (Fig. 1(b)). In the future research, we will find a better way to locate and remove the vowels. In addition, we need to adapt this method on the Tibetan document with non-horizontal direction of text-line.

Acknowledgments. This work was supported by the Science and Technology Project of Qinghai Province (no. 2016-ZJ-Y04) and the Basic Research Project of Qinghai Province (no. 2016-ZJ-740). The authors would like to thank Qilong Sun, the Department of Computer Science, Qinghai Nationalities University for providing the experimental dataset of historical Tibetan document images.

References

1. 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)
2. Manmatha, R., Rothfeder, J.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1212–1225 (2005)

3. Louloudis, G., et al.: Text line detection in handwritten documents. Pattern Recogn. **41**(12), 3758–3772 (2008)
4. Oztop, E., et al.: Repulsive attractive network for baseline extraction on document images. Sig. Process. **75**(1), 1–10 (1999)
5. Huang, H., Da, F.: General structure based collation of Tibetan syllables. J. Inf. Comput. **6**(5), 1693–1703 (2010)
6. Zhang, X., Duan, L., Ma, L.-L.: Text extraction for historical Tibetan document images based on connected component analysis and corner point detection. In: Yang, J., et al. (eds.) CCCV 2017, Part I. CCIS, vol. 771, pp. 545–555. Springer, Heidelberg (2017)
7. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
8. Zahour, A., et al.: Overlapping and multi-touching text-line segmentation by block covering analysis. Pattern Anal. Appl. **12**(4), 335–351 (2009)
9. Epshtein, B., Ofek, E., Wexler, Y.: Stroke width transform. In: Computer Vision and Pattern Recognition (2010)
10. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(8), 2825–2830 (2011)
11. van der Walt, S., et al.: Scikit-image: image processing in Python. PeerJ **2**, e453 (2014)
12. Stamatopoulos, N., et al.: ICDAR 2013 handwriting segmentation contest. In: 2013 12th International Conference Document Analysis and Recognition (ICDAR), pp. 1402–1406. IEEE press (2013)