Text Extraction for Historical Tibetan Document Images Based on Connected Component Analysis and Corner Point Detection

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

[1] Faculty of Information Technology, Beijing University of Technology, Beijing, China
zhangxiqun122@163.com, ljduan@bjut.edu.cn
[2] Beijing Key Laboratory of Trusted Computing, Beijing, China
[3] Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data, Beijing, China
[4] Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
{longlong,wujian}@iscas.ac.cn

Abstract. In this paper, we present a text extraction method for historical Tibetan document images. The task of text extraction is considered as text area detection and location problem. Firstly, the historical Tibetan document image is preprocessed to correct imbalanced illumination, tilt and noises, then get the binary image. Secondly, the regions of interest in historical Tibetan documents are divided into three categories using connected components. The images are divided equally into grids and the grids are filtered by the information of the categories of CCs and corner point density. The remaining grids are used to compute vertical and horizontal grid projections. Thirdly, by analyzing the projections, the approximate location of the text area can be detected. Finally, the text area is extracted accurately by correcting the bounding box of the approximate text area. Experiments on the dataset of historical Tibetan document images demonstrate the effectiveness of the proposed method.

Keywords: Historical Tibetan document · Text extraction
Connected components · Corner point

1 Introduction

Nowadays, some historical Tibetan documents have been digitized and available to the public. Most of them are stored in the form of images and digitized by manually entering corresponding texts into the computer. The historical Tibetan documents stored in the form of images require a lot of storage space. The scanned image cannot be edited, so the research and the use of these images are restricted. Except of the relevant researchers, few people will read the

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 545–555, 2017.
https://doi.org/10.1007/978-981-10-7302-1_45

scanned image of historical documents. If there is an efficient document recognition method to digitize automatically historical Tibetan documents, it is very meaningful for the inheritance and protection of the Tibetan traditional culture.

Text extraction is an important initial step in the automatic digitization of historical documents. In the past decades, researchers have proposed some methods to extract texts from the document images. These methods rely on connected components analysis [7], corner point density [11,13], and feature extraction [1,2,10]. In addition, some researchers use edge detection [6] and similarity matching [5] methods. In the field of text extraction, it is usually impossible to use the same method to process document images with different layout structure and text features. Our goal is to develop a method to extract texts from historical Tibetan document images. The vast majority of Tibetan historical documents are written on Tibetan papers, which are handmade in traditional ways. The layout of historical Tibetan document is irregular and complex. The lines of some frames have double layers, even the same frame will be broken into several parts. Some text areas are surrounded by multiple layers of such frames. Due to the compositional characteristics of Tibetan, the consecutive text lines of Tibetan is often touching or overlapping. The text is also attached to non-text parts. All of the above mentioned document features bring difficulties for text extraction. Figure 1 gives one example of historical Tibetan document images.

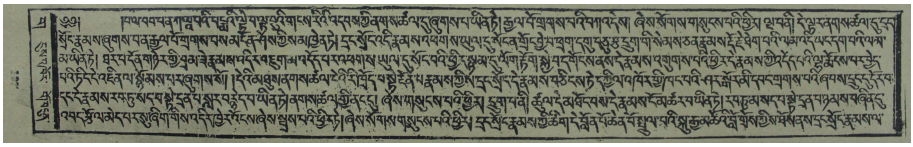


Fig. 1. Historical Tibetan document image.

We transform the text extraction problem into text area detection and location problem. Firstly, the historical Tibetan document image is preprocessed to eliminate the effects of imbalanced illumination, tilt and noises. After this, the image is binarized using the Otsu algorithm. The seed-filling algorithm is used to detect the CCs, and CCs are divided into three categories (*text*, *frame*, *line*) according to the area threshold and the width-height or height-width ratio of CCs. Then, the image is divided equally into $N*N$ grids (non-overlapping) and the grids are filtered by the information of the categories of CCs and the corner point density. The remaining grids are used to compute vertical and horizontal grid projections. By analyzing the projections, we can obtain the approximate location of the text area. Finally, the text area is extracted accurately by correcting the bounding box of the approximate text area.

The rest of the paper is organized as follows. Section 2 presents an overview of the related work. Section 3 describes the proposed method in detail. Section 4 shows our experimental results. Finally, the conclusions are given in Sect. 5.

2 Related Work

There are some works related to text extraction tasks. Researchers used different methods, such as traditional methods, machine-learning methods, etc., to extract texts from different types of documents.

Traditional segmentation methods can be grouped into top-down, bottom-up and hybrid methods. In general, extracting text from the complex layout document is mainly based on the hybrid method. AGORA project [7] developed the user-driven approach based on the hybrid method to perform layout analysis for historical printed books. Their segmentation algorithm includes two maps: a shape map and a background map. The shape map is formed from the bounding box of the CCs in the page. The background map provides information about white areas corresponding to block separations in the page. Their method used the background map of the images to highlight the separation between blocks and the shape map. Then, they segment the image with the predefined rules and the information provided by these two representations. In addition, Winder [9] modified the RAST and Voronoi segmentation algorithms in OCRopus to process mixed content layouts at a variety of resolutions, and make the digitization of standard format historical documents by low-budget organizations feasible. Yu [12] improved the bottom-up page segmentation method based on the connected region of printed newspapers. In Chinese ancient book "Imperial Collection of Four" digitization project, Jiang [14] employed a hybrid method, associated with artificial correction, to analyze the document layout. Singh [8] analyzed the layout of Indian newspapers with the hybrid method.

Machine-learning method regards text extraction as a classification problem. These methods extract the features from different parts of document images to train the classifier, the classifier classifies image regions into text, illustration and so on. Chen [3] developed an unsupervised feature learning method for page segmentation of historical documents with color images. They used the convolutional autoencoder to learn features directly from pixel intensity values. Aiming at reducing the computation time, they present a superpixel-based method [2] to replace the original pixel-based method. Training a support vector machine to classify superpixels based on these features. Bukhari [1] extracted the relevant features in a connected-component level to train a multilayer perceptron. A voting scheme is then applied to refine the resulting segmentation and produce the final classification. Xiao [10] proposed a method to extract text from ancient Yi character documents. They combined edge and texture features to represent the characteristics of text in ancient Yi character documents accurately and adopt the GBDT (Gradient Boost Descent Tree) learning theory to design a classifier to classify text and non-text pixels.

Other methods, such as corner point detection, contour detection, can also extract texts from document images. Zeng [13] proposed a filtering method based on the Harris corner point detection. The algorithm can filter the background, which contains text image commendably of printed documents. Yadav [11] designed a very simple technique based on FAST key points to extract texts from document images. The image is divided into blocks and the blocks

including more points are classified as text blocks. Then, the connectivity of blocks is checked to group and obtain complete text blocks. In order to segment text from degraded historical Indus script images, Kavitha [6] proposed a new combination of Sobel and Laplacian for enhancing degraded low contrast pixels to generate skeletons for text components. The component that gives a less number of branches is considered as a text cluster because text components usually has fewer branches compared to non-texts in historical Indus script images. Ha [4] proposed an adaptive over-split and merge algorithm for page segmentation of printed documents. Firstly, the document image is over-splitted into text blocks or text lines. Then, these text blocks or text lines are merged into text regions using a new adaptive threshold method. The local context analysis method used a set of text line separators to split homogeneous text regions with similar font size and further merged text blocks into paragraphs.

3 System Description

In order to extract texts accurately from historical Tibetan documents, we propose a text extraction method based on connected component analysis and corner point detection. Figure 2 gives the workflow of our method.

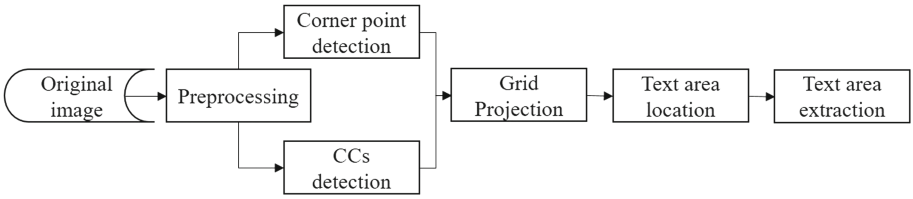


Fig. 2. Text extraction workflow.

3.1 Preprocessing

The preprocessing is to correct imbalanced illumination, tilt, noises. If these problems are not solved properly, will cause unnecessary trouble and affect the subsequent operations.

In order to eliminate the adverse effects mentioned above, the images are preprocessed by the following methods. In order to reduce the amount of calculation, we normalize the original image into a fixed size. The gamma correction algorithm is used to correct illumination. A tilt correction algorithm based on the Hough transform is used to correct the skew of historical Tibetan document images. Finally, the images are transformed into the binary image using the Otsu algorithm.

3.2 Grid Projection

The grid projections are used to locate the approximate text area and highlight the gap between the different parts in the historical Tibetan documents. From [11,13], we can know that text regions have more corner points than non-text regions, which is confirmed by our experiments on the historical Tibetan document images. Inspired by the use of key point density from the grid filter method [11], we combine the CCs classification information with the corner point density to filter the grid.

The CCs in the binary image are detected by the seed-filling algorithm. The CCs are classified using some priori information. By observing the document images, we can find that the frame is larger in area than the text and the width and height ratio or height and width ratio (w/h or h/w ratio) of the line has a greater difference from other categories. The CCs are classified according to the following rules. The area of the image is denoted as S.

- Rule 1: If the area of CC is bigger than $a*S$, where a is the threshold of frame, the CC is classified as frame.
- Rule 2: If the w/h or h/w ratio of CC is smaller than the ratio of R, the CC is classified as a line.
- Rule 3: If the CC does not match the above rules, it is classified as text.

All CC matches each rule in order. The classification labels of CCs are stored in an image named *labelImg*, where all pixel values of CC are equal to its label value. The Harris algorithm detects corner points in the binary image. The image is divided equally into $N*N$ grids, and the number of corner points in each grid is calculated. The maximum number of corner points ($Nmax$) in all grids is recorded. The grids are filtered with the following steps.

- Step 1: If the number of corner points of the grid is less than $c*Nmax$ (c is the threshold of grid filtering), the grid will be deleted directly. Otherwise, do the following steps.
- Step 2: If the grid contains more than two non-text classes in the corresponding location of *labelImg*, the grid is isolated, or the grid is an edge grid that only contains non-text class. The grid is also deleted. Because we thought that these grids are located in a corner-dense non-text area.
- Step 3: If the grid contains text and non-text classes in the corresponding location of *labelImg*, the text part of the grid will be preserved.

Filtered grids using the above rules, gaps between different parts are highlighted. The remaining grids are used to compute vertical and horizontal grid projections.

3.3 Text Area Location

The approximate text area will be located by analyzing the grid projections. The projections are analyzed with the following steps.

- Step 1: Search the change points from left to right according to the vertical grid projection. The two adjacent change points from zero to non-zero and from non-zero to zero are considered as the horizontal start and end positions of an area.
- Step 2: For the horizontal grid projection, the first non-zero point at both ends of the horizontal grid projection is considered as the vertical start and end positions.
- Step 3: The start and end positions of the horizontal and vertical directions are used to search for the first non-text points as the areas pixel point of the bounding box from the inside to the outside of the four edges in the corresponding area of *labelImg*. The pixel points of the bounding box near the broken frames, which are misclassified to texts, are filled with zero-to-nonzero change points. These points can be obtained by searching for the corresponding position of its nearby pixel point of the bounding box. The breakpoints of the broken frames are filled with its nearest bounding boxes pixel point.

Repeating the above steps until the end of vertical projection is searched. Approximate text areas are located, and their bounding boxes are also obtained.

3.4 Text Area Extraction

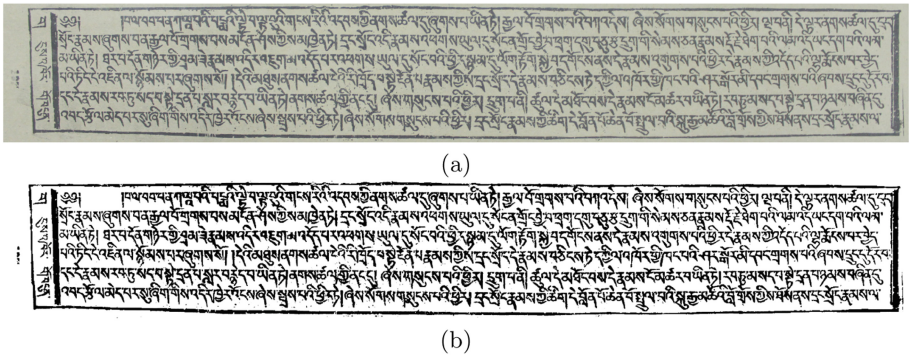
Text area are extracted accurately by correcting the bounding box of the approximate text area. The bounding box is corrected by the following strategy. Take the upper boundary as an example, the boundary points are selected to calculate the average of their horizontal ordinate (denoted as AO), and if any point's horizontal ordinate greater than AO and the difference between its horizontal ordinate and its previous points horizontal ordinate is greater than X, and its horizontal ordinate will be replaced by the previous points horizontal ordinate. Using the same method, continue to correct the horizontal ordinate of the bounding box. From the previous steps, we can know that some points from four corners of bounding box for the approximate text area are not found. These points are searched based on the existing detected neighbor points. After the above operations, the texts attached to the boundary are removed and the text areas are accurately extracted.

4 Experiments

This text extraction method is tested on manuscripts taken from "*The Complete Works of the Panchen Lama*", provided by the Department of Computer Science, Qinghai Nationalities University. The layout structure of the collected images is irregular and complex. Due to the restriction of hardware, the images have serious imbalanced illumination and the quality of the images is poor. The color of documents is not uniform. The size of the images is not uniform, but within a certain range. We use the collected historical Tibetan document images to conduct the experiment, and the collected dataset contains 360 images.

4.1 Preprocessing

In order to obtain the preprocessed images from historical Tibetan documents, several methods of illumination equalization and binarization are experimented. We found that the combination of the gamma correction algorithm and the Otsu algorithm could obtain a better binarization result. Figure 1 shows the original image. The gamma value of gamma correction is set to 0.4. Figure 3(a) shows the result of elimination imbalanced illumination. Figure 3(a) is transformed into grayscale image, the skew of the image is corrected by a tilt correction algorithm based on the Hough transform. Figure 3(b) is the binary image of the grayscale image using the Otsu algorithm.



(a)

(b)

Fig. 3. Preprocessing. (a) Gamma correct result. (b) Binary image.

4.2 Grid Projection

In order to get the vertical and horizontal projections, we perform the operations according to Sect. 3.2. Firstly, the threshold of a, R are set to 1/4 and 0.05, respectively. The classification results of CCs are shown in Fig. 4(a). The regions of frame, line and text are labeled separately using red, green and black color. We can see that the broken parts of the frame are misclassified as other classes. Some sticky texts were misclassified as non-text classes. These misclassifications of sticky texts will be corrected in the following operations. Figure 4(b) shows the corner point detection results. If the grids are filtered using the method of [11], the text area could not be located accurately from historical Tibetan documents. Figure 4(c) shows the results of grid filtering, which are filtered according to the method described in Sect. 3.2. It is not difficult to find that the grids of Fig. 4(c) can not only locate the text area, but also highlight the gap between the different text areas.

The vertical and horizontal grid projections are computed based on the remaining grids. Figure 5(a) and (b) shows the vertical and horizontal projections, respectively. It is possible to determine the approximate text area by analyzing the grid projections.

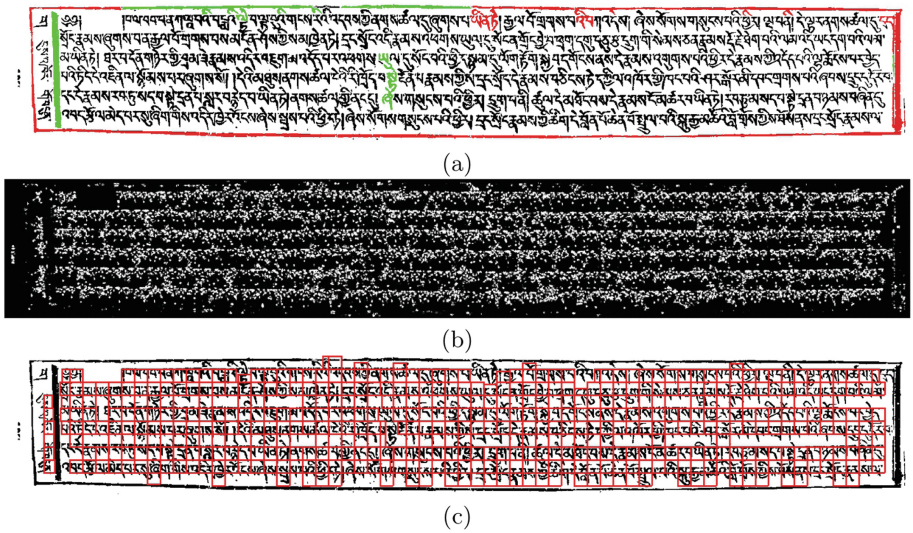


Fig. 4. CCs classification and grid filter. (a) CCs classification results. (b) Corner point detection result. (c) The result of filtered grids. (Color figure online)

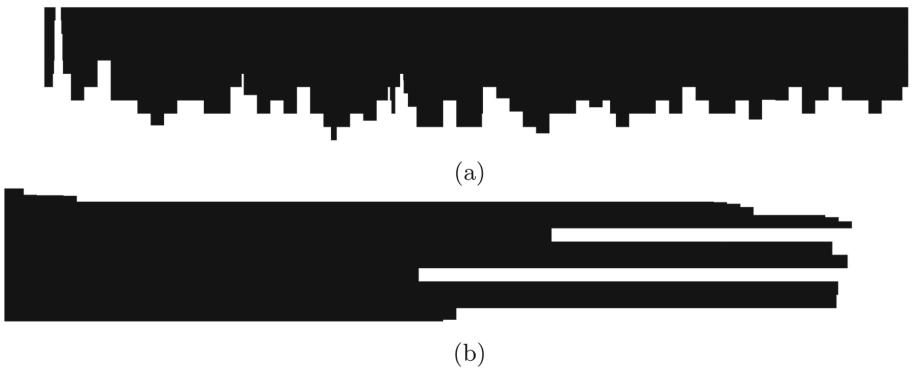


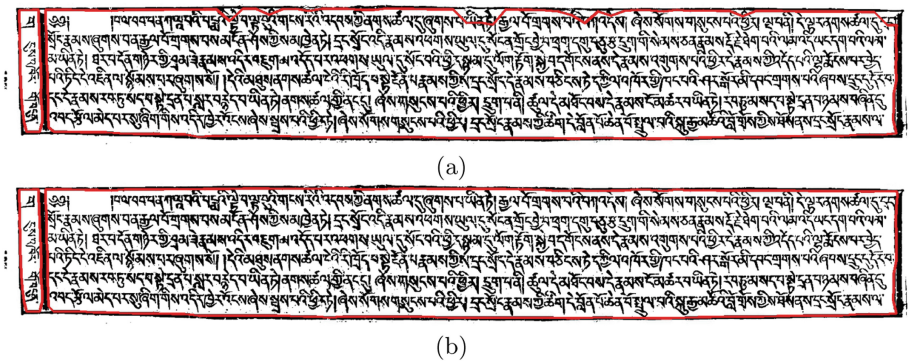
Fig. 5. Grid projections. (a) Vertical projection. (b) Horizontal projection.

4.3 Text Area Location

As we can see from the projections, there is a clear gap between different parts. So the approximate text area can be obtained by the method described in Sect. 3.3. Figure 6(a) shows the result of approximate text area location. The red bounding box is the boundary of the text area.

4.4 Text Area Extraction

Based on the result of previous step, the strategy, described in Sect. 3.4, is used to correct the bounding box of the approximate text area. The threshold of X



(a)

(b)

Fig. 6. The bounding box and final result of text area. (a) The bounding box of the approximate text area. (b) The result of text extraction. (Color figure online)

is set to 3. The results are shown in Fig. 6(b). The points of the extracted text areas bounding box are labeled as red color.

4.5 Experimental Results

As we all know, the text extraction algorithm has a strong correlation with the specific document layout. Usually a layout type of document corresponds to a specific text extraction algorithm. In fact, we have also tried some of the methods in other paper, most of the methods are not suitable for historical Tibetan documents. And there are no papers about extracting text from historical Tibetan document or similar layout historical document. Therefore, only the method proposed in this paper is evaluated. We adopt the F-measure metric, which combines precision with recall values into a single scalar representative, to evaluate the text extraction accuracy. Precision and recall are estimated according to Eqs. 1 and 2, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where True-Positive(TP), False-Positive(FP) and False-Negative(FN) with respect to text area, are defined as following:

TP: Number of the extracted text area correctly
FP: Number of the misclassified text area
FN: Number of the undetected text area

Once we have the precision and recall values. F-measure is calculated according to Eq. 3.

$$F - Measure = \frac{\left(1 + \beta^2\right) \cdot Precision \cdot Recall}{\left(\beta^2 \cdot Recall\right) + Precision} \tag{3}$$

The β is set to 1, it means that precision and recall are equally important in the F-measure estimation.

In order to compare the rigorous evaluation of the experimental results, we labeled ground truth of text areas on our dataset, and the use of pixel accuracy to evaluate our experimental results. If the pixel accuracy of the extracted area bigger than the threshold *acc*, we thought it is a text area. The *acc* for larger text areas and smaller text areas are set to 95% and 90%, respectively. From Table 1, we can see that the F-measure reaches 85.60%, but the recall reaches 98.58%, while the precision is only 75.64%. By analyzing the experimental results, we found that main reasons of lower precision are led by the complexity of the layout and the document image quality. In our dataset, lines of the frame are rough and irregular. Due to the quality of the image and other reasons, these lines are usually broken into shorter lines or even dot sequences, so these broken lines and dots are often extracted as the text and effects the precision of text extraction.

Table 1. Performance evaluation

Precision(%)	Recall(%)	F-Measure(%)
75.64	98.58	85.60

5 Conclusion

In this paper, we presented an efficient method to extract text areas from historical Tibetan documents. It makes a classification of CCs using the priori rules. Corner points are detected in binary images, and images are equally divided into grids. Based on the classification information of CCs and corner points, the grids are filtered by the predefined rules. By analyzing vertical and horizontal projections of remaining grids, the approximate text area is located, and the bounding box of the text area is searched. The text area is extracted accurately in the approximate area through the bounding box correction strategies. The experimental results verify the effectiveness of our method.

Our future work will focus on improving the adaptability of the algorithm. As we all know, different historical documents have different layout structures and characteristics. Our method needs to adapt to historical Tibetan documents with different layout structures. Enhancing the performance of the text extraction is also the direction of our future efforts.

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. Bukhari, S.S., Breuel, T.M., Asi, A., El-Sana, J.: Layout analysis for arabic historical document images using machine learning. In: 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 639–644. IEEE (2012)
2. Chen, K., Liu, C.L., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation for historical document images based on superpixel classification with unsupervised feature learning. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 299–304. IEEE (2016)
3. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1011–1015. IEEE (2015)
4. Dai-Ton, H., Duc-Dung, N., Duc-Hieu, L.: An adaptive over-split and merge algorithm for page segmentation. Pattern Recognit. Lett. **80**, 137–143 (2016)
5. Fu, H., Liu, X., Jia, Y.: Text extraction based on maximum-minimum similarity training method (in Chinese). J. Softw. **19**(3), 621–629 (2008)
6. Kavitha, A., Shivakumara, P., Kumar, G., Lu, T.: Text segmentation in degraded historical document images. Egypt. Inf. J. **17**(2), 189–197 (2016)
7. Ramel, J.Y., Leriche, S., Demonet, M., Busson, S.: User-driven page layout analysis of historical printed books. Int. J. Doc. Anal. Recognit. **9**(2), 243–261 (2007)
8. Singh, V., Kumar, B.: Document layout analysis for indian newspapers using contour based symbiotic approach. In: 2014 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4. IEEE (2014)
9. Winder, A., Andersen, T., Smith, E.H.B.: Extending page segmentation algorithms for mixed-layout document processing. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 1245–1249. IEEE (2011)
10. Xiao, R.: Research on the Method of Extracting Ancient Yi Text from Complex Background (in Chinese). Ph.D. thesis, South-Center University for Nationalities (2011)
11. Yadav, V., Ragot, N.: Text extraction in document images: highlight on using corner points. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 281–286. IEEE (2016)
12. Yu, M., Guo, J., Wang, D., Yu, Y.: Improved page segmentation method based on connected domain (in Chinese). Comput. Eng. Appl. **49**(17), 195–198 (2013)
13. Zeng, F., Zhang, G., Jiang, J.: Text image with complex background filtering method based on Harris corner-point detection. J. Softw. **8**(8), 1827–1834 (2013)
14. Jiang, S.P.Z., Ma, Y.X.: Automatic document layout analysis system for the large scale Chinese antient books 'imperial collection of four' (in Chinese). J. Chin. Inf. Process. **17**(2), 14–20 (2000)