

Received January 23, 2022, accepted February 8, 2022, date of publication February 15, 2022, date of current version March 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3151886

# Character Detection and Segmentation of Historical Uchen Tibetan Documents in Complex Situations

CE ZHANG<sup>1,2</sup>, WEILAN WANG<sup>1</sup>, (Member, IEEE), HUAMING LIU<sup>3</sup>, GUOWEI ZHANG<sup>1</sup>, AND QIANG LIN<sup>4</sup>

<sup>1</sup>Key Laboratory of China's Ethnic Languages and Information Technology, Ministry of Education, Northwest Minzu University, Lanzhou 730030, China

<sup>2</sup>School of Artificial Intelligence, Chongqing University of Education, Chongqing 400065, China

<sup>3</sup>School of Computer and Information Engineering, Fuyang Normal University, Fuyang 236037, China

<sup>4</sup>Key Laboratory of Streaming Data Computing and Application, Northwest Minzu University, Lanzhou 730124, China

Corresponding author: Weilan Wang (wangweilan@xbmu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772430 and Grant 62166036, in part by the Program for Leading Talent of State Ethnic Affairs Commission (SEAC), in part by the Program for Innovative Research Team of SEAC ([2018]98), in part by the Gansu Provincial First-Class Discipline Program of Northwest Minzu University through the "Innovation Star" Project of Excellent Postgraduates of Gansu Provincial Department of Education under Grant 2021CXZX-663, in part by the Science and Technology Research Program of Chongqing Education Commission under Grant KJQN202101608, in part by the Research Program of Chongqing University of Education under Grant KY202118C, and in part by the Key Projects of Natural Science Research in Anhui Colleges and Universities under Grant KJ2020ZD46.

**ABSTRACT** Tibetan is a low-resource language, and Tibetan culture carried by historical Tibetan documents is an important part of Chinese civilization. The study of historical Tibetan documents is of great significance to the protection of Tibetan culture and the promotion of Chinese culture. Character segmentation is an important step in image analysis and recognition of historical Tibetan documents. However, the following three challenges prevent solving problems of character segmentation in historical Tibetan documents: 1) the text lines have different degrees of tilt and twist; 2) there are many complex situations such as overlapping, crossing, touching and breaking character strokes; and 3) these documents are written by different people with different stroke styles. To resolve these problems, we propose a character segmentation method based on key feature information for historical Tibetan documents. The proposed method consists of three parts: 1) projection and syllable point location information are used to shorten the text lines of historical Tibetan documents and establish a character block database; 2) the local baseline of the character block is detected by using the location information of syllable points or combined with horizontal projection and straight line detection, and the character block is divided into two areas above and below the baseline, and different segmentation methods are adopted; and 3) in view of the large difference in stroke styles, three stroke attribution distances are used to complete the attribution. The experimental results show that the method proposed in this paper can effectively solve the problem of character segmentation of historical Tibetan documents and achieve a better character segmentation effect, which also provides a reference for the relevant document character segmentation.

**INDEX TERMS** Historical Tibetan documents, local baseline detection, character detection, character segmentation, stroke attribution.

## I. INTRODUCTION

Historical Tibetan documents record research achievements in literature, history, philosophy, politics, economy, medicine,

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar<sup>1</sup>.

art and other fields. These documents are rich in content and have important reference value for the study of Chinese multiethnic culture. Due to long-time preservation, the materials and ink of historical Tibetan documents have faded to varying degrees, and may even appear damaged or unusable. The protection, development and utilization of historical

Tibetan documents has become an important topic in the field of ethnic language research. Tibetan is a low-resource language, and it is difficult to obtain a large amount of document data, especially for historical Tibetan documents, which has delayed relevant research. Research studies on historical Tibetan documents began in the 1980s. From 1991, Kojima *et al.* [1]–[3] studied the recognition of Tibetan documents in woodcuts. Since 2010, researchers have carried out relevant studies on image preprocessing [4], [5], layout analysis [6], [8], text line segmentation [9]–[12], character segmentation [13], [14], dataset construction [15], [16], character recognition [17], [18] and other aspects of historical Tibetan documents.

Character segmentation is an important part of the study of historical Tibetan documents. To date, there have been few achievements in character segmentation of historical Tibetan documents. Ngodrup and Zhao applied the dripping method to gravity compress the simulated water molecules above the touching characters and obtained the character segmentation path according to the stress of pixels. Zhao *et al.* proposed a character segmentation method based on feature point information to solve the problem of touching character segmentation. First, the foreground contour and skeleton were detected, and feature points and baseline information were extracted. Then, the support vector machine (SVM) classifier and distance rule were combined to remove the useless feature points near the endpoints of the top vowel and consonant skeletons, and all the candidate segmentation points were obtained. Finally, the characters were obtained by using the segmentation graph method.

In terms of character segmentation of documents in other languages, character projection, connected component analysis, character feature information and other commonly used character segmentation methods have been utilized. Zhou *et al.* [19] first used the projection method to perform rough segmentation and divided Chinese characters into touching and untouching categories. Then, the segmentation path was set with the rough segmentation statistics, and the optimal weighted segmentation path was obtained based on the shortest path idea to complete the segmentation of touching characters. Qi *et al.* [20] first used connected component analysis to obtain the initial character segmentation, then solved the attribution of Chinese characters in historical documents by establishing the hesitation fuzzy set, and finally analyzed the pixel jump to achieve the segmentation of touching and overlapping Chinese characters. Tian *et al.* [21] improved the method of Qi *et al.*, taking into account the problem of character undersegmentation and using the improved K-means to solve the problem of touching character segmentation. Zaw and War [22] proposed a method based on character connected component analysis to complete the segmentation of Myanmar touching characters. Thongkanhorn *et al.* [23] proposed a vertical and horizontal segmentation method based on a 4-direction depth-first search algorithm and completed the segmentation of Thai

characters. Tamhankar *et al.* [24] solved the segmentation problem of the text characters of Microsoft office document imaging (MODI) in historical cursive script by using the double-threshold criterion to minimize the segmentation error. Ali and Suresha [25] used character geometry and shape information to solve the problem of vertical segmentation of touching Arabic characters. Ullah *et al.* [26] detected the touching and segmentation of Arabic characters based on morphological methods and character feature information. The projection method and stroke width transform (SWT) were used to solve touching Japanese segmentation [27]. The drop fall algorithm [28] was initially applied to the segmentation of touching numbers, and then it was continuously improved and applied to the segmentation of touching numbers [29], [30] or touching verification code [31]. It was also applied to the segmentation of touching characters in mixed Korean and Chinese historical documents [32].

In addition, the method based on character oversegmentation and machine learning has been a popular segmentation method in recent years. The general idea of the segmentation method is to divide the touching characters into more characters and then attribute the oversegmented characters through various methods. Ji *et al.* [33] segmented historical Chinese documents by graph nodes and assigned the segmented characters according to the optimal path of the graph. Xu *et al.* [34], when solving the problem of touching and segmentation of handwritten Chinese characters, first used character characteristic information to obtain candidate segmentation paths and then used rules and learning-based filters to achieve touching and segmentation. Sahare and Dhok [35] first used the character characteristic information to segment the touching characters and then completed the character segmentation of multilingual Indian documents composed of Latin and Devanagari by using the graph distance. Gao *et al.* [36] proposed a character segmentation method based on fully convolutional neural networks (FCNs) to solve the complex problems of Chinese character touching and breaking, which extracts the spatial features of characters using convolutional neural networks. Two FCNs were used to extract features and form a score map. Finally, character features were used to adjust the exact segmentation points in the fraction graph to achieve character segmentation. Xie *et al.* [37] proposed a weakly supervised character segmentation method with recognition and guidance information in the attention region to solve the problem of touching and segmentation of historical Chinese documents and realized high-precision segmentation under strict cross and union ratios.

Although Ngodrup and Zhao studied the touching and overlapping character problems in the character segmentation of historical Wooden Tibetan documents, the number of character segmentations was small, and the correct rate of optimal segmentation was not high, so there is still space for further research. Zhao *et al.* studied the segmentation and recognition of touching characters in historical Tibetan documents, but this method did not involve the situation of

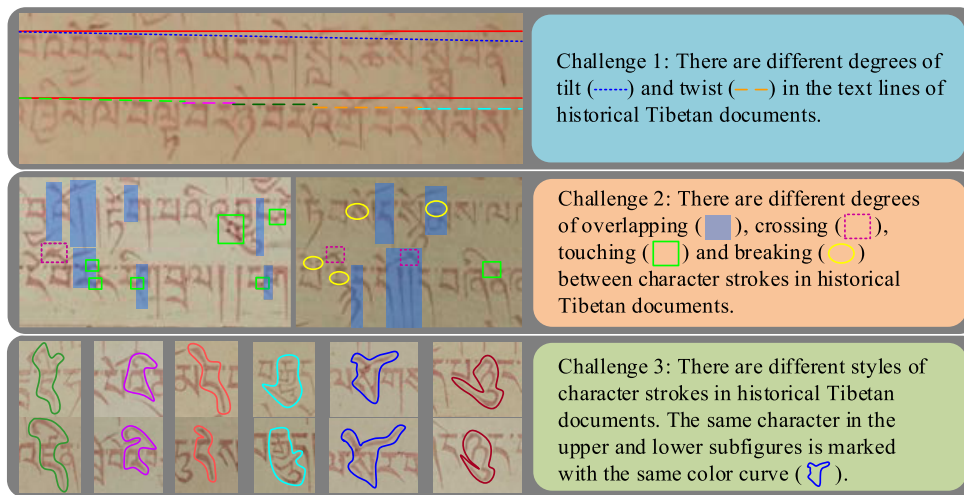


FIGURE 1. Three main challenges of character segmentation in historical Tibetan documents.

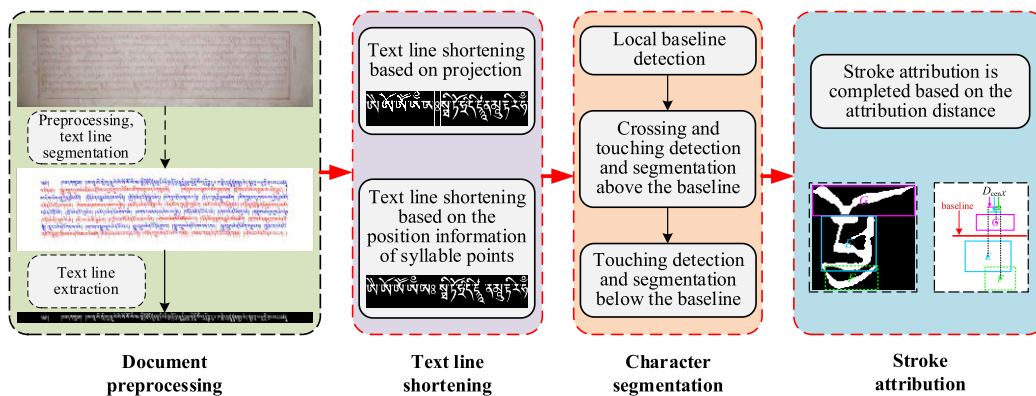


FIGURE 2. Character segmentation framework of historical Tibetan documents.

stroke crossing above the baseline, and the type of touching character was less. Although the methods mentioned above are not directly applicable to the segmentation of historical Tibetan documents, they provide inspiration for the work of this paper. At the same time, we solved the character segmentation problems of stroke overlapping, crossing, touching and breaking above the character baseline of historical Tibetan documents [38] in previous character segmentation work. However, the more complex character segmentation problem below the baseline has not been solved. Thus, character segmentation still faces three challenges (Figure 1):

*Challenge 1:* There are varying degrees of tilt and twist in the text lines of historical Tibetan documents. Historical Tibetan documents are printed from wooden plates, which are hand-carved, resulting in a certain degree of tilt and twist. In addition, there is a certain deviation in the placement of paper during the printing process, which further aggravates the tilt and twist of historical Tibetan documents. The tilt and twist of text lines cause serious interference in detecting the global baseline of text lines.

*Challenge 2:* There are many overlapping, crossing, touching and breaking strokes in historical Tibetan documents. Specific formation reasons are as follows: when carving the original printing plate, some vowel strokes above the character baseline span the left and right characters to create overlapping (vertical projection overlap), crossing and touching. When printing historical Tibetan documents, strokes below the character baseline touch the strokes of the left and right characters. There is considerable separation between the character strokes on the original printing plate, and the uneven color of historical documents during printing causes different degrees of broken strokes.

*Challenge 3:* Historical Tibetan documents are written by different people with different styles of strokes; even the same person’s writing results may not be the same at different times. The difference in stroke style is embodied in the size, position and shape of stroke, which seriously affects stroke attribution after stroke segmentation.

In view of the above character segmentation challenges, we propose a character segmentation method for historical

Uchen Tibetan documents based on key feature information (Figure 2). The proposed method consists of three parts:

First (solution for Challenge 1), a method of line shortening processing of historical Tibetan documents is proposed to reduce the impact of line tilt and twist on baseline detection. There are two methods to shorten the text line: the first method is to use the gap of the vertical projection of the text line for vertical segmentation; and the second method is based on the position coordinates of syllable points without overlapping above and below for vertical segmentation. The text line shortening results obtained in the two methods form two different character block databases. The text lines are shortened into character blocks of different widths, and the global baseline detection of text lines is transformed into local baseline detection of character blocks, which effectively solves the problem of inaccurate baseline detection caused by the tilt and twist of text lines.

Second (solution for Challenge 2), a character segmentation method based on key feature information is proposed to segment the characters above and below the baseline separately. Through the analysis of historical Tibetan document texts, it was found that strokes above the baseline are mainly crossing and touching, and strokes below the baseline are mainly touching and breaking, while there is more overlap between the strokes above and below the baseline. To solve the problem of stroke crossing and touching above the baseline, the method of multidirection and multipath crossing and touching is adopted, while the algorithm of stroke touching below the baseline is based on the position information of the “head” of the character.

Third (solution for Challenge 3), an attribution method based on three attribution distances for character strokes is proposed to attribute breaking strokes. We summarize the morphological features of characters in historical Tibetan documents by attributing breaking strokes and segmented strokes to complete character segmentation.

Experiments on historical Uchen Tibetan documents show that 1) our method only uses the key feature information of the characters without any postprocessing, and the segmentation process is simple; and 2) our method can effectively solve segmentation problems such as character overlapping, crossing, touching and breaking.

In summary, the main contributions of this work are described as follows:

1) We propose a text line shortening method based on the position information of syllable points without overlapping above and below. Using this method, the global baseline detection of text lines is transformed into a local baseline of character blocks, effectively solving the interference of text line tilt and twist on character segmentation.

2) We propose a character segmentation method based on key feature information, which effectively solves the segmentation problems of characters overlapping, crossing, touching and breaking in historical Uchen Tibetan documents. The proposed method can be applied to Tibetan character

segmentation in other application scenarios and can provide a reference for character segmentation in other languages.

3) We propose a character stroke attribution method based on three different attribution distances, which effectively solves the problem of the same character stroke morphological differences caused by writing at will.

## II. CHARACTER BLOCK DATABASE ESTABLISHMENT

According to the characteristics of text lines in historical Tibetan documents, we propose two shortening methods: shortening based on the projection gap of text lines and shortening based on the position information of syllables without overlapping above and below.

**Projection-based shortening method (Pro-SM):** shortening based on text line projection gap

Text lines are shortened to produce character blocks of varying widths. Through the analysis of the characteristics of text lines in historical Tibetan documents, it is found that there are different degrees of overlap between character strokes, which will form gaps with different spacings after vertical projection (Figure 3(a)).

**Syllable points-based shortening method (Syl-SM):** Shortening based on the position information of syllable points without overlapping above and below

Syllable points are an important part of the Tibetan language and a symbol of syllable division. Due to the character writing characteristics of historical Tibetan documents, there are different degrees of stroke overlap between characters, and it is difficult to ensure that each character block has a complete syllable through projection. In Figure 3(b), the syllable points in the circle of the solid line are above the syllable point with overlapping and the syllable points in the circle of the dotted line are below the syllables without overlap. We analyze the connected component of the syllable points without overlapping in a certain height range, and the left and right outer boundaries of the minimum outer rectangle of the syllable point are the start point and the end point of the text line segmentation.

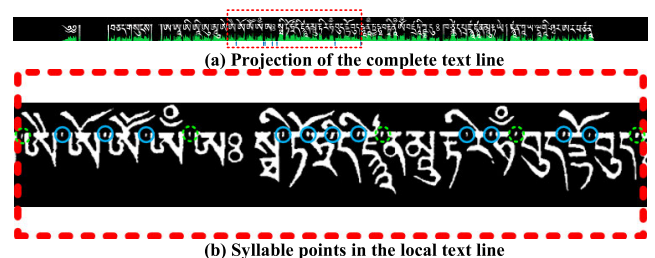


FIGURE 3. Text line projection and syllabic points.

## III. LOCAL BASELINE DETECTION

Local baselines are for character blocks. Local baselines are extremely important for locating strokes above and below baselines. Local baseline detection was achieved by using the



previously proposed baseline detection algorithm based on the location information of syllable points or combined with horizontal projection and straight line detection [38]. The local baseline detection of character blocks is classified as follows: 1) For character blocks with syllable points, the location information of syllable points can be obtained to realize baseline detection indirectly according to the characteristics that syllable points are at the same level as the baseline; 2) for character blocks without syllable points and strokes above the baseline, the distance between ordinates of the connected components of strokes is compared to achieve baseline detection; and 3) for character blocks without syllable points and with strokes above the baseline, the method combining horizontal projection and Hough straight line detection algorithm is used to realize baseline detection, and horizontal segmentation is carried out at the baseline position.

The local baseline position is used to divide the character block into two parts above and below the baseline so that the strokes above and below the baseline of the character can be located. In addition, dividing the character block at the local baseline can also solve the touching problem between the strokes above the baseline of the left and right characters (Figure 4).

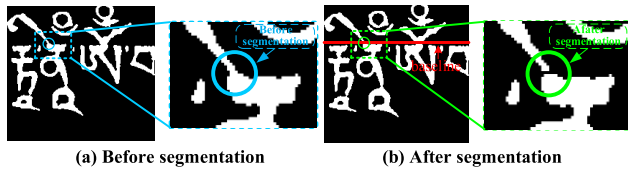


FIGURE 4. Character block before and after segmentation at the local baseline position.

#### IV. TOUCHING STROKE DETECTION AND SEGMENTATION

Crossing and touching between character strokes is the core problem of character segmentation, and crossing is a special case of touching. The strokes above the baseline exhibit both crossing and touching, and the strokes below the baseline are mainly touching. In view of the complexity of the crossing and touching problems between character strokes, we use the key feature information of the characters in historical Tibetan documents (Figure 5) to design different detection and segmentation methods for the strokes above and below the baseline.

##### A. STROKE CROSSING AND TOUCHING ABOVE THE BASELINE

For the detection and segmentation of stroke crossing and touching above the baseline, the improved template matching algorithm (Algorithm 1) and the multidirection and multipath segmentation algorithm (Algorithm 2) proposed in the character segmentation work of the previous stage are adopted. We summarized 14 types of stroke crossing and touching (Table 1).

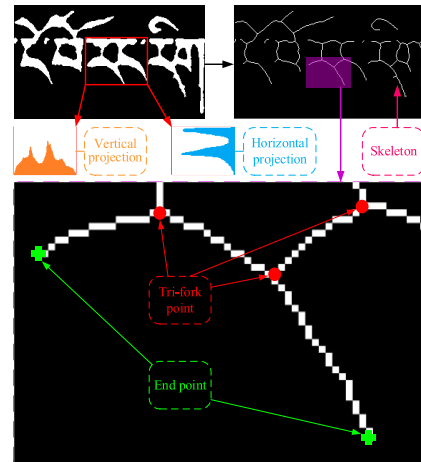


FIGURE 5. Schematic diagram of key feature information.

TABLE 1. Type of stroke crossing and touching above the baseline.

No.	Component	Example
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		

Since the stroke size above the baseline is smaller, there are many types of templates, and the sizes of different template types are not uniform, which makes matching difficult. Therefore, the template matching algorithm based on the error value has been improved to some extent; that is, the stroke size to be matched is normalized to the size of the current template type before the matching calculation, and the stroke size is dynamically adjusted during the matching process. The pixel error value corresponding to the template and the image to be matched is used as the matching criterion.

After detecting the crossing and touching of the strokes above the baseline in the character block by algorithm 1, the number  $N_{touch}$  of crossing and touching in the character block and the position of the corresponding strokes are obtained. According to the obtained crossing and touching information, crossing and touching strokes are segmented.

**Algorithm 1** Algorithm 1: Detection Algorithm for Stroke Crossing and Touching Types Above the Baseline

**Input:** Stroke  $S_{\text{above}}$  above the baseline;  
 Template database  $T_{\text{database}}$ ;

**Output:** Touching type  $S_{\text{touchType}}$

for  $m = 1 \dots 14$

$T_{\text{currentType}} \leftarrow T_{\text{database}}(m)$ ;

$N = \text{count}(T_{\text{currentType}})$ ;

$\text{resize}(S_{\text{above}}) \leftarrow \text{size}(T_{\text{currentType}})$ ;

for  $n = 1 \dots N$

$T_{\text{temp}} \leftarrow T_{\text{currentType}}(n)$ ;

$E_{\text{eachType}}(n) = (S_{\text{above}} - T_{\text{temp}})^2$ ;

endfor

$E_{\text{currentType}}(m) = \min E_{\text{eachType}}$ ;

endfor

$S_{\text{touchType}} \leftarrow \arg \min E_{\text{currentType}}$ .

The space between characters in historical Tibetan documents is smaller, and strokes overlap to varying degrees. Moreover, there are many types and numbers of strokes crossing and touching, which make it very difficult to segment strokes. After observing the crossing and touching of a large number of character strokes in historical Tibetan documents, we found that the crossing and touching between strokes are mainly presented in four types (as shown in Figure 6, the upper part is a schematic diagram, and the lower part is the corresponding example). The black and gray squares in the figure represent different strokes, where subfigure (a) is the intersection of two strokes, subfigures (b) and (c) are the touching between the stroke end and the stroke middle, and subfigure (d) is the touching of the ends of the two strokes.

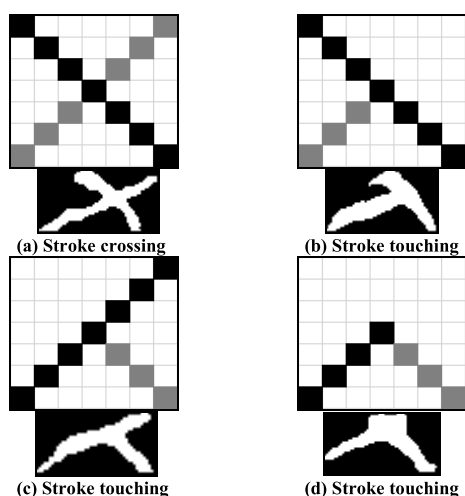


FIGURE 6. Diagrams and examples of stroke crossing and touching types.

Because historical Tibetan documents are handwritten texts, the two strokes of crossing and touching will vary by different tilts. We summarize seven segmentation paths

applicable to most of the strokes of crossing and touching (Figure 7).

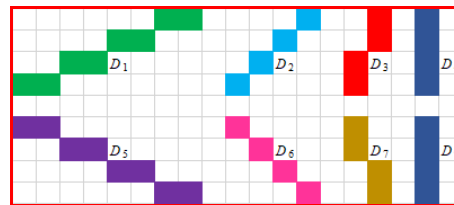


FIGURE 7. Seven kinds of tilts of crossing and touching strokes.

We use the trigonometric formula to calculate the tilt angles corresponding to the seven segmentation paths in Figure 7 as follows:

$$\begin{cases} \theta_{D_1} = \arctan(0.5) \\ \theta_{D_2} = \arctan(1) \\ \theta_{D_3} = \arctan(2) \\ \theta_{D_4} = \pi/2 \\ \theta_{D_5} = \pi - \arctan(0.5) \\ \theta_{D_6} = \pi - \arctan(1) \\ \theta_{D_7} = \pi - \arctan(2) \end{cases} \quad (1)$$

where  $\arctan(\cdot)$  represents the inverse of the tangent function,  $\pi$  represents an angle of  $180^\circ$ , and  $\theta_{D_1} - \theta_{D_7}$  is the angle in the corresponding direction within a period of the tangent function. In crossing and touching stroke segmentation, this angle corresponds to the direction of stroke segmentation.

The formula for calculating the slope is as follows:

$$K = \text{abs} \left( \frac{P_{\text{leftEndY}} - P_{\text{segStartY}}}{P_{\text{leftEndX}} - P_{\text{segStartX}}} \right) \quad (2)$$




**B. STROKE CROSSING AND TOUCHING BELOW THE BASELINE**

In historical Tibetan documents, strokes below the baseline are composed of a prefixed consonant (PFC), a superscript consonant (SPC), a base consonant (BC), a subscript consonant (SBC), a bottom vowel (BV), a suffixed consonant (SFC) and a further suffixed consonant (FSFC) according to the rules of Tibetan construction. Therefore, strokes below the baseline account for most of the total types of character strokes; thus, there are more complex problems of touching stroke below the baseline than above the baseline. The problem of stroke touching below the baseline is mainly distributed in the ‘‘head’’ of character (HC) and below the HC (Figure 8). The baseline is the ‘‘base line’’ in Tibetan writing, and the HC is the stroke head of the character at the baseline position (the stroke inside the rectangular box in Figure 8 (a)). The HC is an important reference stroke for the left and right characters in the base position. Inspired by this, we use the HC as an important basis for whether there is adhesion below the baseline.

We find that the touching position of the strokes below the HC is always below the gap between the two adjacent

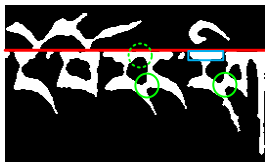
**Algorithm 2** Algorithm 2: Multidirection and Multipath Crossing and Touching Segmentation Algorithm

**Input:** Touching stroke  $S_{touch}$ ;  
 Crossing and touching types  $S_{touchType}$ ;  
 The number  $N_{touch}$  of crossing and touching strokes;  
 The position of crossing and touching strokes;  
 The baseline position of the character block

**Output:** Segmented stroke  $S_{seged}$   
 $S_{skeleton} \leftarrow$  skeleton operation:  $S_{touch}$ ;  
 $N_{trifork} \leftarrow$  count the number of tri-fork points within a certain range of  $S_{skeleton}$ ;  
**for**  $n = 1 \dots N_{touch}$   
   **if**  $N_{trifork} == 0$  **then**  
      $S_{seged} \leftarrow$  segment  $S_{touch}$  (such as composed of “”, “”, “”) in the  $D_4$  direction at different widths;  
   **else**  
      $P_{trifork} \leftarrow$  the tri-fork point with smallest abscissa;  
      $P_{segStart} \leftarrow$  calculate segmentation starting point combining  $P_{trifork}$  and  $S_{touchType}$ ;  
      $P_{leftNum}, P_{rightNum} \leftarrow$  calculate pixels number extending to the left and right from  $P_{trifork}$  in  $S_{skeleton}$ ;  
      $P_{leftEnd}, P_{rightEnd} \leftarrow$  calculate left and right end points in  $S_{skeleton}$ ;  
      $K \leftarrow$  calculate the slope using Formula (2);  
     segmentation direction  $\leftarrow \arg \min_{i=1\dots 7} [\theta_{D_i} - \arctan(K)]$ ;  
     segmentation path  $\leftarrow$  combine  $S_{touchType}, P_{leftNum}, P_{rightNum}$ , and extension threshold  $P_{Thd}$ ;  
      $S_{seged} \leftarrow$  segment the touching strokes from  $P_{segStart}$  with different segmentation directions and segmentation paths.  
   **endif**  
**endfor**



(a) Touching HC (dotted circle)      (b) Touching stroke below the baseline (solid circle)



(c) Touching strokes below the baseline and the HC

**FIGURE 8.** Example of a touching stroke (marked by a circle) below the baseline.

HCS. First, the position of the character HC is used to initially locate the touching range of the stroke. Then, the specific location of the touching point within the touching range is determined. Finally, the touching segmentation path is determined according to the positional relationship between the touching point and the end point of the stroke. Based on this, we propose a touching segmentation algorithm based on the position information of the HC. The algorithm consists of two parts: touching HC detection and segmentation

(Algorithm 3) and stroke touching detection and segmentation below the HC (Algorithm 4).

**Algorithm 3** Algorithm 3: Touching HC Detection and Segmentation Algorithm

**Input:** Character block  $B_{below}$  containing only strokes below the baseline  
**Output:** Character block  $B_{segedHC}$  after touching HC segmentation  
 $B_1 = B_{below}$ ;  
 $S_{width} \leftarrow []$ ;  
 $S_{width} \leftarrow$  calculate widths of the connected component of  $B_1$ ;  
 $B_2 \leftarrow$  shield other connected components if  $S_{width} > 1.5 C_{avgWidth}$ , detect breaking HC in  $B_1$  if Formula (3) is met;  
 $B_3 \leftarrow$  segment HC at a stroke height ( $1.25 S_{avgWidth}$ ) in  $B_2$   
 $B_4 \leftarrow$  detect and connect “point” and “breakpoint” type HC in  $B_3$ ;  
 $B_{segedHC} \leftarrow$  HC of  $B_{below}$  is vertically segmented into  $N_{head}$  equal parts if the number and width of the HC in  $B_4$  meet Formula (4).

The method of detecting the breaking HC is shown in Formula (3), and the method of calculating the actual number of HCs is shown in Formula (4).

$$\begin{cases} S_{cenX} \geq C_{conX} \\ S_{conX} + S_{width} \leq C_{conX} + C_{width} \\ S_{width} \geq 1.5 S_{avgThickness} \\ S_{conY} \geq L_Y \end{cases} \quad (3)$$

$$N_{head} = \begin{cases} 2, & S_{width} \geq 1.2 C_{avgWidth} \text{ and } S_{width} < 1.8 C_{avgWidth} \\ 3, & S_{width} \geq 1.8 C_{avgWidth} \text{ and } S_{width} < 2.4 C_{avgWidth} \\ 4, & S_{width} \geq 2.4 C_{avgWidth} \text{ and } S_{width} < 3 C_{avgWidth} \end{cases} \quad (4)$$

where  $S_{cenX}$  is the abscissa of the centroid of stroke,  $C_{conX}$  is the abscissa of the connected component of character,  $S_{width}$  is the width of the stroke,  $S_{avgThickness}$  is the average thickness of the stroke,  $C_{width}$  is the width of character,  $S_{conY}$  is the ordinate of the connected component of the stroke, and  $L_Y$  is the ordinate of the baseline.

The formula for deleting useless tri-point points and end points is shown in Formula (5), and the stop condition of segmentation is shown in Formula (6).

$$\begin{cases} P_{triforkY} \leq L_Y + 1.25 S_{avgThickness} \\ P_{endY} \leq L_Y + 1.25 S_{avgThickness} \end{cases} \quad (5)$$

$$\begin{cases} C_{segNum} > C_{num} \\ C_{segMaxWidth} < C_{maxWidth} \end{cases} \quad (6)$$

**V. CHARACTER STROKE ATTRIBUTION**

Stroke attribution is the last step in the process of character segmentation, in which the breaking strokes and the segmented strokes are placed according to the position of the original strokes. There are breaking strokes or separating strokes before character segmentation (Figure 9(a)).

#### Algorithm 4 Algorithm 4: Touching Strokes Detection and Segmentation Below the HC Algorithm

**Input:** Character block  $B_{\text{segedHC}}$  after segmentation of HC  
**Output:** Character block  $B_{\text{segedbHC}}$  after touching segmentation below the HC

The number of character stroke  $C_{\text{num}} = 0$ ;  
Maximum width of characters  $C_{\text{maxWidth}} = 0$ ;  
Tri-fork points  $P_{\text{trifork}} \leftarrow []$ ;  
End points  $P_{\text{end}} \leftarrow []$ ;  
Manhattan distance  $D_{\text{Man}} \leftarrow []$ ;  
 $C_{\text{num}}, C_{\text{maxWidth}} \leftarrow$  conduct connected component analysis on  $B_{\text{segedHC}}$ ;  
 $P_{\text{trifork}}, P_{\text{end}} \leftarrow$  calculate tri-fork points and end points after skeletonization on  $B_{\text{segedHC}}$ ;  
 $P_{\text{trifork}}, P_{\text{end}} \leftarrow$  delete useless  $P_{\text{trifork}}$  and  $P_{\text{end}}$  according to Formula (5);  
 $P_{\text{trifork}} \leftarrow$  delete  $P_{\text{trifork}}$  outside the gap between two adjacent HCs;  
**if**  $P_{\text{trifork}} \neq \text{null}$  **then**  
     $(P_{\text{trifork}X}, P_{\text{triforkmax}Y}) \leftarrow$  select tri-fork point of maximum ordinate from  $P_{\text{trifork}}$ ;  
     $D_{\text{Man}} \leftarrow$  calculate Manhattan distance between  $P_{\text{trifork}X}$  and all  $P_{\text{end}X}$ ;  
     $(P_{\text{end}X}, P_{\text{end}Y}) \leftarrow \arg \min D_{\text{Man}}$ ;  
**if**  $\text{abs}(P_{\text{trifork}X}, P_{\text{end}X}) < 2S_{\text{avgThickness}}$  **then**  
    compare the size relationship between  $(P_{\text{trifork}X}, P_{\text{triforkmax}Y})$  and  $(P_{\text{end}X}, P_{\text{end}Y})$ ;  
    **if**  $P_{\text{end}X} < P_{\text{trifork}X}$  and  $P_{\text{end}Y} < P_{\text{triforkmax}Y}$  **then**  
         $B_{\text{segedbHC}} \leftarrow$  the left of tri-fork point  $(P_{\text{trifork}X}, P_{\text{triforkmax}Y})$  is segmented in the  $D_6$  direction, and the right is segmented horizontally, if Formula (6) is met, stop segmenting;  
    **elseif**  $P_{\text{end}X} > P_{\text{trifork}X}$  and  $P_{\text{end}Y} > P_{\text{trifork}Y}$  **then**  
         $(P_{\text{median}X}, P_{\text{median}Y}) \leftarrow$  calculate the median of tri-fork points if there are other tri-fork points near  $(P_{\text{trifork}X}, P_{\text{triforkmax}Y})$ ;  
         $B_{\text{segedbHC}} \leftarrow$  the above of median point  $(P_{\text{median}X}, P_{\text{median}Y})$  is segmented in the  $D_4$  direction, and the below is segmented in the  $D_6$  direction, if Formula (6) is met, stop segmenting;  
    **else**  
         $B_{\text{segedbHC}} \leftarrow$  segment touching strokes at the minimum position of the vertical projection pixel of the gap between the two adjacent HCs in the  $D_4$  direction, if Formula (6) is met, stop segmenting;  
    **endif**  
**endif**  
**endif**

In Figure 9(a), “new” strokes are broken from a complete stroke marked with an elliptical box, and a complete stroke marked with a rectangular box is separated from other strokes of the character. After character segmentation, breaking strokes are added to characters, which are marked with a dotted elliptical box (Figure 9(b)).

The above character strokes are classified as breaking (or separating) strokes above and below the baseline. The strokes above the baseline are mainly vowels with fewer types of strokes, and each vowel has specific geometric features. The types of vowels above the baseline can be determined by detecting geometric features of the strokes, and the strokes above the baseline are used as the location basis for

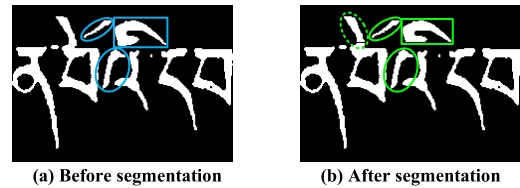


FIGURE 9. Examples of breaking (or separating) strokes before and after character segmentation.

TABLE 2. Vowel stroke types and geometric features above the baseline.

No.	Stroke type	Basic geometric features of a stroke
1		$Y_{\text{right}} > Y_{\text{cen}} > Y_{\text{left}}$
2		composed of two No. 1 strokes, with the same features as No. 1
3		$Y_{\text{left}} > Y_{\text{cen}}$ , $Y_{\text{right}} > Y_{\text{cen}}$ and $Y_{\text{right}} > Y_{\text{left}}$
4		regarded as No. 3, with the same features as No. 3
5		$Y_{\text{left}} > Y_{\text{cen}}$ , $Y_{\text{right}} > Y_{\text{cen}}$ and $Y_{\text{right}} < Y_{\text{left}}$
6		$Y_{\text{cen}} > Y_{\text{left}}$ and $Y_{\text{cen}} > Y_{\text{right}}$
7		composed of two No. 6 strokes, with the same features as No. 6
8		regarded as No. 6, with the same features as No. 6
9		left: $Y_{\text{right}} > Y_{\text{cen}} > Y_{\text{left}}$ ; right: $Y_{\text{right}} < Y_{\text{cen}} < Y_{\text{left}}$

attribution. However, there are many types of strokes below the baseline, and the degree of breaking varies greatly between strokes, so the stroke type cannot be determined by the geometric features of strokes.

#### A. STROKE TYPE ABOVE THE BASELINE

Most of the strokes above the baseline are vowels, whose geometric features are obviously different, so the vowel type can be determined by detecting their geometric features. Let  $Y_{\text{left}}$  and  $Y_{\text{right}}$  be the corresponding maximum ordinate of foreground pixels in the first and last columns on the left and right sides of the stroke connected component, respectively, and let  $Y_{\text{cen}}$  be the ordinate of the centroid of the stroke connected component. There are a large number of Tibetan transliterations of Sanskrit texts in historical Tibetan documents, and the type geometric features of strokes above the baseline will increase accordingly. Therefore, we summarized the stroke types above the baseline that might affect the effect of character segmentation (Table 2) and counted the type and number of strokes.

#### B. STROKE BREAKING TYPE BELOW THE BASELINE

The breaking of a stroke below the baseline seriously affects the effect of character segmentation. Breaking usually occurs at positions with thinner longitudinal strokes, basically presenting four breaking stroke types (Figure 10). As shown in Figure 10, A and B are the centroids of the rectangular box outside the connected component of different breaking strokes.



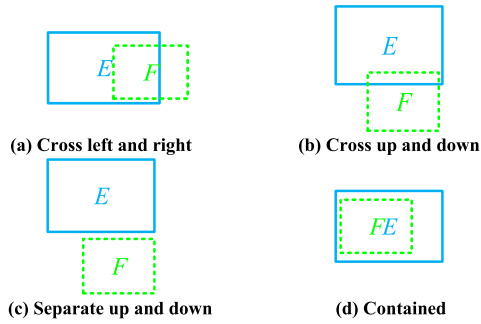


FIGURE 10. Breaking types of strokes below the baseline.

Thus, we conclude the method of determining breaking strokes below the baseline. The abscissa of the centroids, the ordinate of the upper boundary and the area of the connected component of strokes should meet the following conditions at the same time:

$$\begin{cases} abs(E_{cenX} - F_{cenX}) < X_{cenThd} \\ abs(E_{conY} - F_{conY}) > Y_{conThd} \\ E_{area} > P_{areaThd} \\ F_{area} > P_{areaThd} \end{cases} \quad (7)$$

where  $E_{cenX}$ ,  $F_{cenX}$  and  $X_{cenThd}$  are the abscissa of the centroids of the connected components  $A$ ,  $B$  and the horizontal distance threshold of the centroid of connected component, respectively;  $E_{conY}$ ,  $F_{conY}$  and  $Y_{conThd}$  are the ordinate of the connected component and the distance threshold of the connected component, respectively;  $E_{area}$ ,  $F_{area}$  represent the area of the connected component; and  $P_{areaThd}$  is the area threshold of the syllable points. The area threshold of syllable points is set to avoid the influence of syllable points during the determination of breaking strokes.

C. THE ATTRIBUTION DISTANCE OF BREAKING STROKES ABOVE AND BELOW THE BASELINE

By calculating the distance between the breaking stroke and other strokes of the original character, the character belonging to the stroke is realized. There are many types of breaking strokes above and below the baseline, and the characters in historical Tibetan documents are handwritten, with great differences in character styles. For this challenging problem, we summarize three attribution distances between the strokes.

1) CENTROID-BASED ATTRIBUTION DISTANCE (Cen-AD)

Both strokes above and below the baseline use the centroid of the connected component

In historical Tibetan documents, the offset of the left and right positions of the strokes above and below the baseline of some characters is small and can be attributed to the horizontal distance between the centroid of each stroke connected component. According to the horizontal distance  $D_{cenX}$  between the centroid of the strokes (Figure 11(b)), the strokes above and below the baseline are attributed to a complete character.

2) CONNECTED COMPONENT-BASED ATTRIBUTION DISTANCE (Con-AD)

Both strokes above and below the baseline use the abscissa of the connected component external rectangle box

Historical Tibetan documents are always written from left to right along the baseline. Most of the strokes below the baseline tend to tilt slightly to the bottom right, which makes it difficult to attribute strokes. Through the analysis of the characteristics of a large number of historical Tibetan documents, it was found that although most of the strokes have different degrees of right slant, the strokes near the BC hardly slant. Therefore, the horizontal distance  $D_{conX}$  (Figure 11(c)) of the abscissa of the connected component external rectangle box of each stroke is used to attribute each stroke of a character.

3) COMBINED CONNECTED AND CENTROID-BASED ATTRIBUTION DISTANCE (ConCen-AD)

Only the strokes below the baseline use the abscissa of the connected component external rectangle box, and the centroid of the connected component is used for the remainder of the strokes.

In historical Tibetan documents, most of the breaking strokes are located below the baseline, and the types and degrees of the breaking strokes are diverse, which seriously affects the stroke attribution effect. The strokes below the baseline are attributed to the horizontal distance of the abscissa of the connected component external rectangle box, and the horizontal distance of the centroid of the connected component is used for attribution of the remaining strokes (Figure 11(d)).

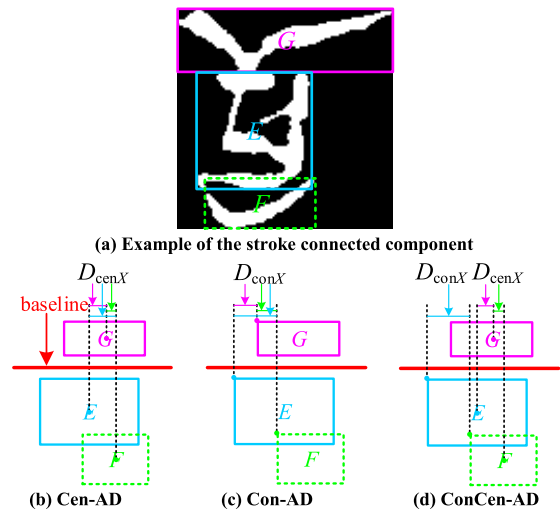


FIGURE 11. Schematic diagram of three kinds of attribution distances.



Combining the above three kinds of stroke attribution distances, we propose a breaking stroke attribution process (Figure 12).

The three attribution distances used in Figure 12 form three stroke attribution methods: *Cen-AM*, *Con-AM*, and *ConCen-AM*.






FIGURE 12. Character stroke attribution process.

The writing styles of the characters in historical Tibetan documents are quite different. If the strokes are severely breaking or strokes with the same features appear repeatedly, then the attribution of strokes will be difficult. For this reason, the following processing needs to be performed for the complicated strokes above the baseline when the strokes are attributed.

(1) The No. 6 “” type breaks to form the No. 9 “” stroke type. The left stroke after breaking has the same geometric features as the No. 1 stroke type, and the right stroke after breaking is often above the baseline of the adjacent character on its right. In this case, the horizontal coordinates of the centroid of the left and right strokes are used to calculate the horizontal coordinates of the overall centroid of the stroke, which are defined as:

$$S_{cenX} = (S_{leftCenX} + S_{rightCenX})/2 \quad (8)$$

where  $S_{leftCenX}$  is the horizontal abscissa of the centroid of the left stroke and  $S_{rightCenX}$  is the horizontal abscissa of the centroid of the right stroke.

(2) When strokes No. 1 and No. 9 appear in the same character block in the left and right adjacent positions and appear alternately in the form of stroke types of “”, “”, and “”, the horizontal distance between the centroid of the stroke above the baseline and stroke types should

be combined, and the strokes of each character should be attributed based on the strokes above the baseline.

To facilitate the statistics of the character segmentation effect, after the character stroke attribution in the character block is completed, the red-green-blue (RGB) color value is used to color each character in the block to represent the result of character segmentation.

## VI. CHARACTER SEGMENTATION EVALUATION METHOD

Historical Tibetan documents belong to the category of handwritten texts. There are differences in writing styles between characters, and some characters have serious stroke deformation. To objectively evaluate the advantages and disadvantages of this method, the results of character segmentation are evaluated by manual statistics.

Manual statistics were carried out using the recall rate ( $R_{seg}$ ), precision rate ( $P_{seg}$ ) and harmonic mean ( $F1_{seg}$ ), which are defined as:

$$R_{seg} = \frac{NCSC}{TNC} \times 100\% \quad (9)$$

$$P_{seg} = \frac{NCSC}{TNCS} \times 100\% \quad (10)$$

$$F1_{seg} = \frac{2 \times R_{seg} \times P_{seg}}{R_{seg} + P_{seg}} \times 100\% \quad (11)$$

where  $NCSC$  is the number of correctly segmented characters,  $TNC$  is the total number of characters, and  $TNCS$  is the total number of segmented characters.

## VII. EXPERIMENTAL RESULTS AND ANALYSIS

We randomly select 212 images from historical Tibetan documents as the basic data of this paper. Each document contains left and right marked text, box lines, and text lines. This paper only studies character segmentation, so we directly use our research results in document binarization and text line segmentation, namely, segmented text lines for relevant research and experimental analysis.

### A. CHARACTER BLOCK DATABASE

This paper uses two methods to shorten text lines, namely, the text shortening methods *Pro-SM* and *Syl-SM*. The part of the text line marked by the dotted box in Figure 3 obtains the character block by the above two text shortening methods, as shown in Figure 13(a) and (b). We calculated the number of different character blocks obtained by the two shortening methods (Figure 3) and formed the corresponding character block library. In terms of the number of character blocks, the shortening method *Pro-SM* segments out more character blocks, and the average width of the character blocks is smaller.

### B. TOUCHING STROKE DETECTION AND SEGMENTATION PROCESS

In view of the complexity of character stroke crossing and breaking in historical Tibetan documents, crossing is

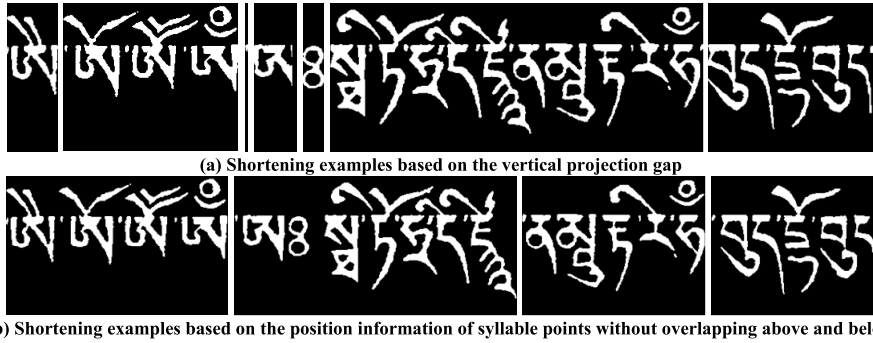


FIGURE 13. Examples of character blocks after different shortening methods.

TABLE 3. Statistics on the number of character blocks formed by two text line shortening methods.

Text line shortening method	Number of documents	Number of text lines	Total number of character blocks	Containing only Syllable points or punctuation marks ( “   ”, “   ”, “   ” )	Containing ordinary characters
<i>Pro-SM</i>	212	1696	109603	22418	87185
<i>Syl-SM</i>	212	1696	47632	-	-

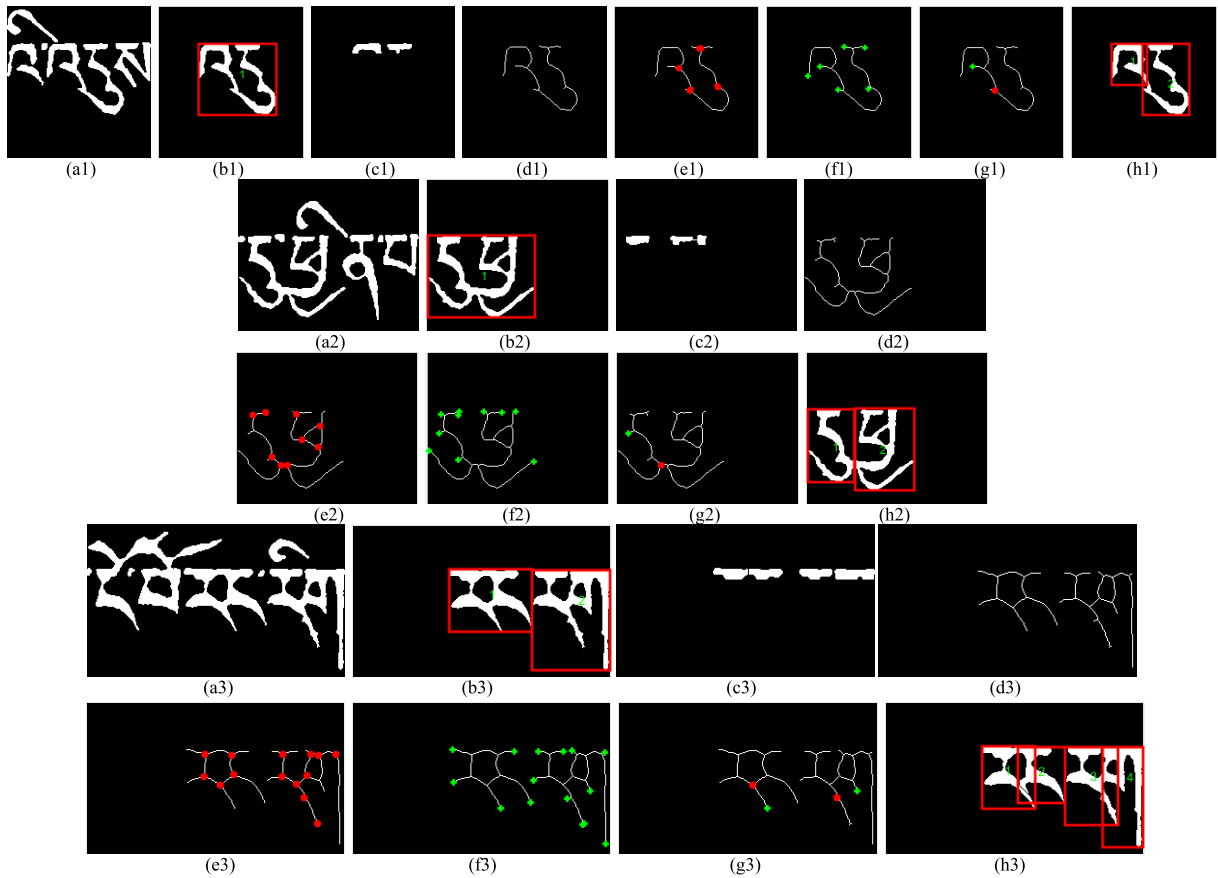


FIGURE 14. Example of touching stroke detection and segmentation below the HC.

a special case of touching. The crossing and touching of the strokes above and below the baseline are dealt with separately.

1) STROKE ABOVE THE BASELINE

Detecting the touching of strokes above the baseline. If there is touching stroke, then the touching type is obtained.

TABLE 4. Character segmentation results obtained by systematic sampling statistics.

Text line shortening method	Stroke attribution method	NCB	NCSCB	NISCB	TNC	NSC	NCSC	$R_{seg}$	$P_{seg}$	$FI_{seg}$
Pro-SM	Cen-AM	10961	9916	1045	22503	23454	20609	0.9158	0.8787	0.8969
	Con-AM	10961	9971	990	22503	23018	20653	0.9178	0.8973	0.9074
	ConCen-AM	10961	9868	1093	22503	23406	20511	0.9115	0.8763	0.8936
Syl-SM	Cen-AM	4764	4012	752	22726	23389	21390	<b>0.9412</b>	<b>0.9145</b>	<b>0.9277</b>
	Con-AM	4764	3931	833	22726	23433	21016	0.9248	0.8969	0.9106
	ConCen-AM	4764	3909	855	22726	23597	21030	0.9254	0.8912	0.9080

TABLE 5. Character segmentation results obtained by random sampling statistics.

Text line shortening method	Stroke attribution method	NCB	NCSCB	NISCB	TNC	NSC	NCSC	$R_{seg}$	$P_{seg}$	$FI_{seg}$
Pro-SM	Cen-AM	10961	9963	998	22683	23473	20967	0.9243	0.8932	0.9085
	Con-AM	10961	10056	905	22882	23234	21266	0.9294	0.9153	0.9223
	ConCen-AM	10961	9927	1034	22573	23445	20719	0.9179	0.8837	0.9005
Syl-SM	Cen-AM	4764	3994	770	22369	23083	21049	<b>0.9410</b>	<b>0.9119</b>	<b>0.9262</b>
	Con-AM	4764	3937	827	22786	23325	21092	0.9257	0.9043	0.9149
	ConCen-AM	4764	3922	842	23058	23665	21518	0.9332	0.9093	0.9211

Otherwise, there was no touching stroke above the baseline. Algorithm 3 is used to segment the complex touching strokes above the baseline.

2) STROKE BELOW THE BASELINE

There are many types of touching strokes below the baseline, and the difference is large. Algorithm 4 and algorithm 5 are used to detect and segment the HC and below the HC, respectively (Figure 14).

C. CHARACTER SEGMENTATION RESULTS

Stroke attribution is the last step of character segmentation, which is to form a complete character image according to the corresponding position of the original complete stroke. When attributing breaking strokes, the condition of broken strokes above and below the baseline should also be considered at the same time. If there are strokes above the baseline, then the attribution is based on the strokes above the baseline. According to the morphological features of characters in historical Tibetan documents, we adopt three attribution methods: Cen-AM, Con-AM and ConCen-AM.

After the establishment of the above character block database, the detection and segmentation of crossing and breaking, and stroke attribution, the amount of character image data obtained reaches more than 230,000. To count the result of character segmentation more efficiently and accurately, 10% of the total result of character segmentation is sampled. Since there are differences in writing styles between pages of historical Tibetan documents to varying degrees or even between different lines on the same page, systematic sampling is adopted to obtain statistics on the character segmentation results to ensure uniform sampling coverage (Table 4). At the same time, to reflect the

reliability of sampling more objectively, random sampling is also adopted to conduct statistics on the segmentation results (Table 5). As shown in Table 4 and Table 5, NCB is the number of character blocks, NCSCB is the number of correctly segmented character blocks, NISCB is the number of incorrectly segmented character blocks, TNC is the total number of characters, NSC is the number of segmented characters, NCSC is the number of correctly segmented characters,  $R_{seg}$  is the character segmentation recall rate,  $P_{seg}$  is the character segmentation precision rate, and  $FI$  is the harmonic mean of  $R_{seg}$  and  $P_{seg}$  ( $R_{seg}$  and  $P_{seg}$  are equally important).

After analyzing Table 4 and Table 5 obtained by different sampling methods, the following conclusions can be drawn:

- 1) There is little difference between the statistical results of systematic sampling and random sampling, and 10% of the total number of samples can be feasibly used for statistics;
- 2) The character segmentation results obtained by text shortening based on syllable point information are better than those of the projection-based shortening method on the whole;
- 3) The performance results of the three stroke attribution methods on the character block obtained by the shortening method based on projection are similar, while the performance result of the attribution method Cen-AM on the character block obtained by the shortening method based on syllable points is slightly better.

Character segmentation speed is also an important indicator to evaluate the effect of character segmentation.

The character segmentation time was tested on a personal desktop computer (CPU: Core I7-9700 3.00 GHz) (Table 6). We counted the time consumed by the character block character databases established by two different shortening





FIGURE 15. Example of the character segmentation result.

TABLE 6. Statistics of character segmentation time.

Text line shortening method	Stroke attribution method	Total number of character blocks (blocks)	Total segmentation time (seconds)	Average time for each block segmentation (seconds)	Average time of the three stroke attribution methods (seconds)
Pro-SM	Cen-AM	109603	10331	0.0943	0.1087
	Con-AM	109603	12187	0.1112	
	ConCen-AM	109603	13218	0.1206	
Syl-SM	Cen-AM	47632	13558	0.2846	0.2922
	Con-AM	47632	13916	0.2922	
	ConCen-AM	47632	14273	0.2997	

methods after segmentation. different attribution methods were used to attribute the breaking strokes, and finally independent characters were formed. Table 6 shows that the average time of character block segmentation based on projection shortening and without overlapping syllable point shortening is 0.1087 seconds and 0.2922 seconds, respectively, which is faster and can meet the character segmentation task of large-scale data.

Figure 15 shows character blocks before and after character segmentation, and character segmentation problems are marked in the character blocks. Different colors in the character blocks after segmentation represent different characters.

D. LIMITATION ANALYSIS

Although character detection and segmentation have achieved good results, the proposed method has certain limitations in the face of extremely complex touching characters. Figure 16 shows examples of detection and segmentation failure. The upper, middle, and lower subpictures in each group of pictures are the character block, each touching (or breaking) character or touching syllable point mark in the character block, and the character block after character segmentation, respectively. The touching (or breaking) characters are marked with a dotted curve circle, and the touching syllable points are marked with a solid circle. The dotted line



**FIGURE 16.** Examples of the limitations of character detection and segmentation algorithm. (a) The touching point is located in the HC. (b) and (c) The touching point is located in the concave stroke. (d) The handwriting is at will or the original image is of low quality and the character stroke features are not obvious. (e) Complex touching and breaking among strokes of multiple characters.

in the box in the following lower subpicture is the correct segmentation path. The details are described as follows:

1) The width of some characters is small, and the touching position is special. The touching point is located in the HC or in the concave stroke, so the touching point cannot be accurately detected or accurately segmented. Figure 16(a) shows that the touching point is located in the HC, and the touching cannot be detected, resulting in the failure of character segmentation. Figure 16(b) and (c) show that the touching point is located in the concave stroke, which cannot be accurately segmented.

2) Characters are written at will or the original image quality is not high, features of the character stroke are not obvious, and the character segmentation path cannot be determined. Figure 16(d) shows that when the handwriting is at will or

the original image is of low quality and the character stroke features are not obvious, incorrect segmentation follows.

3) The touching between strokes of multiple characters is complex and simultaneously breaks, so the position of the touching point cannot be accurately determined. Figure 16(e) shows complex touching and breaking among strokes of multiple characters, and the touching position detection fails.

### VIII. CONCLUSION

We proposed a character detection and segmentation method based on key feature information for historical Uchen Tibetan documents, which consists of four parts: 1) shortening the text line based on vertical projection gap and position information of syllable points without overlapping above and below; 2) baseline detection based on the location information of syl-

lable points or combined with horizontal projection and linear detection; 3) a multidirection and multipath crossing and touching stroke segmentation based on key feature information; and 4) attribution of breaking strokes based on Tibetan structure, stroke form and stroke position. Experiments on historical Uchen Tibetan documents show that our method can solve the three challenges in the process of character segmentation and achieve a higher accuracy rate in character segmentation, providing data support for subsequent research on document analysis and recognition of historical Uchen Tibetan documents. In future work, we will consider the detection and segmentation of the touching of syllable points and characters.

Although our method is only suitable for character detection and segmentation of historical Tibetan documents, it can provide a reference for similar character detection and segmentation challenges as follows: 1) small-scale datasets; 2) characters composed of fixed structural units; and 3) characters that need to be segmented more accurately than those segmented more roughly. The proposed method can be applied to the following scenarios: a) digitization of historical Tibetan documents and cultural relics; b) character detection and segmentation of other historical Tibetan documents; and c) Tibetan handwritten signature character detection and segmentation.

## REFERENCES

- [1] M. Kojima, Y. Kawazoe, and M. Kimura, "Automatic character recognition for Tibetan script," *J. Indian Buddhist Stud.*, vol. 39, no. 2, pp. 844–848, 1991, doi: [10.4259/ibk.39.848](https://doi.org/10.4259/ibk.39.848).
- [2] M. Kojima, Y. Akiyama, Y. Kawazoe, and M. Kimura, "Extraction of characteristic features in Tibetan wood-block editions," *J. Indian Buddhist Stud.*, vol. 42, no. 2, pp. 866–869, 1994, doi: [10.4259/ibk.42.869](https://doi.org/10.4259/ibk.42.869).
- [3] M. Kojima, Y. Kawazoe, and M. Kimura, "Character recognition of wooden blocked Tibetan similar manuscripts by using Euclidean distance with deferential weight," *IPSJ SIGNotes Comput. Humanities*, vol. 42, pp. 13–18, May 1996.
- [4] Y. Han, W. Wang, H. Liu, and Y. Wang, "A combined approach for the binarization of historical Tibetan document images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 33, no. 14, May 2019, Art. no. 1954038, doi: [10.1142/S0218001419540387](https://doi.org/10.1142/S0218001419540387).
- [5] Z. J. Li, W. L. Wang, and Z. Q. Cai, "Historical document image binarization based on edge contrast information," in *Proc. Adv. Comput. Vis. (CVC)*, 2019, pp. 614–628, doi: [10.1007/978-3-030-17795-9\\_44](https://doi.org/10.1007/978-3-030-17795-9_44).
- [6] X. Q. Zhang, L. L. Ma, L. J. Duan, Z. Y. Liu, and J. Wu, "Layout analysis for historical Tibetan documents based on convolutional denoising autoencoder," *J. Chin. Inf. Process.*, vol. 32, no. 7, pp. 67–73, Jul. 2018. [Online]. Available: <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=MESS201807009&DbName=CJFQ2018>
- [7] H. Liu, X. Bi, and W. Wang, "Layout analysis of historical Tibetan documents," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 74–78, doi: [10.1109/ICAIBD.2019.8837040](https://doi.org/10.1109/ICAIBD.2019.8837040).
- [8] P. Zhao, W. Wang, Z. Cai, G. Zhang, and Y. Lu, "Accurate fine-grained layout analysis for the historical Tibetan document based on the instance segmentation," *IEEE Access*, vol. 9, pp. 154435–154447, 2021, doi: [10.1109/ACCESS.2021.3128536](https://doi.org/10.1109/ACCESS.2021.3128536).
- [9] L.-J. Duan, X.-Q. Zhang, L.-L. Ma, and J. Wu, "Text extraction method for historical Tibetan document images based on block projections," *Optoelectron. Lett.*, vol. 13, no. 6, pp. 457–461, Nov. 2017, doi: [10.1007/s11801-017-7197-0](https://doi.org/10.1007/s11801-017-7197-0).
- [10] Y. X. Li, L. L. Ma, L. J. Duan, and J. Wu, "A text-line segmentation method for historical Tibetan documents based on baseline detection," in *Proc. CCCF Chin. Conf. Comput. Vis. (CCCV)*, Tianjin, China, 2017, pp. 356–367, doi: [10.1007/978-981-10-7299-4\\_29](https://doi.org/10.1007/978-981-10-7299-4_29).
- [11] F. Zhou, W. Wang, and Q. Lin, "A novel text line segmentation method based on contour curve tracking for Tibetan historical documents," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 10, Oct. 2018, Art. no. 1854025, doi: [10.1142/S0218001418540253](https://doi.org/10.1142/S0218001418540253).
- [12] Z. Li, W. Wang, Y. Chen, and Y. Hao, "A novel method of text line segmentation for historical document image of the uchen Tibetan," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 23–32, May 2019, doi: [10.1016/j.jvcir.2019.01.021](https://doi.org/10.1016/j.jvcir.2019.01.021).
- [13] D. C. Zhao, "Research on wooden blocked Tibetan character segmentation based on drop penetration algorithm," in *Proc. Chin. Conf. Pattern Recognit. (CCPR)*, Oct. 2010, pp. 1–5, doi: [10.1109/CCPR.2010.5659181](https://doi.org/10.1109/CCPR.2010.5659181).
- [14] Q. C. Zhao, L. L. Ma, and L. J. Duan, "A touching character database from Tibetan historical documents to evaluate the segmentation algorithm," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Guangzhou, China, 2018, pp. 309–321, doi: [10.1007/978-3-030-03341-5\\_26](https://doi.org/10.1007/978-3-030-03341-5_26).
- [15] W. L. Wang, X. B. Lu, Z. Q. Cai, W. T. Shen, J. Fu, and Z. X. Caike, "Online handwritten sample generated based on component combination for Tibetan-Sanskrit," *J. Chin. Inf. Process.*, vol. 31, no. 5, pp. 64–73, Sep. 2017. [Online]. Available: <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=MESS201705010&DbName=CJFQ2017>
- [16] Z. J. Li and W. L. Wang, "Tibetan historical document recognition of uchen script using baseline information," in *Proc. 10th Int. Conf. Graph. Image Process. (ICGIP)*, May 2019, pp. 11069–110693H, doi: [10.1117/12.2524234](https://doi.org/10.1117/12.2524234).
- [17] F. Hedayati, J. Chong, and K. Keutzer, "Recognition of Tibetan wood block prints with generalized hidden Markov and kernelized modified quadratic distance function," in *Proc. Joint Workshop Multilingual OCR Anal. Noisy Unstructured Text Data*, Sep. 2011, pp. 1–14, doi: [10.1145/2034617.2034631](https://doi.org/10.1145/2034617.2034631).
- [18] L. Ma, C. Long, L. Duan, X. Zhang, Y. Li, and Q. Zhao, "Segmentation and recognition for historical Tibetan document images," *IEEE Access*, vol. 8, pp. 52641–52651, 2020, doi: [10.1109/ACCESS.2020.2975023](https://doi.org/10.1109/ACCESS.2020.2975023).
- [19] S. F. Zhou, C. P. Liu, G. Liu, and S. R. Gong, "Multi-step segmentation method based on minimum weight segmentation path for ancient handwritten Chinese character," *J. Chin. Comput. Syst.*, vol. 33, no. 3, pp. 614–620, Mar. 2012, doi: [10.3969/j.issn.1000-1220.2012.03.032](https://doi.org/10.3969/j.issn.1000-1220.2012.03.032).
- [20] Y. M. Qi, X. D. Tian, and L. N. Zuo, "Segmentation method of ancient Chinese character images based on hesitant fuzzy sets," *Sci. Technol. Eng.*, vol. 19, no. 30, pp. 232–240, Oct. 2019. [Online]. Available: <https://kns.cnki.net/kcms/detail/detail.aspx?FileName=KXJS201930037&DbName=CJFQ2019>
- [21] X. D. Tian, T. Y. Sun, and Y. M. Qi, "Ancient Chinese character image segmentation based on interval-valued hesitant fuzzy set," *IEEE Access*, vol. 8, pp. 146577–146587, 2020, doi: [10.1109/ACCESS.2020.3014219](https://doi.org/10.1109/ACCESS.2020.3014219).
- [22] K. P. Zaw and N. War, "Y-position based Myanmar touching character segmentation and sub-components based character classification," in *Proc. IEEE 17th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, Honolulu, HI, USA, May 2019, pp. 76–83, doi: [10.1109/SERA.2019.8886810](https://doi.org/10.1109/SERA.2019.8886810).
- [23] K. Thongkanhorn, S. Kanchanapreechakorn, P. Borwaringinn, and W. Kusakunniran, "Thai character segmentation in handwriting images using four directional depth first search," in *Proc. 11th Int. Conf. Technol. Electr. Eng. (ICITEE)*, Pattaya, Thailand, Oct. 2019, pp. 1–5, doi: [10.1109/ICITEE.2019.8929972](https://doi.org/10.1109/ICITEE.2019.8929972).
- [24] P. A. Tamhankar, K. D. Masalkar, and S. R. Kolhe, "A novel approach for character segmentation of offline handwritten Marathi documents written in MODI script," *Proc. Comput. Sci.*, vol. 171, pp. 179–187, Jan. 2020, doi: [10.1016/j.procs.2020.04.019](https://doi.org/10.1016/j.procs.2020.04.019).
- [25] A. A. Ali and M. Suresha, "An efficient character segmentation algorithm for recognition of Arabic handwritten script," in *Proc. Int. Conf. Data Sci. Commun. (IconDSC)*, Bengaluru, India, Mar. 2019, pp. 1–6, doi: [10.1109/IconDSC.2019.8817037](https://doi.org/10.1109/IconDSC.2019.8817037).
- [26] I. Ullah, M. Sanusi, M. Ishak, and Y. M. Alomari, "Segmentation of touching Arabic characters in handwritten documents by overlapping set theory and contour tracing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 5, pp. 155–160, 2019, doi: [10.14569/IJACSA.2019.0100519](https://doi.org/10.14569/IJACSA.2019.0100519).
- [27] K. C. Nguyen and N. Masaki, "Enhanced character segmentation for format-free Japanese text recognition," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Shenzhen, China, Oct. 2016, pp. 138–143, doi: [10.1109/ICFHR.2016.0037](https://doi.org/10.1109/ICFHR.2016.0037).
- [28] G. Congedo, G. Dimauro, S. Impedovo, and G. Pirlo, "Segmentation of numeric strings," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, Aug. 1995, pp. 1038–1041, doi: [10.1109/ICDAR.1995.602080](https://doi.org/10.1109/ICDAR.1995.602080).
- [29] M. Rui, D. Jie, G. Yunhua, and Y. Yunyang, "An improved drop-fall algorithm based on background analysis for handwritten digits segmentation," in *Proc. WRI Global Congr. Intell. Syst.*, Xiamen, China, 2009, pp. 374–378, doi: [10.1109/GCIS.2009.60](https://doi.org/10.1109/GCIS.2009.60).
- [30] J. Y. Yang, J. Guo, and W. W. Jiang, "A novel drop-fall algorithm based on digital features for touching digit segmentation," in *Proc. IEEE 7th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Vancouver, BC, Canada, Oct. 2016, pp. 1–4, doi: [10.1109/IEMCON.2016.7746350](https://doi.org/10.1109/IEMCON.2016.7746350).

[31] X. G. Li and W. Gao, "Segmentation method for merged characters in CAPTCHA based on drop fall algorithm," *Comput. Eng. Appl.*, vol. 50, no. 1, pp. 163–166, Jan. 2014, doi: [10.3778/j.issn.1002-8331.1208-0310](https://doi.org/10.3778/j.issn.1002-8331.1208-0310).

[32] X. C. Liu and X. F. Jin, "Characters segmentation method of historical documents mixed in Korean and Chinese," *Comput. Eng. Appl.*, vol. 56, no. 11, pp. 135–141, Jun. 2020, doi: [10.3778/j.issn.1002-8331.1902-0119](https://doi.org/10.3778/j.issn.1002-8331.1902-0119).

[33] J. Ji, L. Peng, and B. Li, "Graph model optimization based historical Chinese character segmentation method," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst.*, Tours, France, Apr. 2014, pp. 282–286, doi: [10.1109/DAS.2014.57](https://doi.org/10.1109/DAS.2014.57).

[34] L. Xu, F. Yin, Q.-F. Wang, and C.-L. Liu, "An over-segmentation method for single-touching Chinese handwriting with learning-based filtering," *Int. J. Document Anal. Recognit. (IJ DAR)*, vol. 17, no. 1, pp. 91–104, Mar. 2014, doi: [10.1007/s10032-013-0208-1](https://doi.org/10.1007/s10032-013-0208-1).

[35] P. Sahare and S. B. Dhok, "Multilingual character segmentation and recognition schemes for Indian document images," *IEEE Access*, vol. 6, pp. 10603–10617, 2018, doi: [10.1109/ACCESS.2018.2795104](https://doi.org/10.1109/ACCESS.2018.2795104).

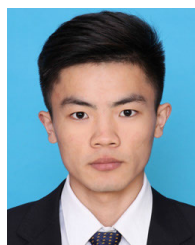
[36] Z. Gao, J. Liu, Y. Li, Y. Yang, and H. He, "A novel semantic segmentation model for Chinese characters," *IEEE Access*, vol. 8, pp. 179083–179093, 2020, doi: [10.1109/ACCESS.2020.3027019](https://doi.org/10.1109/ACCESS.2020.3027019).

[37] Z. Xie, Y. Huang, L. Jin, Y. Liu, Y. Zhu, L. Gao, and X. Zhang, "Weakly supervised precise segmentation for historical document images," *Neurocomputing*, vol. 350, pp. 271–281, Jul. 2019, doi: [10.1016/j.neucom.2019.04.001](https://doi.org/10.1016/j.neucom.2019.04.001).

[38] C. Zhang and W. L. Wang, "Character segmentation for historical Uchen Tibetan document based on structure attributes," *Laser Optoelectron. Prog.*, vol. 58, no. 20, Mar. 2021, Art. no. 2010020, doi: [10.3788/LOP202158.2010020](https://doi.org/10.3788/LOP202158.2010020).



**HUAMING LIU** received the Ph.D. degree in signal and information processing from the Nanjing University of Posts and Telecommunications, in 2020. He was a Visiting Scholar with the University of Science and Technology of China, Hefei, China, in 2015. He is currently an Associate Professor with the School of Computer and Information Engineering, Fuyang Normal University, Fuyang, China. His research interests include image processing, pattern recognition, and software engineering.



**GUOWEI ZHANG** is currently pursuing the master's degree with Northwest Minzu University, China. His research interests include artificial intelligence and computer vision.



**CE ZHANG** is currently pursuing the Ph.D. degree in Chinese language and literature with Northwest Minzu University, Lanzhou, China. His research interests include image processing, pattern recognition, especially focus on analysis and recognition of historical Tibetan document image.



**WEILAN WANG** (Member, IEEE) received the B.S. degree in mathematics from Northwest Normal University, Lanzhou, China, in 1983. She was a Visiting Scholar with Sun Yat-sen University, Guangzhou, China, in 1987. From 2001 to 2002, she was a Visiting Scholar with Tsinghua University, Beijing, China. From 2006 to 2007, she was a Visiting Scholar with Indiana University, Bloomington, IN, USA. She is currently a Professor and a Doctoral Supervisor with the Key Laboratory of China's Ethnic Languages and Information Technology, Ministry of Education, Northwest Minzu University, Lanzhou. Her research interests include image processing, pattern recognition, Tibetan information processing, and computer vision.



**QIANG LIN** received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, in 2014. He is currently an Associate Professor with the School of Mathematics and Computer Science, Northwest Minzu University. His research interests include medical image computing, pervasive computing, intelligent information processing, and human behavior sensing.

...