第 58 卷 第 20 期/2021 年 10 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于结构属性的乌金体藏文古籍字符切分

张策[1,2]，王维兰[1*]

[1]西北民族大学中国民族语言文字信息技术教育部重点实验室，甘肃 兰州 730030；

[2]重庆第二师范学院数学与信息工程学院，重庆 400065

摘要 字符切分是藏文古籍文档图像分析与识别中重要的一环，针对乌金体藏文古籍文本行倾斜，字符之间笔画交叠、交叉、粘连以及不同程度的笔画断裂、噪声干扰等问题，提出了一种基于结构属性的乌金体藏文字符切分方法。首先，建立了乌金体藏文古籍字符区块库。然后，利用音节点位置信息或结合水平投影与直线检测的方法检测出字符区块的局部基线，并根据基线将字符区块切分为上下两部分；利用改进的模板匹配算法检测基线上方笔画的粘连及其类型，利用多方向、多路径粘连切分算法切分交叉、粘连笔画。最后，根据藏文结构属性对各笔画进行归属，完成字符切分。实验结果表明，本方法能有效解决字符切分中遇到的问题，字符切分的召回率、精确率以及 F-Measure 可分别达到 96.52%、98.24%、97.37%。

关键词 图像处理；藏文古籍文档；字符区块；局部基线；粘连检测与切分；笔画归属

中图分类号 TP391.1 **文献标志码** A **doi**：10.3788/LOP202158.2010020

Character Segmentation for Historical Uchen Tibetan Document Based on Structure Attributes

Zhang Ce[1,2], Wang Weilan[1*]

[1] *Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu* 730030, *China*;

[2] *School of Mathematics and Information Engineering, Chongqing University of Education, Chongqing* 400065, *China*

Abstract Character segmentation is an important part in image analysis and recognition of historical Tibetan document. Aiming at the problems of text line slanting, stroke overlapping, crossing, touching between characters, stroke breaking and noise interference of historical Uchen Tibetan document, a character segmentation method for historical Uchen Tibetan document based on structure attributes is proposed in this paper. First, a character block dataset of historical Uchen Tibetan document is established. Then, the local baseline of character block is detected by using syllable point position information or combining horizontal projection and linear detection, and the character block is divided horizontally into two parts above and below the baseline. The improved template matching algorithm is used to detect touching strokes and touching type above the baseline. The multi-direction and multi-path touching character segmentation algorithm is used to realize crossing and touching strokes segmentation. Finally, according to Tibetan structure attribute, to complete the attribution of each stroke. Experimental results show that the proposed method can effectively solve the challenge problem in character segmentation. The recall rate, precision rate and F-Measure of character segmentation reached 96.52%, 98.24% and 97.37%, respectively.

Key words image processing; historical Tibetan document; character block; local baseline; touching strokes detection and segmentation; strokes attribution

OCIS codes 100.2000; 100.2960; 150.1135

收稿日期：2021-01-08；**修回日期**：2021-02-03；**录用日期**：2021-03-03

基金项目：国家自然科学基金(61772430)、国家民委创新团队计划((2018)98 号)、优秀研究生“创新之星”项目(2021CXZX-663)、重庆市教育委员会科学技术研究计划项目(KJQN202101608)、重庆第二师范学院校级科研项目(KY202118C)

通信作者：*wangweilan@xbmu.edu.cn

1 引 言

民族语言信息化处理是铸牢中华民族共同体意识的重要体现。历史久远且存量丰富的藏文古籍文档是藏文化的重要载体，对研究藏族历史、政治、经济、文化、医药等方面有重要的参考价值。让珍贵的藏文古籍文字“活”起来，也成为藏文古籍研究领域的一项重要任务。因此，2017 年起，北京工业大学、中国科学院、西北民族大学等逐渐开展了藏文古籍文档的分析与识别研究。

目前，人们对藏文古籍文档的分析与识别进行了大量研究[1-4]，包括对藏文古籍文档的预处理[5-6]、文本行切分[7-11]以及字符识别[12]等方面。其中，文本行切分成果为字符切分奠定了一定的基础，但该领域的研究还处于起步阶段。字符切分是藏文古籍文档分析与识别研究中的重点和难点内容，目前，针对藏文古籍字符切分的研究成果较少。Zhao 等[3]提出了一种基于特征点信息的字符切分方法，首先，检测出前景轮廓、骨架，提取特征点和基线信息；然后，利用支持向量机（SVM）分类器和距离规则分别移除上元音和辅音骨架端点附近的无用特征点，得到所有候选切分点并切分；最后，利用图论方法获得粘连的字符。此外，人们在其他文种的古籍字符切分方面也进行了相关研究。刘星辰等[13]在解决朝汉混排古籍文字的切分问题时，采用连通域投影方法完成列切分，并利用连通域删除、合并、拆分以及改进的滴水算法对字符或粘连字符进行切分。齐艳媚等[14]对古籍汉字版面图像进行连通域搜索，实现了对古籍汉字图像的切分，并通过建立犹豫模糊集判断过切分区域的隶属度，采用分段像素跳跃数突变分析方法切分粘连和重叠的汉字。周双飞等[15]先采用投影方法进行粗切分，将图像中的汉字分为粘连与非粘连两类；然后用粗切分统计信息设置切分路径，并基于最短路径思想获得最佳加权切分路径，完成粘连字符的切分。Sahare 等[16]提出了一种利用字符结构获得基本路径的方法，完成了拉丁文和梵文组成的多语种印度文档图像切分。Zaw 等[17]提出了一种基于字符连通域分析的方法，完成了缅甸语粘连字符的切分。Thongkanchorn 等[18]提出了一种基于 4 方向深度优先搜索算法的垂直和水平切分方法，完成了泰语字符的切分。Tamhankar 等[19]利用双阈值准则最大限度减少了切分误差，并提出了一种新的字符切分方法，解决了古代手写体草书 MODI（Microsoft office document imaging）文本字符的切分问题。

针对乌金体藏文古籍文本行倾斜，字符之间笔画交叠、交叉、粘连以及断裂等复杂的字符切分问题，Zhao 等[3]提出的藏文古籍字符切分方法未考虑到基线上方笔画交叉的情况，其他文种古籍的字符切分方法也不能很好地解决乌金体藏文古籍的字符切分问题。通过对文本行投影图的观察分析发现，文本行存在不同间距的空隙，因此，本文将长文本行切分成单独含有音节点、标点或其与字符组合的字符区块，在一定程度上减少长文本行整体倾斜对字符切分的影响，同时将字符之间的交叠（垂直投影重叠）、交叉、粘连以及断裂等问题分散到字符区块内，提出了一种基于结构属性的乌金体藏文字符切分方法。实验结果表明，本方法切分出的字符可以满足藏文古籍文档的进一步识别研究需求，对藏文古籍文档图像的分析与识别具有重要意义，同时为处理其他类古籍文档图像的字符切分提供了参考。

2 研究方法

2.1 藏文结构与古籍文本行

藏文句子由音节组成，音节具有“字”的意义，音节之间用音节点（SP）隔开。音节有严格的左右拼写及上下叠加规则，1 个音节最多包含前加字（PFC）、基字（BC）、上加字（SPC）、下加字（SBC）、上元音（TV）、下元音（BV）、后加字（SFC）、再后加字（FSFC）中的 7 个字母，且最多出现 1 个元音（上元音或下元音），基字所在字符的字母最多叠加 4 层。藏文总是沿着一条基线（BL）书写，基线是藏文的重要位置信息。在藏文古籍中，有大量的梵音藏文，梵音藏文只有上下叠加，最多可达 7 层，基线上方除上元音外，还有其他符号。因此，本方法将基线上方所有字母统称为基线上方笔画，基线下方字母或所有字母的叠加组合统称为基线下方笔画。藏文的音节结构以及梵音藏文如图 1 所示。

实验中的藏文古籍文档来源于北京版木刻《甘珠尔》，字体为乌金体。用专业相机对古籍文档进行拍照，形成电子文档图像，每张文档图像有 8 行文本，在文本框外左右两侧标记相关信息，如图 2 所示。

2.2 字符区块库

实验仅对字符切分进行研究，研究对象为对原始文档图像进行预处理（二值化、去噪、去边框、去文本行间粘连等）得到的文本，如图3所示。选取212

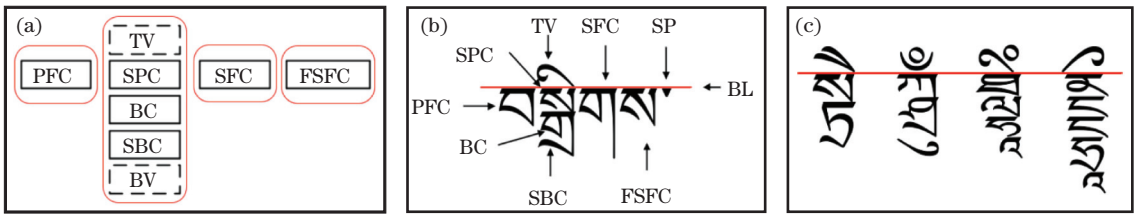


图 1　藏文音节。(a)藏文音节的结构;(b)藏文音节实例;(c)梵音藏文实例

Fig. 1　Syllable of the Tibetan. (a) Structure of the Tibetan syllable; (b) example of the Tibetan syllable; (c) examples of Tibetan transliteration of Sanskrit

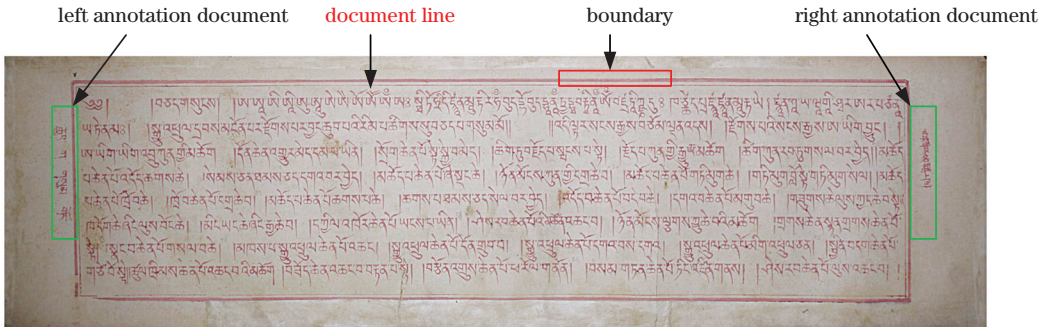


图 2　藏文古籍文档的原始图像

Fig. 2　Original image of the historical Tibetan document

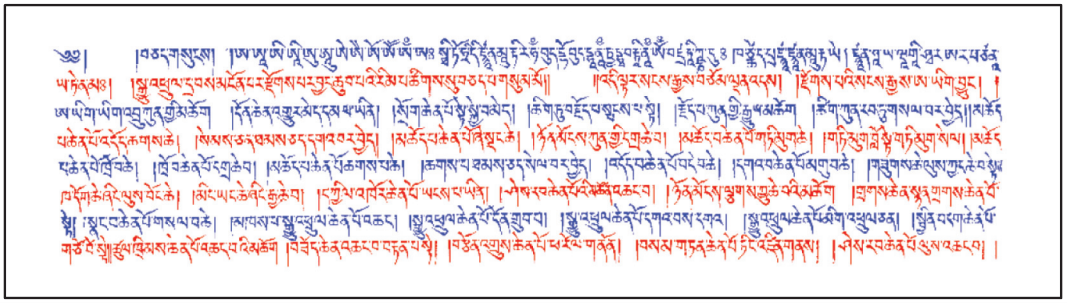


图 3　预处理后的藏文文本

Fig. 3　Tibetan document after pre-processing

张预处理后的文档图像(共 1696 行文本),利用水平投影空隙对文档进行水平切分,得到文本行。

得到文本行图像后,对每张文本图像进行垂直投影,将文本行切分成单独含有音节点、标点或其与字符组合的字符区块,共得到 109603 个字符区块,具体流程如图 4 所示。对投影图像进行垂直切分得到字符区块,并建立字符区块库,结果如图 5 所示。

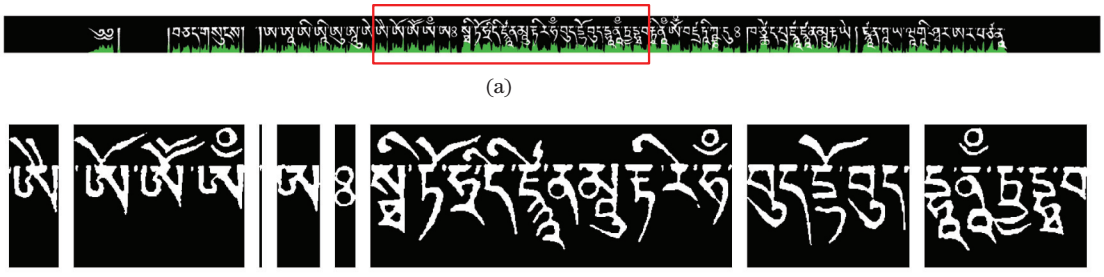


图 4　藏文古籍的投影垂直切分过程。(a)文本行及其垂直投影;(b)矩形区域的字符区块

Fig. 4　Vertical segmentation process by projection of the historical Tibetan document. (a) Document line and its vertical projection; (b) character blocks in rectangular area

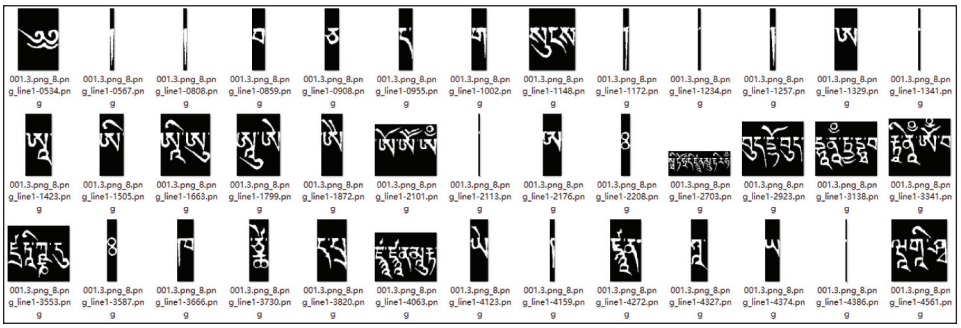


图 5　部分字符区块库

Fig. 5　Part of character block dataset

2.3　方法流程

藏文古籍文本行中字符之间具有笔画交叠、交叉、粘连以及断裂等问题，给字符切分带来挑战，具体的字符切分问题分类如表 1 所示，部分实例如图 6 所示。

表 1　字符切分问题的分类

Table 1　Classification of the character segmentation challenges

Label	Description
C1	overlapping strokes above the baseline
C2	crossing strokes above the baseline
C3	touching strokes above the baseline
C4	broken strokes above the baseline
C5	overlapping strokes below the baseline
C6	touching strokes below the baseline
C7	broken strokes below the baseline

从图 6 中可以发现，基线上方出现了 4 种切分问题，基线下方出现了 3 种切分问题，这些问题均会使字符切分变得极其困难。为了解决基线上方出现的 C1、C2、C3、C4 及基线下方出现的 C5、C7 字符切分问题，提出了一种基于结构属性的藏文古籍字符切分方法，具体流程如图 7 所示。其中，输入为字符区块，输出为藏文古籍字符。首先，获取输入字符区块的宽度 C_{width}，并对比 C_{width} 与平均字符宽度 C_{avgWidth} 的大小。若 $C_{\mathrm{width}}<0.5C_{\mathrm{avgWidth}}$，表明字符区块内为一个音节点或标点符号，利用连通域分析可得到一个完整符号；若 $C_{\mathrm{width}}\geqslant 0.5C_{\mathrm{avgWidth}}$ 且 $C_{\mathrm{width}}<1.5C_{\mathrm{avgWidth}}$，表明绝大多数字符区块内仅有一个字符，可能存在笔画断裂，需对字符的各笔画进行归属，完成字符切分；若 $C_{\mathrm{width}}\geqslant 1.5C_{\mathrm{avgWidth}}$，表明字符区块内至少存在一个字符，可能出现交叠、交叉、粘连以及断裂等现象，需进入多字符切分步骤。在多字符切分步骤中，先对字符区块的基线位置信息进行检测，并在基线处进行水平切分，形成基线上方笔画和下方笔画；然后获取基线上方笔画的粘连类型，对存在粘连的笔画进行切分，确定切分后笔画的类型，并用相同方法对基线下方的笔画断裂情况进行统计。最后，对字符各笔画进行归属，完成字符切分。

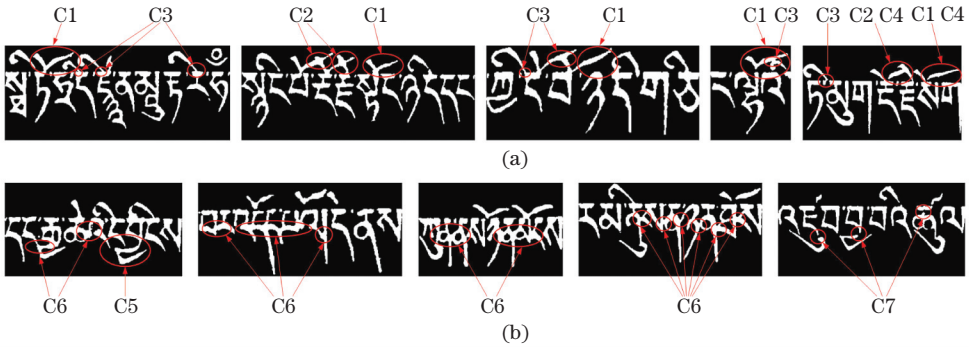


图 6　字符切分问题的实例。(a)基线上方的切分问题；(b)基线下方的切分问题

Fig. 6　Examples of the character segmentation challenges. (a) Segmentation challenges above the baseline; (b) segmentation challenges below the baseline

2.4　局部基线检测与水平切分

字符区块的局部基线位置信息可以通过字符水平投影检测得到，藏文字符结构的多样性导致检测出的基线位置信息并不准确。因此，提出了一种基于音节点位置信息或结合水平投影与直线检测的字符区块局部基线检测流程，如图8所示。其中，输入

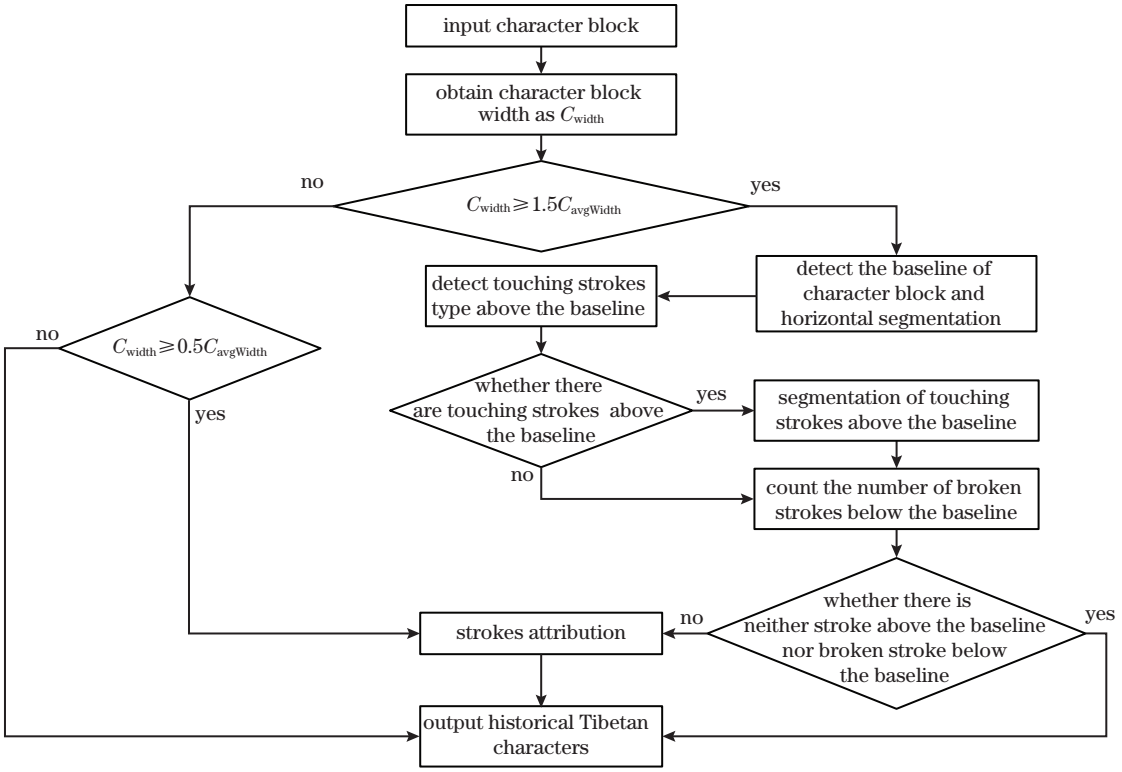


图 7 藏文古籍字符切分的流程图

Fig. 7 Flow chart of the character segmentation for historical Tibetan document

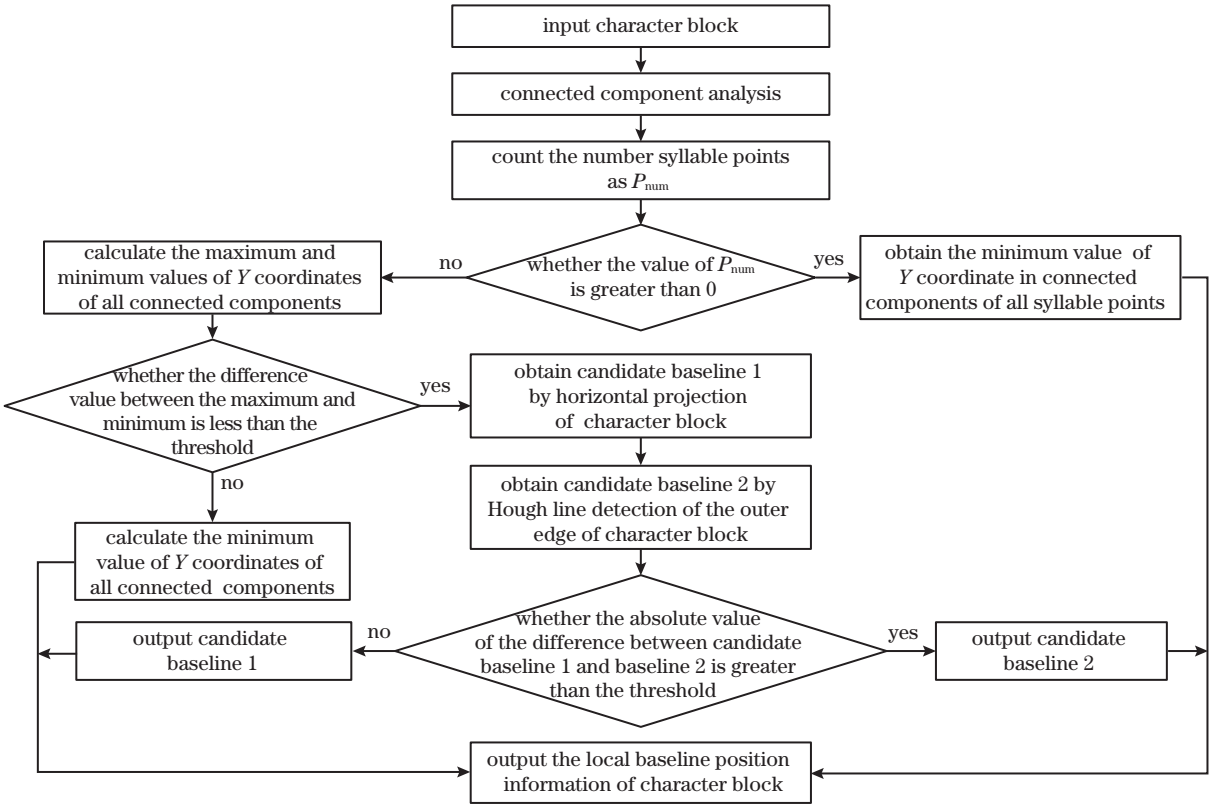


图 8 局部基线检测的流程图

Fig. 8 Flow chart of the local baseline detection

为字符区块，输出为局部基线。首先，对字符区块进行连通域分析，得到连通域的位置坐标、宽度、高度、面积及质心坐标等信息，并统计音节点的数量 P_{num}。然后，判断音节点数量 P_{num} 是否大于 0，若大于 0，表明该字符区块存在音节点，将所有音节点连通域的 Y 坐标最小值作为基线位置；否则，计算所有连通域 Y 坐标的最大值与最小值，判断最大值与最小值的差是否小于设定阈值。如果小于设定阈值，将最小值作为基线位置；否则，对字符区块进行水平投影和 Hough 直线检测，得到候选基线 1 和候选基线 2。最后，对比候选基线 1 和候选基线 2 的 Y 坐标差绝对值与阈值的大小，若绝对值大于阈值，则将候选基线 2 作为基线位置；否则，将候选基线 1 作为基线位置。获得字符区块局部基线后，在基线位置处进行水平切分，得到基线上方和下方两部分图像。

2.5　粘连笔画及其类型检测

笔画交叉是笔画粘连的一种特殊情况，对局部基线处进行水平切分后，基线上方部分由上元音及其他符号组成。从基线上方笔画中挑选出所有粘连笔画，并进行分类，结果如表 2 所示。其中，粘连组成表示粘连类型的组成笔画，粘连类型实例包括不同粘连方向及粘连程度的实例。可以发现，藏文基线上方的笔画尺寸较小，给区分粘连与非粘连造成一定困难，因此，将基线上方非粘连的笔画作为一类模板，以提高模板的匹配精度。利用每一类模板的平均尺寸对该类型中的所有模板进行尺寸归一化处理，形成模板库。

表 2　基线上方的粘连类型

Table 2　Touching type above the baseline

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

采用改进的模板匹配算法对基线上方笔画的粘连及类型进行检测。由于基线上方笔画尺寸较小，因此，直接用模板与待匹配图像对应的像素误差值作为匹配评判标准。但模板的种类较多，且不同模板类型之间的尺寸不统一，给匹配带来困难。因此，对传统基于误差值模板的匹配算法进行了改进，在匹配计算前先将待匹配的笔画尺寸调整为当前模板类型的尺寸，以实现匹配过程中尺寸的动态调整。粘连笔画及类型检测算法的具体步骤如下。

1）输入基线上方笔画 $S_{upperStrokes}$ 以及模板路径 $S_{temptTypePath}$，然后获取当前模板类型 $S_{temptType}$ 的尺寸，并以此对输入的 $S_{upperStrokes}$ 尺寸进行调整。

2）计算步骤 1）调整后的 $S_{upperStrokes}$ 与当前模板类型 $S_{temptType}$ 中所有模板 S_{tempt} 对应像素的误差值 S_{error}，可表示为

$$S_{error}=\sum_{m=1}^{M}\sum_{n=1}^{N}\left[S_{upperStrokes}(m,n)-S_{tempt}(m,n)\right]^2, \tag{1}$$

式中，M、N 分别为基线上方笔画和模板的高度与宽度。

3）计算步骤 2）得到的所有误差值 S_{error} 的最小值 $S_{minEachType}$。

4）重复步骤 2）和步骤 3），计算输入基线上方笔画 $S_{upperStrokes}$ 与所有模板类型 $S_{temptType}$ 对应模板 S_{tempt} 的误差值。

5）计算步骤 4）得到的所有模板类型误差值的最小值 $S_{minAllType}$。

6）根据步骤 5）得到的 $S_{minAllType}$ 从模板类型中查找与之对应的粘连类型 $S_{touchingType}$ 并输出，进而获得基线上方笔画的粘连数量、所在位置、粘连类型等信息。

2.6　粘连笔画切分

针对藏文古籍出现的复杂粘连问题，提出了一种多方向、多路径粘连切分算法。对藏文古籍粘连类型的统计观察发现，虽然粘连类型多达 14 种，但多数粘连类型能够在 45°、90°或 135°方向被正确切分。为解决少数粘连类型在这 3 个方向不能被正确切分的问题，进一步对坐标系内的切分方向进行细化，得到 7 个切分方向。粘连切分所用坐标系和切

分方向如图 9 所示，其中，O-XY 为图像坐标系，o-xy 为切分方向坐标系。在切分坐标系内，45°、90°及 135°方向分别对应 D_2、D_4 及 D_6，其他方向分别由 45°、135°对应正切函数值的 0.5 倍和 2 倍计算得到。在粘连切分 o-xy 坐标系的一个象限内，相邻 2 个切分方向的角平分线为切分方向的界线（如图中的虚线所示）。

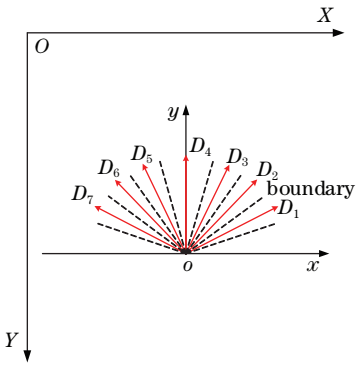


图 9　坐标系与切分方向的示意图

Fig. 9　Schematic diagram of coordinate system and segmentation direction

根据分支点左右延伸的像素量 P_{left} 和 P_{right} 与延伸阈值 $P_{\text{threshold}}$ 组合成 4 种大小关系，可表示为

$$\begin{cases} P_{\text{left}} > P_{\text{threshold}}, P_{\text{right}} > P_{\text{threshold}} \\ P_{\text{left}} > P_{\text{threshold}}, P_{\text{right}} \leqslant P_{\text{threshold}} \\ P_{\text{left}} \leqslant P_{\text{threshold}}, P_{\text{right}} > P_{\text{threshold}} \\ P_{\text{left}} \leqslant P_{\text{threshold}}, P_{\text{right}} \leqslant P_{\text{threshold}} \end{cases}。\tag{2}$$

根据藏文古籍字符的结构特点，用不同的组合关系形成不同的切分路径，多路径切分示例如图 10 所示。其中，图 10(a)为 P_{left} 和 P_{right} 大小组合实例，图 10(b)为已完成分支点和端点标记的骨架图，图 10(c)为粘连笔画的切分示意路径。

值得注意的是，图 10 中展示的路径并非该笔画的实际切分路径，在实际粘连笔画切分过程中，需结合切分起点、切分方向与路径完成切分。多方向、多路径粘连切分算法的具体步骤如下。

1）输入基线上方粘连笔画、粘连数量、粘连位置、粘连类型及字符区块的基线位置信息，然后对基线上方粘连笔画进行骨架化处理，得到骨架图。在骨架图的一定范围内查找分支点，若不存在分支点，则进入步骤 2）；否则，进入步骤 3）。

2）结合粘连类型，在 D_4 方向对粘连笔画进行切分。如“  ”粘连类型在其宽度的 1/3 处进行切分，“  ”粘连类型在其宽度的 1/2 处进行切分，“”粘连类型在其宽度的 2/3 处进行切分。

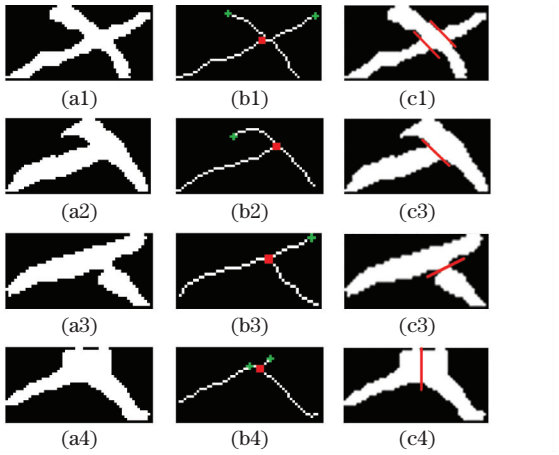


图 10　多路径切分示例。(a)组合实例；(b)标记的骨架图；(c)切分路径

Fig. 10　Examples of multipath segmentation. (a) Combination example; (b) marked skeleton diagram; (c) segmentation path

3）记录分支点的坐标信息。若在笔画分支处检测出多个候选分支点，则取 X 坐标最小的分支点作为该笔画的分支点 $P_{\text{branchPoint}}$。

4）结合粘连类型和分支点 $P_{\text{branchPoint}}$ 确定切分起点 $P_{\text{segStartPoint}}$。

5）以骨架图的分支点 $P_{\text{branchPoint}}$ 为起点，在一定范围内计算骨架向左和向右延伸的像素量 P_{left} 和 P_{right}。

6）计算切分起点 $P_{\text{segStartPoint}}$ 分别与左右延伸像素端点 $P_{\text{leftEndPoint}}$ 和 $P_{\text{rightEndPoint}}$ 构成的直线斜率 K，以 $P_{\text{rightEndPoint}}$ 为例，K 可表示为

$$K = \text{abs}\left(\frac{P_{\text{leftEndPointY}} - P_{\text{segStartPointY}}}{P_{\text{leftEndPointX}} - P_{\text{segStartPointX}}}\right)。\tag{3}$$

7）根据步骤 4）计算的直线斜率 K 选择相应方向作为切分方向，其中，arctan(K)为切分起点 $P_{\text{segStartPoint}}$ 与左右延伸笔画端点 $P_{\text{leftEndPoint}}$ 和 $P_{\text{rightEndPoint}}$ 构成的直线斜率对应的度数。根据 arctan(K)值和界线 P_{boundary}，选择与 arctan(K)最邻近的方向作为切分方向。

8）结合粘连类型 $S_{\text{touchingType}}$、P_{left} 和 P_{right}、$P_{\text{threshold}}$ 的大小及不同组合，从切分起始点 $P_{\text{segStartPoint}}$ 出发以不同切分路径对粘连字符进行切分。

9）若笔画存在多处粘连，则重复步骤 3）～步骤 8），并输出粘连切分后的笔画。

粘连切分后，需确定字符区块基线上方笔画的类型，为后续笔画归属提供依据。设 Y_{left} 为笔画连通域左侧第一列前景像素中对应的最大 Y 坐标，Y_{right} 为笔画连通域右侧最后第一列前景像素对应

的最大 Y 坐标，Y_{centroid} 为笔画连通域质心的 Y 坐标。藏文古籍中有大量的梵音藏文，基线上方笔画类型也会相应增多，但大部分笔画不会影响字符切分。因此，总结了基线上方可能影响字符切分效果的笔画类型及其几何特征，如表 3 所示。

表 3　基线上方的笔画类型及几何特征

Table 3　Stroke types and geometric characteristics above the baseline

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

根据表 3 中各类型笔画的几何特征，统计基线上方笔画的类型和数量。若字符区块中同时出现表 3 中的 1 号和 9 号笔画，则需要进一步增加判断条件，原因是 1 号笔画与 9 号笔画的左侧笔画具有相同的几何特征。即 1 号笔画单独出现，无右侧笔画与其配对，而 9 号笔画的左右两个笔画需成对出现。

2.7　断裂笔画统计

除基线上方的粘连问题外，断裂也是藏文古籍文本的普遍现象，会严重影响字符切分效果。基线上方的断裂（如 9 号笔画）可通过表 3 的基线上方笔画类型及几何特征确定。观察基线下方断裂的情况可知，断裂常出现在纵向笔画较细位置，因此，总结出了 4 种断裂情况，如图 11 所示。其中，A 和 B 分别为不同断裂笔画连通域外接矩形框的质心。

根据 4 种断裂类型，归纳出判断基线下方笔画断裂的方法，除笔画连通域位于基线下方的基本条件外，笔画连通域质心的 X 坐标、上边界的 Y 坐标以及面积需同时满足

$$\begin{cases}\text{abs}(A_X^{\text{centroid}}-B_X^{\text{centroid}})<X_{\text{centroidThresold}}\\ \text{abs}(A_Y^{\text{upper}}-B_Y^{\text{upper}})>Y_{\text{upperThresold}}\\ A_{\text{area}}>P_{\text{areaThreshold}}\\ B_{\text{area}}>P_{\text{areaThreshold}}\end{cases},\tag{4}$$

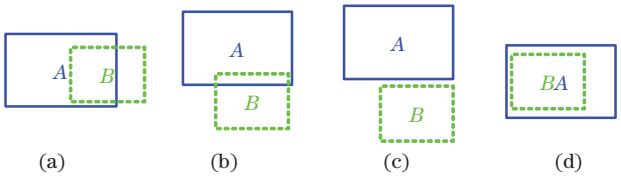


图 11　基线下方笔画的断裂情况。(a)左右交叉；(b)上下交叉；(c)上下相离；(d)包含

Fig. 11　Broken strokes type below the baseline. (a) Cross left and right; (b) cross up and down; (c) separate up and down; (d) contain

式中，A_X^{centroid}、B_X^{centroid} 及 $X_{\text{centroidThresold}}$ 分别为连通域质心 A、B 的 X 坐标及连通域质心的水平距离阈值，A_Y^{upper}、B_Y^{upper} 及 $Y_{\text{upperThresold}}$ 分别为连通域上边界的 Y 坐标及连通域上边界的距离阈值，A_{area}、B_{area} 为连通域的面积，$P_{\text{areaThreshold}}$ 为音节点的面积阈值。设置音节点的面积阈值可避免音节点对断裂判断的影响。

2.8　笔画归属

笔画归属是字符切分的最后阶段，具体是将字符各笔画按照正确的文字结构放到对应的位置。根据字符区块的宽度，将笔画归属初次划分为三类进行处理，即不需要归属、单字符的归属以及多字符的归属。对于多字符的归属，根据基线上方笔画类型、数量及基线下方笔画断裂数量等情况，再次划分为两类。第一类是基线上方无笔画且基线下方无断裂，各连通域均为字符；第二类是除第一类以外的所有情况组合，需计算字符区块内各笔画之间的质心水平距离 $D_{\text{centroid}\,X}$，若小于距离阈值 $D_{\text{threshold}}$，则将其存入待归属数组 A_{merge} 中；反之，则连通域为字符。根据藏文古籍字符的特点，归属前需对数组 A_{merge} 进行修正，并对数组 A_{merge} 中的笔画进行归属，笔画归属分类实例如图 12 所示。

笔画归属算法的具体步骤如下。

1）输入字符区块的各个连通域，并获取输入字符区块的宽度 C_{width}。

2）对比当前字符宽度 C_{width} 与平均字符宽 C_{avgWidth} 的大小。若 $C_{\text{width}}<0.5C_{\text{avgWidth}}$，则字符区块内的连通域为音节点或标点符号；若 $C_{\text{width}}\geqslant 0.5C_{\text{avgWidth}}$，且 $C_{\text{width}}<1.5C_{\text{avgWidth}}$，则将各笔画归属为一个完整字符；若 $C_{\text{width}}\geqslant 1.5C_{\text{avgWidth}}$，表明字符区块内有多个字符，进入步骤 3）的多字符归属。

3）判断字符区块是否满足基线上方无笔画且基线下方无笔画断裂。若满足，则各连通域均为字符；否则，进入步骤 4）。

4）对比字符区块所有笔画的质心水平距离 $D_{\text{centroid}\,X}$，将 $D_{\text{centroid}\,X}<D_{\text{threshold}}$ 的笔画存入待归属数组 A_{merge} 中；否则，认为连通域为字符。

5）根据藏文古籍字符的结构特点，对数组 A_{merge} 进行修正。

6）先确认 A_{merge} 中是否存在基线上方笔画，若存在，则以该笔画为基础根据质心水平距离阈值 $D_{\text{threshold}}$ 在基线上方和下方查找同属于一个字符的笔画，完成归属；否则，只需对基线下方的断裂笔画进行归属，并输出藏文古籍字符。

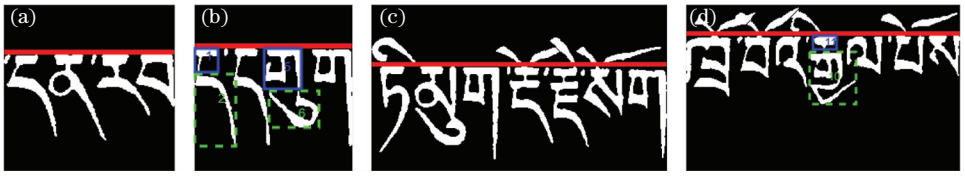


图 12　笔画归属的分类实例。(a)基线上方无笔画且基线下方无断裂；(b)基线上方无笔画且基线下方有断裂；(c)基线上方有笔画且基线下方无断裂；(d)基线上方有笔画且基线下方有断裂

Fig. 12　Examples of strokes attribution classification. (a) With no stroke above the baseline and with no broken stroke below the baseline; (b) with strokes above the baseline and with no broken stroke below the baseline; (c) with no stroke above the baseline and with no broken stroke below the baseline; (d) with strokes above the baseline and with broken strokes below the baseline

2.9　字符切分评价方法

为了客观评价本方法的优劣，采用召回率(Recall)、精确率(Precision)、加权调和平均度量(F-Measure)以及字符切分错误率(Error)评价字符切分效果，可表示为

$$N_{\text{Recall}}=\frac{N_{\text{CSC}}}{N_{\text{TCC}}}\times 100\%,\tag{5}$$

$$N_{\text{Precision}}=\frac{N_{\text{CSC}}}{N_{\text{TSC}}}\times 100\%,\tag{6}$$

$$N_{\text{F-Measure}}=\frac{2\times N_{\text{Recall}}\times N_{\text{Precision}}}{N_{\text{Recall}}+N_{\text{Precision}}}\times 100\%,\tag{7}$$

$$N_{\text{Error}}=\frac{N_{\text{WSC}}}{N_{\text{TCC}}}\times 100\%,\tag{8}$$

式中，N_{CSC} 为正确切分数量，N_{TCC} 为总字符数量，N_{TSC} 为切分的数量，N_{WSC} 为错误切分的数量。

3　实验结果与分析

3.1　实验过程

1）局部基线检测与水平切分

对于有音节点的字符区块，根据音节点与基线处于同一水平位置的特点，通过获取音节点的位置信息实现基线检测；对于无音节点且基上方无笔画的字符区块，通过对比各笔画连通域 Y 坐标的距离实现基线检测；对于无音节点且基上方有笔画的字符区块，利用水平投影与 Hough 直线检测算法相结合的方法实现基线检测。基线检测后在基线位置进行水平切分。字符区块局部基线检测与水平切分的过程如图 13 所示。

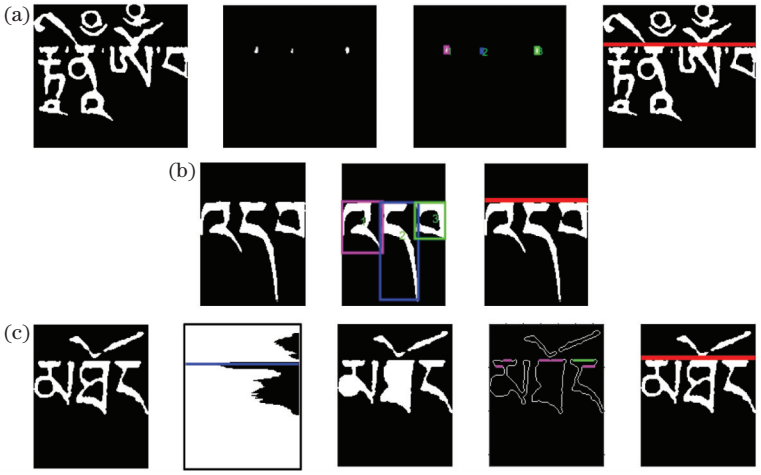


图 13　基线检测与水平切分的过程。(a)有音节点的字符区块；(b)无音节点且基线上方无笔画的字符区块；(c)无音节点且基线上方有笔画字的字符区块

Fig. 13　Process of local baseline detection and horizontal segmentation of character block. (a) Character blocks with syllable points; (b) character blocks with no syllable point and with no stroke above the baseline; (c) character blocks with no syllable point and with strokes above the baseline

2）粘连笔画及类型检测

对基线水平切分后的上方笔画进行粘连及类型检测，若存在粘连，则得到粘连笔画及其类型；否则，认为基线上方笔画不存在粘连。基线上方粘连及类型检测实例如表 4 所示，其中，图像中的矩形框为字符区块基线上方笔画，序号与表 1 中的序号相对应，误差最小值对应的粘连类型为粘连及类型检测算法的输出结果。可以发现，该字符区块基线上方笔画检测的误差最小值对应的序号为 3，对应表 1 中的粘连类型“  ”。

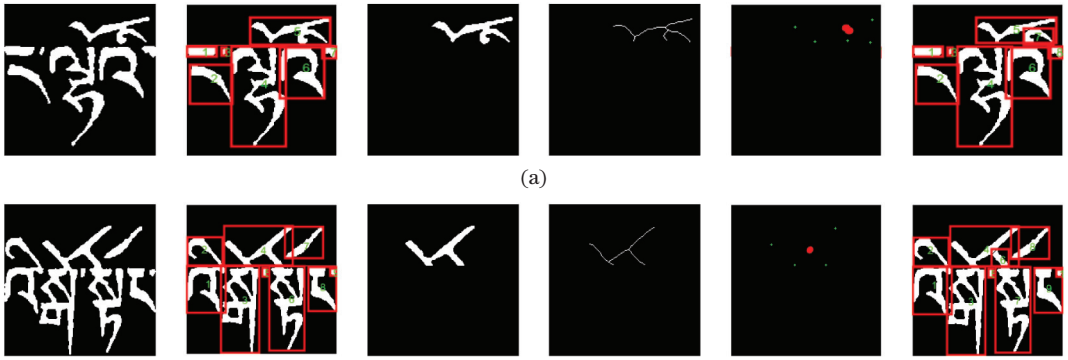
表 4　基线上方笔画粘连及类型检测

Table 4　Touching stroke and type detection above the baseline

	No.	1	2	3	4	5	6	7
	T_{error}	939	555	**106**	1125	585	364	1137
	No.	8	9	10	11	12	13	14
	T_{error}	893	1183	1155	682	1577	1363	1889

3）粘连笔画切分

用多方向、多路径粘连切分算法对基线上方笔画出现的各种复杂粘连、交叉等情况进行切分，结果如图 14（字符区块有一处粘连）和图 15（字符区块有多处粘连）所示。



(a)

(b)

图 14　有一处粘连的字符切分。(a)切分方向为 D_1；(b)切分方向为 D_2

Fig. 14 Character segmentation with a touching stroke. (a) Character direction is D_1; (b) character direction is D_2

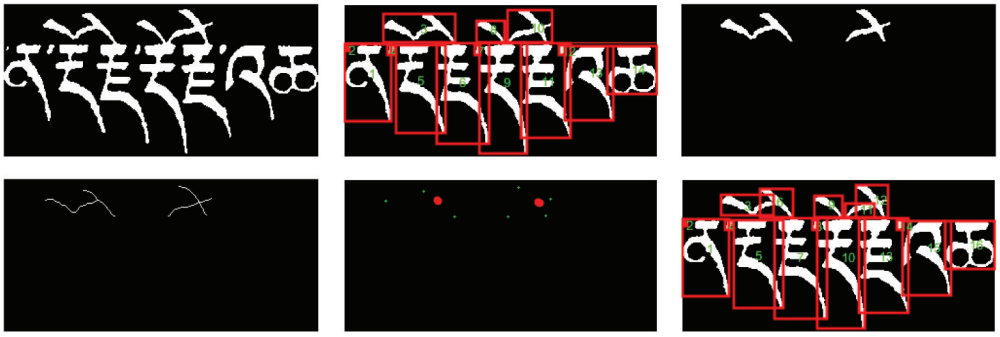


图 15　有多处粘连的字符切分

Fig. 15　Character segmentation with multiple touching strokes

交叉是粘连的一种特殊情况，交叉笔画会在分支点左上方和右上方有不同长度的延伸笔画。根据延伸笔画的长度判断是否对其进行多路径切分并删除，是一种简单且有效的交叉切分方法。基线上方笔画交叉切分效果的实例如图 16（字符区块有多处交叉）所示。

4）断裂笔画统计

根据字符区块内各笔画的连通域质心、位置坐标及边界信息对基线下方的断裂笔画进行统计。常见的几种断裂情况如图 17 所示，其中，实线方框和虚线方框代表不同笔画的最小外接矩形框。

5）笔画归属

表 3 中 3 号至 8 号笔画类型在藏文古籍文本中与其基线下方笔画左右位置的偏移较小，因此，可利用各笔画的质心水平距离完成归属，如图 18 所示。根据质心水平距离 $D_{centroid\,X}$ 将基线上方和下方的笔

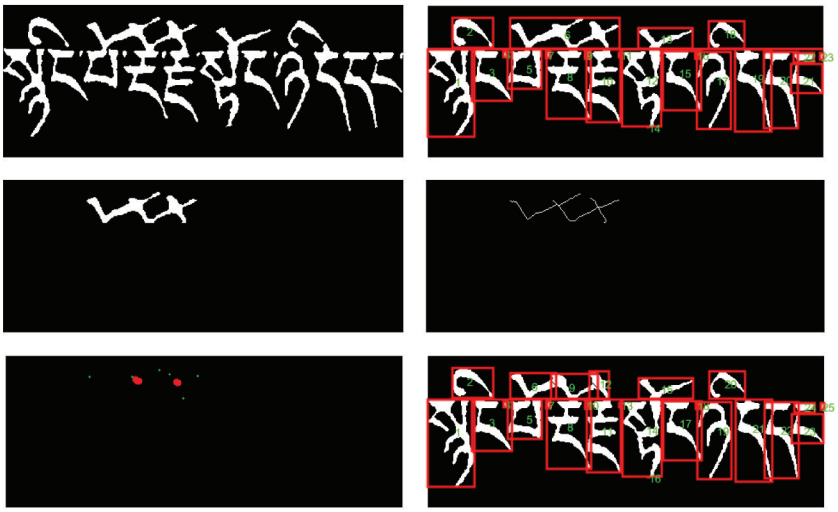


图 16　有多处交叉的字符切分

Fig. 16　Character segmentation with multiple crossing strokes

图 17　基线下方笔画的断裂统计结果。(a)左右交叉;(b)上下交叉;(c)上下相离;(d)包含

Fig. 17　Statistical results of broken strokes below the baseline. (a) Cross left and right; (b) cross up and down; (c) separate up and down; (d) contain

图 18　基于质心水平距离的归属。(a)字符区块;(b)归属后笔画的质心

Fig. 18　Attribution based on the horizontal distance of the centroid. (a) Character block; (b) centroid of strokes after attribution

画归属为一个完整的字符,归属后基线上下笔画的质心用同一符号标记。

表 3 中的 1 号“”、2 号“”及 9 号“”类型具有一定的书写特点或断裂问题,导致质心坐标信息不能将所有包含此类笔画的字符都归属正确。其中,2 号笔画是由两个 1 号笔画左右叠加而成,归属方法相同,因此,仅阐述 1 号笔画和 9 号笔画类型的归属结果。1 号笔画类型在藏文古籍字符中与其基线下方笔画左右偏移的大小不稳定,在一些字符中处于靠左的位置,而在另一些字符中处于靠右的位置。因此,利用质心水平距离归属此类笔画并不能完全解决归属问题,如图 19 所示。根据质心水平距离归属可知,L_{left} 为“”质心与左侧字符笔画质心的水平距离,L_{right} 为“”质心与其原字符基线下方笔画质心的水平距离。其中,图 19(a)中“”的质心与其原字符基线下方笔画的质心水平距离小于与左侧字符质心的水平距离,图 19(b)的情况恰好相反。可以发现,利用字符笔画连通域右侧边界的坐标信息与其基线下方笔画进行归属,可以减小笔画左右偏移对这类笔画归属的影响。

9 号“”笔画类型是由 6 号“”类型断

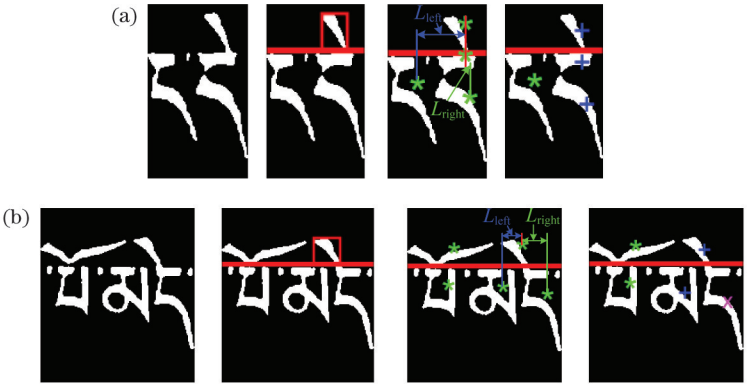


图 19　笔画的归属分析。(a)1 号;(b)9 号

Fig. 19　Attribution analysis of the stroke. (a) No. 1; (b) No. 9

裂而成，断裂后的左笔画与 1 号“ ”笔画类型具有相同的几何特征，且断裂后的右笔画常处于其右侧相邻字符的基线上方，增加了归属的难度。因此，利用左右笔画的质心水平坐标计算该笔画类型的整体质心水平坐标 A_X^{centroid}，可表示为

$$A_X^{\text{centroid}}=(A_X^{\text{leftCentroid}}+A_X^{\text{rightCentroid}})/2, \tag{9}$$

式中，$A_X^{\text{leftCentroid}}$ 为左侧笔画质心的水平坐标，$A_X^{\text{rightCentroid}}$ 为右侧笔画质心的水平坐标。此类笔画的归属过程如图 20 所示，可以发现，基线上虚线方框中笔画的质心与原字符基线下方笔画质心的水平距离明显大于与右侧字符笔画的质心水平距离，直接利用笔画质心不能将其正确归属。通过(9)式的质心水平距离中值，可以计算“ ”笔画的左侧笔画(实线方框标记)质心与右侧笔画(虚线方框标记)质心的水平距离中值，使更新后的质心(用字母“c”标记)“前移”，从而解决归属问题。

图 20　“ ”笔画的归属分析

Fig. 20　Attribution analysis of “ ” stroke

1 号笔画与 9 号笔画以左右相邻位置关系出现在同一个字符区块内，使笔画归属变得更加困难。如图 21 所示，字符区块基线上方同时出现了 1 号笔画和 9 号笔画类型，且以“ ”、“ ”、“ ”笔画类型交替的方式出现。其中，“ ”左笔画距离其原字符基线下方笔画质心水平距离较远，右笔画质心水平距离更远，且右笔画的右侧紧挨着“ ”，这些情况同时出现后，容易导致归属错误。利用(9)式对基线上方笔画质心进行更新，将更新后的质心分别用字母“c”和“v”标记。解决此类问题时需结合基线上方笔画的质心水平距离与笔画类型，且以基线上方笔画为基础对字符各笔画进行归属。

图 21　“ ”和“ ”两种类型同时出现的归属分析

Fig. 21　Attribution analysis of both “ ” and “ ” stroks

藏文古籍字符切分的难点集中在字符切分阶段，特别是字符之间存在交叠、交叉、粘连等现象时，切分难度更大。字符切分的最终结果如图 22 所示。

图 22　字符切分结果。(a)字符区块；(b)切分后的字符区块

Fig. 22　Results of character segmentation. (a) Character block; (b) block after character segmentation

3.2　实验结果

藏文古籍字符切分先后经过字符区块库建立和字符切分两个阶段。建立字符区块库阶段时，对 212 张藏文古籍文档(包含 1696 行文本)采用垂直投影完成切分，共得到 109603 个字符区块，并建立字符区块库。字符区块库中仅有一个音节点或标点符号的字符区块有 22418 个，基线下方存在粘连的字符区块有 5685 个。字符切分阶段可解决基线上方出现的 C1、C2、C3、C4 及基线下方出现的 C5、C7 字符切分问题，共计 81500 个字符区块。依次对字符区块进行局部基线检测与水平切分、粘连及类型检测、粘连切分以及笔画归属等处理步骤，完成字符切分。

1) 建立字符区块库阶段的相关数据

正确切分数量 N_{CSC} 为文本行经过垂直投影切分后的所有笔画能全部正确归属为字符的字符区块数量，切分数量 N_{TSC} 为投影垂直切分后的所有笔画不能全部正确归属为字符的字符区块个数。文本行垂直投影切分前无实际字符区块数量，因此无召回率。建立字符区块数据库阶段的正确切分数据如表 5 所示。

表 5　字符区块数据库的正确切分数据

Table 5　Correct segmentation data of character block dataset

N_{CSC}	N_{TSC}	N_{Recall}/%
109354	109603	99.77

2) 字符切分阶段的相关数据

切分数量 N_{TSC} 为切分出的字符总数，正确切分数量 N_{CSC} 为正确切分出的字符数量，实际数量 N_{TCC} 为字符区块内所有的字符数量。字符切分阶段的正确切分数据如表 6 所示。

表 6　字符切分阶段正确切分的数据

Table 6　Data of the correct segmentation in the character segmentation stage

N_{CSC}	N_{TC}	N_{TSC}	N_{Recall}/%	$N_{Precision}$/%	$N_{F\text{-}Measure}$/%
176802	183987	180379	96.09	98.02	97.05

为了进一步评估字符切分阶段的字符切分效果，对字符切分各步骤出现的错误字符切分数量占比 $N_{Proportion}$ 及对应的错误切分率 N_{Error} 进行统计，结果如表 7 所示。

表 7　字符切分阶段各步骤的 N_{Error}

Table 7　N_{Error} for each step during character segmentation

Character segmentation steps	N_{WSC}	$N_{Proportion}$/%	N_{Error}/%
Build character block dataset	249	3.46	0.14
Detect the local baseline and horizontal segmentation	962	13.39	0.52
Detection of touching strokes type	267	3.72	0.15
Segmentation of touching strokes	25	0.35	0.01
Strokes attribution	5682	79.08	3.09

结合字符区块建立阶段和字符切分阶段的切分数据，对字符切分效果进行整体评估，结果如表 8 所示。建立字符区块库阶段得到 22418 个字符区块中仅含有音节点或标点符号，不需要进行字符切分。

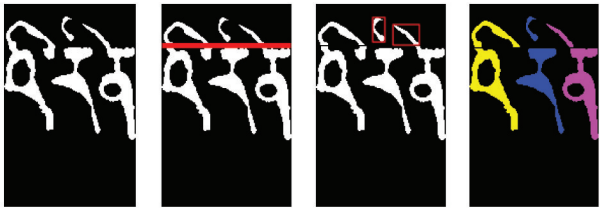
表 8　正确切分的数据

Table 8　Correctly segmented data

N_{CSC}	N_{TCC}	N_{TSC}	N_{Recall}/%	$N_{Precision}$/%	$N_{F\text{-}Measure}$/%
199220	206405	202797	96.52	98.24	97.37

3.3　实验结果分析

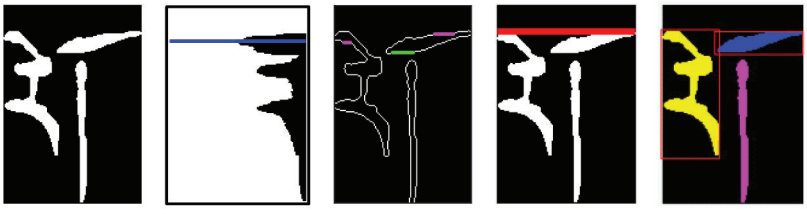
本方法虽然能有效解决乌金体藏文古籍文本行倾斜、字符之间笔画交叠、交叉、粘连、断裂及噪声干扰等问题，但仍存在部分问题，如乌金体藏文古籍字符笔画出现严重断裂，且断裂后的笔画与其他笔画极其相似时，会导致字符切分错误，如图 23 所示。其中，图 23(c)中矩形框标记的两个笔画由上元音“ ”断裂而成，左侧部分被检测为其他笔画，右侧部分被检测为上元音“ ”。此外，乌金体藏文古籍字符的固有特点以及有些笔画的书写近乎水平位置时，本方法也不能准确检测出局部基线，如图 24 所示。该字符区块无音节点，需结合水平投影和直线检测完成基线检测。因基线上方“ ”类型的右侧笔画近乎水平，导致水平投影产生的候选基线 1 偏离实际基线位置，而 Hough 直线检测产生的候选基线 2 在右侧笔画的下端，影响了局部基线检测的准确性。



(a)　(b)　(c)　(d)

图 23　笔画归属导致的字符切分错误。(a)字符区块；(b)局部基线与水平切分；(c)断裂笔画标记；(d)字符切分结果

Fig. 23　Wrong character segmentation caused by strokes attribution. (a) Character block; (b) local baseline and horizontal segmentation; (c) broken stroke mark; (d) result of character segmentation



(a)　(b)　(c)　(d)　(e)

图 24　基线检测导致的字符切分错误。(a)字符区块；(b)水平投影；(c)Hough 直线检测；(d)局部基线；(e)字符切分结果

Fig. 24　Wrong character segmentation caused by the baseline detection. (a) Character block; (b) horizontal projection; (c) Hough straight line detection; (d) local baseline; (e) result of character segmentation

4　结　　论

根据乌金体藏文古籍字符结构属性对藏文古籍文档图像的文本行字符进行切分。首先，采用垂直投影将文本切分成字符区块，建立字符区块库；然后，依次对基线下方无粘连的字符区块完成局部基线检测、粘连及其类型检测、粘连切分以及断裂笔画统计；最后，对字符各笔画进行归属，得到藏文古籍

字符，完成字符切分。实验结果表明，本方法能有效解决基线上方字符之间笔画交叠、交叉、粘连以及不同程度的笔画断裂、噪声干扰等复杂情况，但仍然存在极少数字符不能被正确切分的情况。下一步工作将对本方法中未解决的问题进行深入研究，并结合深度学习技术与传统方法提高字符切分的召回率和精确率。

参 考 文 献

[1] Duan L J, Zhang X Q, Ma L L, et al. Text extraction method for historical Tibetan document images based on block projections [J]. Optoelectronics Letters, 2017, 13(6): 457-461.

[2] Zhang X Q, Ma L L, Duan L J, et al. Layout analysis for historical Tibetan documents based on convolutional denoising autoencoder[J]. Journal of Chinese Information Processing, 2018, 32(7): 67-73, 81.
张西群, 马龙龙, 段立娟, 等. 基于卷积降噪自编码器的藏文历史文献版面分析方法[J]. 中文信息学报, 2018, 32(7): 67-73, 81.

[3] Zhao Q C, Ma L L, Duan L J. A touching character database from Tibetan historical documents to evaluate the segmentation algorithm[M]//Lai J H, Liu C L, Chen X L, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2018, 11259: 309-321.

[4] Li Y X, Ma L L, Duan L J, et al. A text-line segmentation method for historical Tibetan documents based on baseline detection[M]//Yang J F, Hu Q H, Cheng M M, et al. CCCV 2017: computer vision. Communications in computer and information science. Singapore: Springer, 2017, 771: 356-367.

[5] Han Y H, Wang W L, Liu H M, et al. A combined approach for the binarization of historical Tibetan document images[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(14): 1954038.

[6] Li Z J, Wang W L, Cai Z Q. Historical document image binarization based on edge contrast information[M]//Arai K, Kapoor S. CVC 2019: advances in computer vision. Advances in intelligent systems and computing. Cham: Springer, 2019, 943: 614-628.

[7] Zhou F M, Wang W L, Lin Q. A novel text line segmentation method based on contour curve tracking for Tibetan historical documents[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2018, 32(10): 1854025.

[8] Wang Y Q, Wang W L, Li Z J, et al. Research on text line segmentation of historical Tibetan documents based on the connected component analysis[M]//Lai J H, Liu C L, Chen X L, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2018, 11258: 74-87.

[9] Li Z J, Wang W L, Chen Y, et al. A novel method of text line segmentation for historical document image of the Uchen Tibetan[J]. Journal of Visual Communication and Image Representation, 2019, 61: 23-32.

[10] Li J C, Wang X J, Wang W L, et al. Text line segmentation of Tibetan historical documents based on text core regions combined with expansion growth[J]. Laser & Optoelectronics Progress, 2021, 58(2): 021008.
李金成, 王筱娟, 王维兰, 等. 结合文字核心区域和扩展生长的藏文古籍文本行切分[J]. 激光与光电子学进展, 2021, 58(2): 021008.

[11] Li Z J, Wang W L. Tibetan historical document recognition of uchen script using baseline information[J]. Proceedings of SPIE, 2019, 11069: 110693H.

[12] Wang X J, Wang W L, Li Z J, et al. A recognition method of the similarity character for uchen script Tibetan historical document based on DNN[M]//Lai J H, Liu C L, Chen X L, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2018, 11258: 52-62.

[13] Liu X C, Jin X F. Characters segmentation method of historical documents mixed in Korean and Chinese[J]. Computer Engineering and Applications, 2020, 56(11): 135-141.
刘星辰, 金小峰. 朝汉混排古籍的文字切分方法[J]. 计算机工程与应用, 2020, 56(11): 135-141.

[14] Qi Y M, Tian X D, Zuo L N. Segmentation method of ancient Chinese character images based on hesitant fuzzy sets[J]. Science Technology and Engineering, 2019, 19(30): 232-240.
齐艳媚, 田学东, 左丽娜. 基于犹豫模糊集的古籍汉字图像切分方法[J]. 科学技术与工程, 2019, 19(30): 232-240.

[15] Zhou S F, Liu C P, Liu G, et al. Multi-step segmentation method based on minimum weight segmentation path for ancient handwritten Chinese character[J]. Journal of Chinese Computer Systems, 2012, 33(3): 614-620.
周双飞, 刘纯平, 柳恭, 等. 最小加权分割路径的古籍手写汉字多步切分方法[J]. 小型微型计算机系统, 2012, 33(3): 614-620.

[16] Sahare P, Dhok S B. Multilingual character segmentation and recognition schemes for Indian

document images[J]. IEEE Access, 2018, 6: 10603-10617.

[17] Zaw K P, War N. Y-position based Myanmar touching character segmentation and sub-components based character classification[C]//2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), May 29-31, 2019, Honolulu, HI, USA. New York: IEEE Press, 2019: 76-83.

[18] Thongkanchorn K, Kanchanapreechakorn S, Borwarnginn P, et al. Thai character segmentation in handwriting images using four directional depth first search[C]//2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), October 10-11, 2019, Pattaya, Thailand. New York: IEEE Press, 2019: 1-5.

[19] Tamhankar P A, Masalkar K D, Kolhe S R. A novel approach for character segmentation of offline handwritten Marathi documents written in MODI script[J]. Procedia Computer Science, 2020, 171: 179-187.