第 58 卷 第 14 期/2021 年 7 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

基于自适应游程平滑算法的藏文文档图像版面分割与描述

陈园园[1]，王维兰[2*]，刘华明[3]，蔡正琦[1]，赵鹏海[2]

[1]西北民族大学数学与计算机科学学院，甘肃 兰州 730030；

[2]西北民族大学中国民族语言文字处理教育部重点实验室，甘肃 兰州 730030；

[3]阜阳师范大学计算机与信息工程学院，安徽 阜阳 236041

摘要 版面分割是文档图像分析与识别过程中的重要基础步骤，为了探索适用于藏文文档图像版面分割与描述的方法，提出一种基于自适应游程平滑算法的研究方法。根据藏文文档图像的版面结构，利用 K 均值聚类分析得到适用于版面的游程阈值，进行游程平滑，寻找连通区域，实现版面分割；根据各版面元素的外轮廓特征，简单区分文本区域与非文本区域；利用藏文文本识别器识别文本区域，再用可扩展标记语言记录版面信息，实现版面描述。在藏文中小学教材文档和铅印版藏文文档图像上的实验表明，该方法能够取得较好的版面分析结果。

关键词 图像处理；藏文文档图像；版面分割；版面描述；自适应游程平滑

中图分类号 TP391.1 **文献标志码** A **doi**：10.3788/LOP202158.1410006

Layout Segmentation and Description of Tibetan Document Images Based on Adaptive Run Length Smoothing Algorithm

Chen Yuanyuan[1], Wang Weilan[2*], Liu Huaming[3], Cai Zhengqi[1], Zhao Penghai[2]

[1] *College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, Gansu* 730030, *China*;

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

[3] *College of Computer and Information Engineering, Fuyang Normal University, Fuyang, Anhui* 236041, *China*

Abstract Layout segmentation is an important basic step in the process of document image analysis and recognition. In order to explore a suitable method for layout segmentation and description of Tibetan document images, a research method based on the adaptive run length smoothing algorithm is proposed. Firstly, according to the layout structure of Tibetan document images, K-means clustering analysis is used to get the run length threshold suitable for the layout, smooth the run length, find the connected component, and realize the layout segmentation. Then, according to the external contour characteristics of each layout element, the text area and non-text area are simply distinguished. Finally, the text area is recognized by a Tibetan text recognizer, and then the extensible markup language is used to record layout information and realize layout description. Experiments on Tibetan primary and secondary school teaching materials and stereotyped Tibetan document images show that this method can achieve good layout analysis results.

Key words image processing; Tibetan document image; layout segmentation; layout description; adaptive run length smoothing

OCIS codes 100.2000; 100.2960; 100.3008

收稿日期：2020-09-21；**修回日期**：2020-10-19；**录用日期**：2020-11-12

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)、甘肃省双一流学科建设项目-西北民族大学(11080304)、甘肃省高等学校创新基金项目(2020B-069)

通信作者：*wangweilan@xbmu.edu.cn

1 引 言

藏文作为中文的一大语言文字分支，是仅存的几种少数民族语言文字之一。藏文的历史悠久程度、文献丰富程度都仅次于汉文。藏文从左到右书写，字体分为“有头字”和“无头字”两大类，常用的是有头字，相当于汉文中的楷书。对藏文文档图像进行版面分析并以数字化方式存储版面信息，不仅可以提供一种高效查阅、检索的方式，也方便后续进行高效的版面复原工作。因此，藏文文档图像的版面分析是一件非常有意义的工作。

在过去的几十年中，文档图像版面分析工作在多文种、多字体上展开，国内外的研究者针对印刷或手写的文档提出了许多不同的版面分析方法。文档图像版面分析方法多依赖于所处理图像的版面特点，传统方法常利用角点信息、边缘信息、连通域信息等确定版面的各部分，随着深度学习的风靡，神经网络也被广泛用于文档图像的版面分析[1-4]。现有的版面分析研究及开发工作主要用来处理一些主流语言（如汉文、英文、法语等）的文档图像，只有少量针对少数民族语言文档图像特点的版面分析方法被提出[5-6]。针对藏文文档图像版面分析技术，也仅有少数基于古籍的相关研究被提出。其中，Ma 等[7]研究出一种应用于藏文历史文档图像分割和识别的框架，提出基于块投影的版面分割方法，将藏文文档图像分割成文本、线条和框架，利用基于图模型的文本行分割方法解决文本与边框之间的粘连问题。Liu 等[8]提出一种基于边界信息的藏文历史文献版面分析方法，采用中值滤波、高斯平滑、Sobel 边缘检测和边缘平滑、去除小区域、获取边界位置等一系列处理，根据边界和区域之间的位置关系，确定各个区域位置，例如文本区域、左注释、右注释等。最后以 XML 页面信息的格式保存文档图像。张西群等[9-11]提出一种基于连通分量分析和角点检测的历史藏文文档图像文本提取方法，利用关联成分把藏文历史古籍的文档区域划分为三类，将图像等分为网格，利用连通域分类信息和角点密度信息对网格进行滤波，计算垂直和水平网格投影，通过投影分析可以检测出文本区域的大致位置，通过校正近似文本区域的包围盒，准确地提取文本区域。Duan 等[12]给出一种基于块投影的历史藏文文档图像文本提取方法，将图像平均分块，并根据连通分量的类别和角点密度信息进行滤波，通过块投影分析找到近似的文本区域，并提取文本区域。上述基于传统方法的版面分析研究在藏文古籍文档图像上取得了较好的效果，而目前尚未有针对印刷版面的藏文文档图像版面分析的方法被提出。

本文针对这一现状，以中小学藏文教材文档图像为例，将藏文文档图像版面分析划分为 7 个阶段：预处理、自适应游程平滑算法（ARLSA）处理、连通域分析、目标连通域过滤、版面分割、版面元素分类以及版面描述，形成一个系统的藏文文档图像版面分析方法。

2 方 法

2.1 方法的流程

藏文文档一般由大量藏文文本行，少量图、符号以及装饰物等组成，如图 1 所示。文本行之间（垂直方向）不存在明显粘连，但元音与基字之间存在断开的情况，如图 1 右侧矩形框部分所示。字符与字符、字符与音节点（水平方向）可能存在粘连，如图 1 中左侧矩形框部分所示。根据藏文文档图像版面结构特点，提出版面分析方法的流程，如图 2 所示。首先对图像进行二值化、去噪等预处理，对二值图像进行 8 方向连通域分析，依据得到的各连通域的宽、高属性，利用 K-means 聚类分析确定游程（run-length）阈值，通过对二值图像进行游程平滑，使得单个文本行形成单独或数量较少的连通分量；然后进行连通域分析，生成连通域的矩形外接框，过滤过小的连通域（噪点）以及矩形外接框重叠的连通域，归属与基线分离的元音，以实现各版面元素的定位；再依据矩形边界框的宽、宽高比等属性，利用聚类分析方法确定阈值，对各版面元素进行文本、非文本的简单分类，并在原图像中标出，分割出所有矩形框；最后将文本类型的版面元素送入文本识别器，得到识别结果，利用可扩展标记语言将识别结果及各矩形边界框的位置信息进行整合，生成版面描述文件。图 2 中 C_1、C_2 满足 2.1 节给出的过滤及归属条件，C_3 满足 2.3 节中不等式组的条件。

2.2 自适应游程平滑算法

传统的游程平滑算法应用于文档图像分割和文本检测的前期处理[13-14]，该算法对同一扫描行（列）上的黑色像素点之间的距离进行检测，当两个相邻黑色像素点之间的空白游程长度小于阈值 T 时，将这两点之间的空白游程全部填黑[15]。算法示例如图 3 所示，其中每一个网格代表一个像素点，图 3(a)经过水平阈值 $T_{hor}=4$、垂直阈值 $T_{ver}=1$ 的游程平滑算法（RLSA）处理后，得到的结果为图 3(b)。

图 1　藏文文档图像

Fig. 1　Tibetan document image

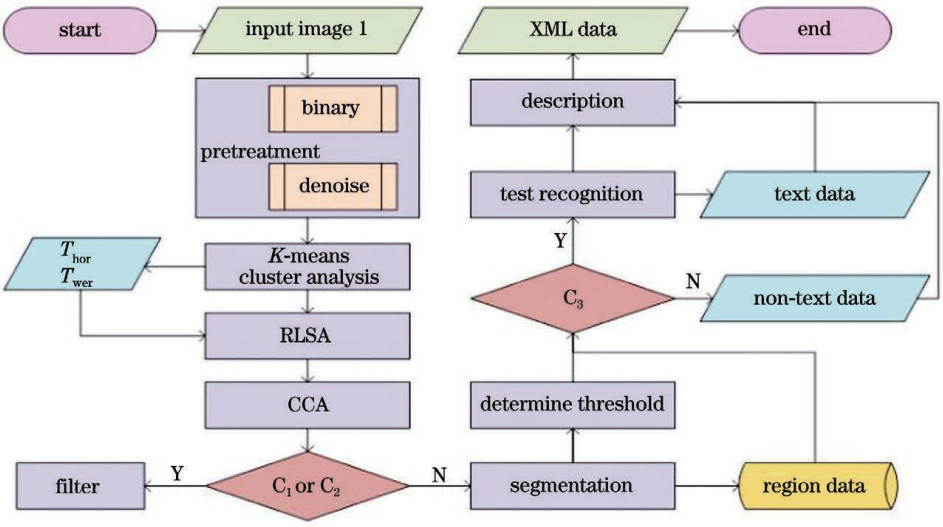


图 2　版面分析方法流程示意图

Fig. 2　Flow chart of layout analysis method

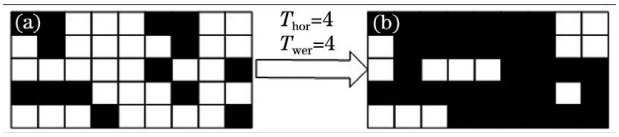


图 3　RLSA 示例图。(a)像素点未平滑；(b)像素点平滑后

Fig. 3　RLSA samples. (a) Pixels before smoothing; (b) pixels after smoothing

本文提出的自适应游程平滑算法是依据藏文文档图像版面结构进行连通域分析，自动确定适用于不同版面的游程阈值，对作为二值图像前景的白色像素点进行平滑，使之形成数目较少或者单个的连通域。藏文文档图像中的版面元素依据连通域的宽度可以大致划分为 4 类：1)音节点、小断裂、噪点；2)正文字符、不粘连元音；3)标题字符、小图标；4)大图标、线条。藏文文档图像中的版面元素依据连通域的高度也可以划分为 4 类：1)音节点、小断裂、噪点、线条；2)正文字符；3)标题字符、小图标；4)大图标。对连通域的宽度、高度分别进行 $K=4$ 的 K-means 聚类分析，得到的聚类中心数据表示如表 1 所示。

表 1　连通域的聚类中心数据表示

Table 1　Data representation of cluster centers in connected components

Cluster center	Center 1	Center 2	Center 3	Center 4
Width	w_1	w_2	w_3	w_4
Height	h_1	h_2	h_3	h_4

令阈值 $T_{\mathrm{hor}}=w_2$、$T_{\mathrm{ver}}=h_1$，将其送入 ARLSA。ARLSA 的输入是一张前景为白色、背景为黑色的二值文档图像和阈值 T，输出是算法处理结果。

ARLSA 示例如图 4 所示，其中图 4(a)为藏文文档图像版块的二值化图，图 4(b)是图 4(a)经过 ARLSA 处理之后的结果。

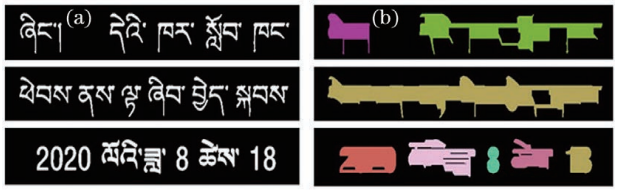


图 4　藏文文档图像文本行 ARLSA 处理。(a) 二值图；(b) ARLSA 处理结果图

Fig. 4　ARLSA process of text lines in Tibetan document images. (a) Binary figures; (b) ARLSA processing results

2.3　目标连通区域形成

经过 ARLSA 处理的图像存在由噪点形成的连通域，如图 5(a)所示，且仍然存在少数元音与基线分离的情况，如图 6(a)所示，这时需要对连通域进行进一步的过滤及归属。

第一步：过滤噪点以及归属外接矩形框重叠的连通域。通过开源图像处理库(OpenCV)的相关方法，检测所有连通域的外轮廓，然后通过计算取得每个轮廓的垂直边界最小矩形，根据其左上角和右下角坐标标出矩形框，如图 5(b)所示。可以看到，图 5(b)中存在重叠矩形框以及噪点连通域的矩形框，因此需要对矩形框进行过滤。对于整张文档图像，过滤的基本思想是：1)过滤噪点，通过对宽、高、宽高比、面积设定阈值，过滤不满足阈值条件的矩形框，根据 2.1 节中的聚类结果，确定宽、高、宽高比、面积的阈值分别为 $T_{\mathrm{w}}=w_1$、$T_{\mathrm{h}}=h_1$、$T_{\mathrm{ar}}=w_1/h_1$、$T_{\mathrm{area}}=w_1\times h_1$，当矩形框的 4 个属性值都小于阈值时，过滤此矩形框；2)归属重叠矩形框，若矩形一的几何中心包含于矩形二，则将矩形一归属于矩形二，当且仅当两矩形的几何中心重合时，将面积小的矩形归属给面积大的矩形。根据以上思想，令 $B=\{b_0,b_1,\cdots,b_n\}$ 表示所有的矩形边界框的集合，G_{C} 表示矩形边界框的几何中心。当 b_i 满足 $G_{\mathrm{C},b_i}\subseteq b_j$，且当 $G_{\mathrm{C},b_i}=G_{\mathrm{C},b_j}$，若 $A_{b_i}<A_{b_j}$(A 表示面积)时，将 b_i 归属于 b_j，其中 $i\neq j$ 且 $i,j\in(0,\cdots,n)$。过滤得到的图像如图 5(c)所示。

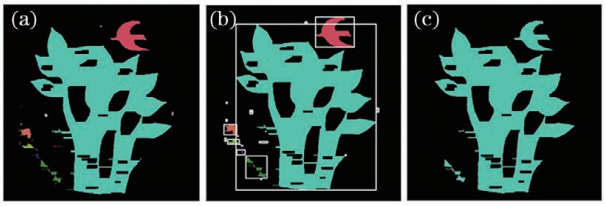


图 5　连通域过滤结果。(a) ARLSA 处理结果；(b)连通域的矩形外接框；(c)过滤结果图

Fig. 5　Filtering results in connected domains. (a) ARLSA processing result; (b) rectangular outer box for connected domains; (c) filtering result

第二步：归属分离元音。通过计算得到每个连通域的质心，如图 6(b)中白点部分。如图 6(c)所示，比较质心之间的垂直距离，当距离小于阈值 T_{v} 时，将质心所属面积小的连通域归属给质心所属面积大的连通域，通过此方法使得与基线分离的元音归属到该基线所在文本行，归属结果见图 6(d)。T_{v} 可表示为

$$T_{\mathrm{v}}=\frac{w_2+w_3}{4}。\tag{1}$$

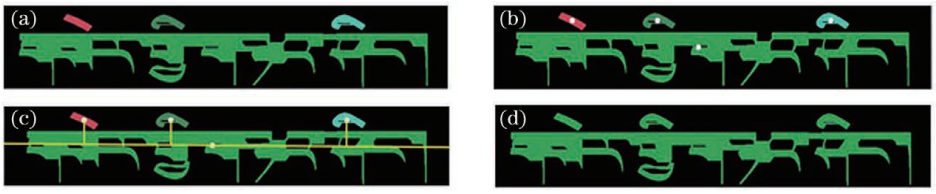


图 6　与基线分离的元音归属。(a)文本行的 ARLSA 处理结果；(b)连通域的质心；(c)质心间的垂直距离；(d)过滤结果图

Fig. 6　Vowel attribution separated from baseline. (a) ARLSA processing result of text line; (b) centroids of connected components; (c) vertical distance between centroids; (d) filtering result

2.4　版面分割及分类

本文采用 OpenCV 中分割图像矩阵的方法在原图上将检测出的各矩形框分割出来，以藏文中小学教材文档以及铅印版藏文文档等图像为例，分割

目标连通域。由于版面特点，分割得到的大于 90% 的版面元素都是文本行，非文本行的数量很少。因此，将分割得到的区域划分为两类：文本(text)和非文本(non-text)，通过 K-means 聚类分析发现，设定合适的高度 H 和长宽比 A_R 的阈值可以有效区分一类文档图像的文本区域和非文本区域，如图 7 所示。图 7(a)是随机挑选的 165 张文本区域分割块以及 110 张非文本区域分割块的(H，A_R)数据的分布图，其中沿边框分布的数据为非文本数据，左下角密集分布的数据为文本数据。图 7(b)是对所选数据进行 3 分类分析的 K-means 聚类结果图，将沿边框分布的数据作为非文本，左下角密集分布的数据作为文本数据。根据图 7 确定连通域宽度、宽高比的最大值及最小值：$H_{min}=40$，$H_{max}=120$，$A_{R,min}=0.4$，$A_{R,max}=35$。其中满足不等式

$$\begin{cases} H_{min} \leqslant H \leqslant H_{max} \\ A_{R,min} \leqslant A_R \leqslant A_{R,max} \end{cases}, \tag{2}$$

即为文本，否则为非文本。

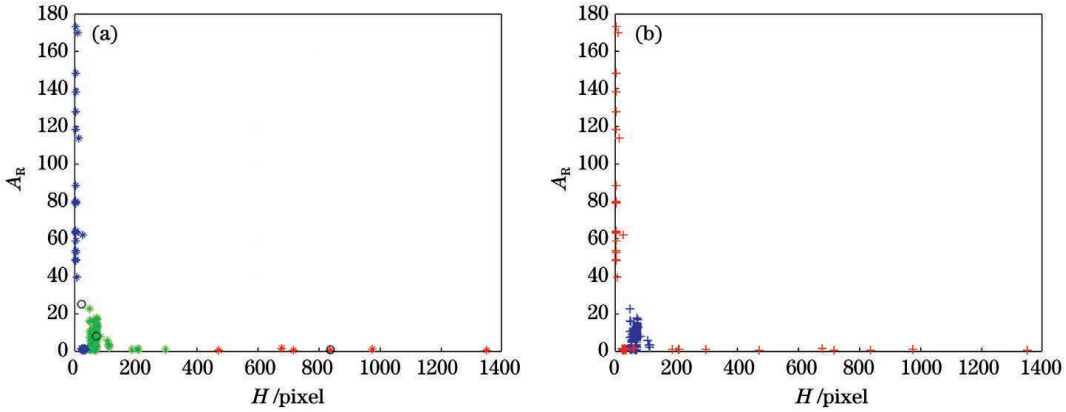


图 7　随机分割块样本数据聚类分析图。(a)随机样本数据分布；(b) $K=3$ 聚类

Fig. 7　Cluster analysis graphs of random segmentation block sample data. (a) Random sample data distribution; (b) $K=3$ cluster

2.5　版面描述

版面描述可以说是一种高级解释性的页面描述语言，体现在本文的版面分析过程中即通过藏文文档图像的版面元素，对其文本、图形以及其他区域的外观以及内容进行描述，并以一定的存储形式进行保存。对于上文中分割以及分类得到的版面元素，采用 XML 对其进行描述，描述的方法是：1)对于文本元素，将其送入文本识别器，得到识别结果，将其与位置信息等一同描述；2)对于非文本元素，将其统一存储为图片格式，描述其存储路径、位置信息。由此将版面描述数据分为 4 类：页面信息(PageInfo)、创建信息(MetaData)、文本区域信息(TextRegionInfo)、非文本区域信息(NonTextRegionInfo)。其中页面信息包括图像名称、图像的宽和高；创建信息包括操作者、操作时间以及最后一次修改时间；文本区域信息包括位置信息、文本信息以及编码；非文本区域信息包括位置信息和区域图像路径信息。版面描述数据对应的结构如图 8 所示。

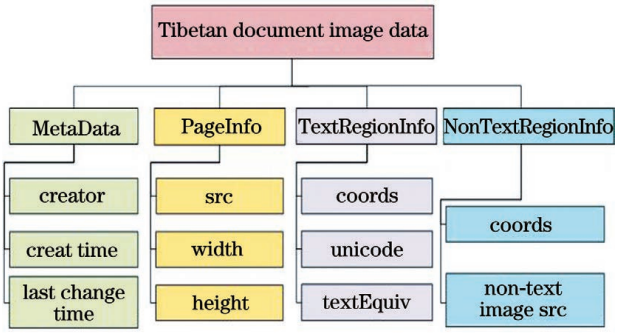


图 8　版面数据结构示意图

Fig. 8　Structural diagram of layout data

要识别文本类型的版面元素，首先要对文本区域进行字切分，然后将切分出的藏文字符送入字符识别器，得到识别结果。采用深度神经网络模型 CovNet 对藏文字符进行识别[16]。

字切分要解决的问题主要是与基字分离元音的归属和字粘连的切分，如图 9(a)所示。采取的相对应方法是：

1) 对文本区域进行连通域分析，求取连通域的平均高度，此时的平均高度受音节点与分离元音的影响，相对于真实的字符高度来说偏小，因此先过滤掉小于平均高度的连通域(音节点、分离的元音)，再求一次平均高度，作为最终的字符平均高度。将平均高度的 50% 的值作为阈值 T_{ver}，对文本区域图像做垂直方向 ARLSA 处理，归属与基字分离的元音。

2) 计算除去音节点、单垂线连通域的平均宽

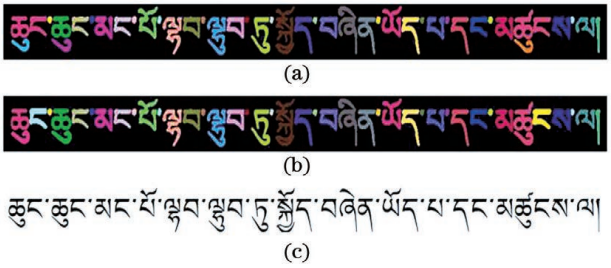


图 9　字切分与识别。(a) 元音与基字分离、字粘连示例；(b) 切分处理结果；(c) 识别结果

Fig. 9　Word segmentation and recognition. (a) Separation of vowels and base words, and word adhesion; (b) segmentation result; (c) recognition result

度，将宽度大于 1.5 倍平均宽度的连通域判定为粘连区域，利用平均宽度切分粘连区域。

切分结果如图 9(b)所示。将切分出的藏文字符送入字符识别器，得到的识别结果如图 9(c)所示。

3　实验结果及分析

3.1　实验结果

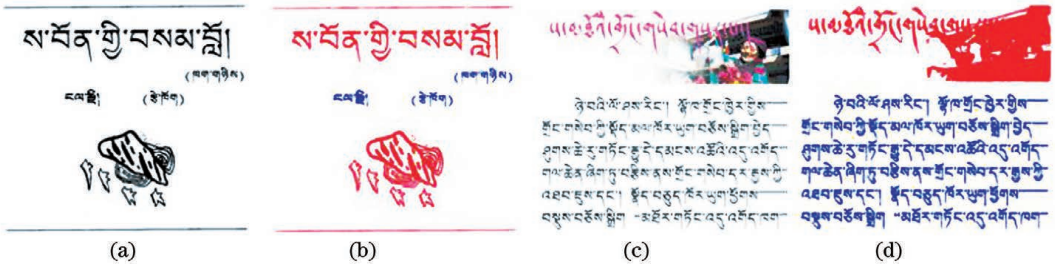
本文以藏文中小学教材文档和铅印版藏文文档图像作为实验样本，挑选文档版面 306 张，用所提方法对其进行了版面分割，分割得到版面元素 13188 张，其中文本元素 12869 张、非文本元素 319 张。实验结果错分 47 张，分割正确率为 99.64%。其中部分有图版面的实验结果如图 10 所示。图 10(a)是原图，图 10(b)是图 10(a)经过 ARLSA 处理的版面分割结果，图 10(c)是版面元素分类结果，图 10(d)是版面描述。可以看到，用该方法进行版面分析可以得到较好的结果。

图 10　版面分析结果。(a)原图；(b)目标连通区域；(c)版面元素分类结果；(d)版面描述

Fig. 10　Layout analysis results. (a) Original image; (b) target connected region; (c) classification result of layout elements; (d) layout description

3.2 实验结果分析

通过实验发现，本文所提出的版面分析方法对于藏文文档图像中的文本区域以及非文本区域具有较好的定位、分割和识别效果，所有的目标连通区域都能被正确分割出来。然而本文方法还不能很好地检测与分割页面中粒度差异较大的文本块、艺术字形式的文本块或与图像叠加的文字块。图 11(a)、(c)为原文档图像，分割与分类结果如图 11(b)、(d)所示。



(a)　(b)　(c)　(d)

图 11 类别错分图像。(a)(c)原图；(b)(d) 错误分类结果

Fig. 11 Wrong classification images. (a)(c) Original images; (b)(d) wrong classification results

分析其原因，主要有以下两个方面：

1)文档图像样本中含非文本数量较少，且非文本包含图、符号、装饰物、页码等，种类较多，容易错分；在使用阈值进行简单分类的情况下，只能区分出大部分在阈值区间内的版面元素，例如存在单个字符或者少量几个字符形成一个文本区域的情况，容易与符号、页码混淆，还有一些文本区域由于字体大小与整篇文档区别较大，单靠阈值还不能很好地区分。

2)对于藏文文档图像中存在背景图的标题文本行，通过连通域分析的方法还不能检测出图中的文本，因而被分割并错分为图像。

4 结　论

提出了一种基于 ARLSA 的藏文文档图像版面分析方法，包括藏文文档图像预处理，利用 ARLSA 实现版面元素定位，进行连通域分析、过滤非目标连通域，然后进行版面分割，利用 K-means 聚类分析分割后的版面元素得到合适的阈值并实现分类，最后用 XML 存储版面信息。通过在藏文中小学教材文档图像以及铅印版藏文文档图像上的实验表明，所提方法能够准确定位以及分割目标连通区域，对目标区域的分类也有较好的效果，而且版面描述有利于文档图像的查阅与检索。在后续的研究中，将继续扩充样本数量，优化版面元素分类方法，对非文本区域进行进一步的划分和分类，以达到更好的版面分析效果。

参 考 文 献

[1] Li Y X, Zou Y J, Ma J W. DeepLayout: a semantic segmentation approach to page layout analysis[M]// Huang D S, Gromiha M M, Han K, et al. Intelligent computing methodologies. Lecture notes in computer science. Cham: Springer, 2018, 10956: 266-277.

[2] Oliveira S A, Seguin B, Kaplan F. dhSegment: a generic deep-learning approach for document segmentation[C]//2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), August 5-8, 2018, Niagara Falls, NY, USA. New York: IEEE Press, 2018: 7-12.

[3] Chen K, Seuret M, Hennebert J, et al. Convolutional neural networks for page segmentation of historical document images[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), November 9-15, 2017, Kyoto, Japan. New York: IEEE Press, 2017: 965-970.

[4] Zulfiqar A, Ul-Hasan A, Shafait F. Logical layout analysis using deep learning[C]//2019 Digital Image Computing: Techniques and Applications (DICTA), December 2-4, 2019, Perth, WA, Australia. New York: IEEE Press, 2019: 1-5.

[5] Wang Y W. Research and implementation of layout analysis and post-processing for Mongolian document images[D]. Hohhot: Inner Mongolia University, 2017.
王艳文. 蒙古文文档图像版面分析及识别后处理的研究与实现[D]. 呼和浩特: 内蒙古大学, 2017.

[6] Chen X, He J J, Li H J, et al. Manchu document layout analysis based on mask R-CNN[J]. Journal of Dalian Minzu University, 2019, 21(3): 240-245.
陈璇, 贺建军, 李厚杰, 等. 基于 Mask R-CNN 的满文文档版面分析[J]. 大连民族大学学报, 2019, 21(3): 240-245.

[7] Ma L L, Long C J, Duan L J, et al. Segmentation and recognition for historical Tibetan document images[J]. IEEE Access, 2020, 8: 52641-52651.

[8] Liu H M, Bi X H, Wang W L. Layout analysis of

historical Tibetan documents[C]//2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), May 25-28, 2019, Chengdu, China. New York: IEEE Press, 2019: 74-78.

[9] 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.

[10] Zhang X Q. Research on layout segmentation method for historical Tibetan documents[D]. Beijing: Beijing University of Technology, 2018.
张西群. 面向藏文历史文献的版面分割方法研究[D]. 北京: 北京工业大学, 2018.

[11] Zhang X Q, Duan L J, Ma L L, et al. Text extraction for historical Tibetan document images based on connected component analysis and corner point detection[M]//Yang J F, Hu Q H, Cheng M M, et al. Communications in computer and information science. Communications in computer and information science. Singapore: Springer, 2017, 772: 545-555.

[12] 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.

[13] Rais M, Goussies N A, Mejail M. Using adaptive run length smoothing algorithm for accurate text localization in images[M]//Martin C S, Kim S W. Progress in pattern recognition, image analysis, computer vision, and applications. Lecture notes in computer science. Heidelberg: Springer, 2011, 7042: 149-156.

[14] Papamarkos N, Tzortzakis J, Gatos B. Determination of run-length smoothing values for document segmentation[C]//Proceedings of Third International Conference on Electronics, Circuits, and Systems, October 16, 1996, Rhodes, Greece. New York: IEEE Press, 1996: 684-687.

[15] Zhang L, Zhu Y, Wu G W. English document segmentation based on run-length smearing algorithm[J]. Acta Electronica Sinica, 1999, 27(7): 102-104.
张利, 朱颖, 吴国威. 基于游程平滑算法的英文版面分割[J]. 电子学报, 1999, 27(7): 102-104.

[16] Chen Y. Design and implementation of printed Tibetan recognition software on android platform[D]. Lanzhou: Northwest University for Nationalities, 2020.
陈洋. 安卓平台上印刷体藏文识别软件的设计与实现[D]. 兰州: 西北民族大学, 2020.