

中图法分类号: TP391. ; 4 文献标识码: A 文章编号: 1006-8961(2024)06-1685-29

论文引用格式: Wang W L, Hu J S, Wei H X, Ubul Kurban, Shao W Y, Bi X J, He J J, Li Z J, Ding K, Jin L W and Gao L C. 2024. Survey on text analysis and recognition for multiethnic scripts. Journal of Image and Graphics, 29(06):1685-1713(王维兰, 胡金水, 魏宏喜, 库尔班·吾布力, 邵文苑, 毕晓君, 贺建军, 李振江, 丁凯, 金连文, 高良才. 2024. 少数民族文字文本分析与识别的研究进展. 中国图象图形学报, 29(06): 1685-1713)[DOI:10.11834/jig.240015]

少数民族文字文本分析与识别的研究进展

王维兰^{1*}, 胡金水², 魏宏喜³, 库尔班·吾布力⁴, 邵文苑⁵, 毕晓君⁶,
贺建军⁷, 李振江⁸, 丁凯⁹, 金连文¹⁰, 高良才¹¹

1. 西北民族大学数学与计算机科学学院, 兰州 730030; 2. 科大讯飞研究院, 合肥 230001;
3. 内蒙古大学计算机学院, 呼和浩特 010021; 4. 新疆大学计算机科学与技术学院, 乌鲁木齐 830046;
5. 上海大学社会学院, 上海 200000; 6. 中央民族大学信息工程学院, 北京 100081; 7. 大连民族大学信息与
通信工程学院, 大连 116605; 8. 甘肃政法大学网络空间安全学院, 兰州 730000; 9. 上海合合信息科技股份有限公司,
上海 200000; 10. 华南理工大学电子与信息学院, 广州 510641; 11. 北京大学王选计算机研究所, 北京 100871

摘要: 对于少数民族古籍的保护与传承, 国家予以高度重视, 并强调了对这些不可再生文化资源透彻数字化的重要性。随着文档图像分析与识别技术的不断进步, 对少数民族文字的文本分析与识别研究受到广泛关注, 并取得显著成就, 成为人工智能应用研究的一个热点领域。然而, 由于少数民族文字种类繁多、应用场景多样及数据集的稀缺性等问题, 这一研究领域仍面临诸多挑战。本文旨在总结先前的工作, 并为未来的研究提供支持, 重点讨论了印刷体文本、联机手写、古籍文档及场景文字识别等任务, 概述了国内外在少数民族文种识别领域的发展和最新成果。首先阐明了少数民族文字文本分析与识别的重要性及其价值, 介绍了特定少数民族文字及其古籍文档的特征。然后, 回顾了这一领域的发展历史和现状, 分析并总结了传统方法的代表性成果及其应用; 详细讨论了研究重点向深度神经网络模型和深度学习方法的全面转移, 这一转变使得各文种的识别性能得到了显著提升。最后, 基于相关分析, 本文指出了在不同文种文档分析与识别中存在的精度和泛化能力等方面的不足, 以及与汉文文本分析与识别的差异; 面对少数民族文字文本识别领域的主要困难与挑战, 展望了未来的研究趋势和技术发展目标。

关键词: 少数民族文字; 文档分析与识别; 印刷体文本识别; 手写识别; 古籍文档识别; 场景文字识别

Survey on text analysis and recognition for multiethnic scripts

Wang Weilan^{1*}, Hu Jinshui², Wei Hongxi³, Ubul Kurban⁴, Shao Wenyuan⁵, Bi Xiaojun⁶,
He Jianjun⁷, Li zhenjiang⁸, Ding Kai⁹, Jin Lianwen¹⁰, Gao Liangcai¹¹

1. School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, China;
2. iFLYTEK Research Co., Ltd., Hefei 230001, China; 3. College of Computer Science-College of Software, Inner Mongolia University,
Hohhot 010021, China; 4. School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China;
5. School of Sociology and Political Science, Shanghai University, Shanghai 200000, China; 6. School of Information Engineering,
Minzu University of China, Beijing 100081, China; 7. College of Information and Communication Engineering,
Dalian Minzu University, Dalian 116605, China; 8. School of Cyberspace Security, Gansu University of Political Science and Law,

收稿日期: 2024-01-06; 修回日期: 2024-01-26; 预印本日期: 2024-02-02

* 通信作者: 王维兰 wangweilan@xbmu.edu.cn

基金项目: 国家自然科学基金项目(62166036, 61772430, 62266044, 62236011); 内蒙古自治区科技计划项目(2019GG281)

Supported by: National Natural Science Foundation of China (62166036, 61772430, 62266044, 62236011); Inner Mongolia Autonomous Region Science and Technology Program (2019GG281)

Lanzhou 730000, China; 9. INTSIG Information Co., Ltd., Shanghai 200000, China;

10. School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China;

11. Wangxuan Computer Institute, Peking University, Beijing 100871, China

Abstract: China's ethnic scripts differ in their structure types, creation periods, and regions of usage and scope. The historical documents and various literary materials written, recorded, and printed in ethnic scripts are even more voluminous, which leave an invaluable wealth for exploring the civilization and development history of different ethnic groups. Compared with mainstream languages, the study of ethnic minority scripts often faces low-resource conditions. In recent years, the protection and inheritance of the intangible cultural heritage of ethnic minorities have attracted increased attention from the country, which has great importance and application value for the protection of irreparable diverse cultural resources. By applying traditional image processing, pattern recognition, and machine learning methods, certain results have been achieved in text recognition and document recognition in Mongolian, Tibetan, Uyghur, Kazakh, Korean, and other major languages. Compared with mainstream languages such as English and Chinese, the research on the character recognition of minority languages, the analysis of document images, and the development of application systems is relatively lagging behind. Since the 21st century, the research and application of ethnic script text analysis and recognition have received extensive attention and made remarkable progress due to the continuous development and application of technologies in the field of document image analysis and recognition. They have become the research hotspots in the field of document analysis and recognition and artificial intelligence. However, a large number of problems still need to be solved in the field of minority script text and recognition research due to the large number of minority scripts, the wide range of application scenarios, and the scarcity of datasets. This study reviews the development history and recent progress in this field at home and abroad to better summarize previous works and provide support for the subsequent research. It focuses on four subtasks: printed text recognition, handwriting recognition, historical document recognition, and scene text recognition of several minority texts. It mainly includes Tibetan, Mongolian, Uighur, Yi, Manchu, and Dongba. These studies are mainly related to the following areas. 1) In the document image preprocessing stage, the system performs a series of operations on the input image, such as binarization, noise removal, skew correction, and image enhancement. The goal of preprocessing is to improve the accuracy of subsequent analysis and recognition. 2) Layout analysis, such as layout segmentation, text line segmentation, and character segmentation, helps understand the organizational structure of documents and extract useful information. 3) Text recognition is one of the core tasks of document image analysis, which identifies the text in a document through various technical approaches. This task may involve traditional methods such as text recognition based on single character classifiers, or it may include end-to-end text line recognition in deep learning methods. 4) Dataset construction involves constructing various datasets for training and evaluating algorithms, such as document image binarization datasets, layout analysis datasets, text line datasets, and character datasets. By contrast, analysis and recognition of historical documents are difficult due to the complexities of rough, degraded, and damaged historical book papers, which result in severe background noise in the document image layout, sticky text strokes, unclear handwriting, and damage. At present, a practical recognition system for historical documents is lacking. First, the importance and value of minority script text analysis and recognition are explained, and some minority script texts, especially historical documents, and their characteristics are introduced. Then, the history of the development of the field and the current state of the research are reviewed, and the representative results of the research of the traditional methods and the progress of the research of the deep learning methods are analyzed and summarized. Current research objects are expanding in depth and breadth, with processing methods comprehensively shifting to deep neural network models and deep learning methods. The recognition performance is also greatly improved, and the application scenarios are constantly expanding. One of the studies realizes effective modeling under low resources. It further proposes a unified multilingual joint modeling technology to identify multiple languages through one model, greatly reduce the overhead of hardware resources, and significantly improve the image and text recognition effect and generalization in multilingual scenarios. At present, it can recognize images and texts in 18 key languages or ethnic languages, including English, French, German, Japanese, Russian, Korean, Arabic, Uyghur, Kazakh, and Inner Mongolian. Based on relevant analyses, obvious deficiencies are observed in recognition accuracy and

generalization ability, and differences with Chinese text recognition of ethnic script text recognition are found. The characteristics of the characters and documents of each language are completely different from those of Chinese characters and Chinese documents. For example, in the development of the Yi language, variant characters are particularly abundant due to various factors, and “one-to-many, many-to-one” characters and interpretations are the norm. The arbitrariness and diversity of historical Yi handwriting have brought great challenges to the recognition of historical Yi script. Moreover, the Tibetan script uses arabesque, the shape of the letters is complex, the black plum script is intertwined with each other, some strokes even span several characters before and after, and the connection between the letters is also relatively unique. Thus, the multi-style Tibetan recognition with high complexity and difficulty needs to be solved to achieve true multi-font text recognition. Finally, the main difficulties and challenges faced in the field of minority text recognition are discussed, and the future research trends and technical development goals are prospected. For example, research and application system development are conducted in combination with the characteristics of different languages, layout formats, and varying application scenarios. A certain gap still exists between the recognition of most ethnic languages and the development of Chinese recognition, especially in applications related to education, security, and people’s livelihood. This gap can be addressed by actively expanding new application directions. Opportunities for expansion are abundant, such as migrating large language models to ethnic minority scripts and text recognition and developing a unified multilingual joint modeling and application system.

Key words: multiethnic scripts; document analysis and recognition; print recognition; handwriting recognition; historical document recognition; scene image text recognition

0 引言

中国是统一的多民族国家,民族文字和民族语言是丰富多彩的中华文化载体。虽然各民族文字在结构类型、创制年代、使用地区和范围等方面都有所不同,但民族文字撰写、记录、印制的历史文献和各种图书资料却浩如烟海,为探讨不同民族的文明史、发展史留下了极其宝贵的财富。在中国的30多种少数民族语言文字中,有些语言文字不仅有较长的历史且至今广泛应用,如彝文,被称为史前文明留下的文字,而蒙古文、藏文、维吾尔文、哈萨克文、柯尔克孜文、朝鲜文等还是跨境民族语文。如藏语除中国的西藏、青海、四川、甘肃、云南等省区外,印度、尼泊尔、不丹等国家也在使用。蒙古语主要分布于中国的内蒙古和邻近的蒙古国。满语主要分布于黑龙江、吉林、辽宁,满文源于传统蒙古文,可追溯到更早的回鹘文。维吾尔语和哈萨克语同属突厥语族,主要分布于中国的新疆,以及俄罗斯、哈萨克斯坦、吉尔吉斯斯坦、乌兹别克斯坦等国,维吾尔语和哈萨克语都以“阿拉伯字母”作为书写文字。朝鲜语除在中国的延边朝鲜族聚居区使用外,也是韩国与朝鲜两国的官方语言,朝鲜文又称“韩文”。彝语与南亚国家语言有亲属关系,在中国的四川、云南、贵州、广西、

广东等省区均有分布,其书写符号“方块壮字”近似汉字,都是单音节、方块形,且造字原理有重合之处。东巴文是目前世界上唯一存在并仍被使用的成体系的象形文字。主流语言文字的文档识别方法甚至表格识别研究与应用受到广泛关注并取得巨大进展(刘成林等,2023;高良才等,2022)。少数民族文字文本分析与识别,不仅对我国民族语言文字的抢救、保护、传承和信息技术发展,而且对边疆民族地区的稳定具有极其重要的意义,此外对于构建中国周边跨境文档图像分析研究的国际文化交流和学术话语权,提升学术实力、学术地位和影响力,以及交往、交流、交融等都有特殊的价值。

各个民族文字、文本的图像识别,因各自的特殊性与主流语言文字如汉文和英文都不相同。进行系统的民族语文印刷体识别研究和应用平台开发是2000年左右,清华大学丁晓青教授团队在对汉文、韩文、日文等文档识别研究和应用开发“HMS-89多字体汉字识别系统”、“多字体多字号印刷体汉字识别系统”、“高性能东方文字文档智能全信息数字化系统”等成果的基础上,利用其技术优势与相关大学又联合进行了藏文、蒙文、维文、哈萨克文、柯尔克孜文、朝鲜文等少数民族文字文本的识别研究与应用系统开发,包括“统一平台上少数民族文字(藏维哈柯朝)文档识别系统”、“统一平台英汉混排民族文字

(蒙藏维哈柯朝)文档识别系统”等(丁晓青等, 2017)。其中藏文识别的单位是字丁、字符集大小为592个,在乌金体类的6种字体样本测试集的平均识别率为99.8%,藏汉英混排文本的测试集上平均识别率为98.7%,藏文文本测试集上识别率为97.3%;维吾尔文字符类别数147、维汉英混排文本的测试集上平均识别率为96%;蒙文识别子系统,字符集大小是160个、单字识别率为99.4%。韩文字符集大小为7 238个;维哈柯文的字符集大小分别为147、156和158;阿拉伯文字符集大小为163个等。通过文通科技公司推出商业化的软件“民族语言文字识别系统”,得到广泛的应用并取得了良好社会效益。

该领域的研究也伴随着相关技术的发展,特别是深度神经网络时代的到来而不断发展。2023年6月,西北民族大学研究团队开发的网络版藏文识别软件正式上线(<https://ocr.sys303.com/>),解决了大规模文档、多种图像格式、PDF(portable document format)、手机拍照等图像文档的识别,在个人计算机端实现批量文本、手机端单页文本的藏文文档识别,获得用户好评、具有一定的社会影响力。在彝文文本识别方面,合合信息公司联合上海大学、华南理工大学共同推进,并于2023年发布业内首个形义相关的古彝文基础编码数据库。中国科大讯飞研究院科研团队针对多语种识别面临的训练数据稀缺和不同语种独立部署带来的硬件资源开销过大的问题,提出了场景风格可迁移的图文生成技术、感受野可控的对比式预训练技术,实现低资源下的有效建模,并进一步提出编码统一的多语种联合建模技术,通过一个模型识别多个语种,大幅降低硬件资源开销,显著提高多语种文档图像识别效果与泛化性。

本文围绕少数民族文档图像处理、版面分析与识别等几大任务,从传统方法到深度学习方法,主要就藏、蒙、维、彝、满、东巴等多个文种印刷体文本识别、手写识别、古籍文本识别和场景文本识别等,综述该领域国内国外研究的发展历史和最新进展,对未来的趋势和技术发展目标进行展望。由于该领域涉及文种多,在回顾相关进展时只按作者的理解选择引用和介绍部分代表性工作,做不到将所有少数民族文字文本、所有成果进行全面综述,肯定会遗漏部分研究结果或者技术成就。

1 主要少数民族文字文档及其特点

1.1 主要少数民族文字和字符识别单位

中国有多种民族文字,其结构类型主要分为表音文字和表意文字两种。表音文字分为音素文字和音节文字。音素文字以字母呈现语言中的元音和辅音,又称“拼音文字”,其中采取线形拼写的有蒙古文、维文和满文等,采取块形拼写的有朝鲜文。表意文字在我国常见的有图符文字和音意文字,图符文字如东巴文,用一连串的符号记录语言中的实词,但虚词要靠读者的联想来补充,是图画走向文字的过渡形态。传统彝文与汉字的造字结构具有可比性,属音意文字。以下对本文涉及的几种少数民族文字作进一步介绍。

藏文由30个辅音字母和4个元音,左右拼写、上下叠加形成音节字,音节的结构如图1(a)所示。音节中的每个纵向单位称为字丁,1个音节至少1个字丁、最多4个字丁。部分上加和下加字母存在变形。藏文字符集包括:《ISO/IEC 10646-1: Tibetan character collection[S]》、《信息技术 藏文编码字符集(扩充集A)》、《信息技术 藏文编码字符集(扩充集B)》,共7 000多个字丁,其中现代藏文500多个,其余均为梵音藏文,即梵文的藏文转写。梵音藏文没有像音节一样的拼写规则,可以前后任意搭配组合,上下叠加时没有层数的限制。如图1(b)所示的梵音藏文,从左到右纵向叠加层数分别是7到2。古籍中还大量出现扩充集A、扩充集B之外的梵音藏文,由于其叠加的灵活性和不确定性给藏文古籍文档识别带来较大的挑战,就像开放集文字识别(杨春等, 2013)的问题一样,藏文也涉及开放集的文字识别问题。以字丁为识别单位的藏文古籍识别中,类别多、相似字丁多以及不同字体风格差异大等给识别带来了难度。

蒙古文由8个元音字母和27个辅音字母构成。Unicode编码的0x1820-0x1842区域分配给了蒙古文字母。虽然蒙古文字母(名义字符)只有35个,但字母根据其在单词中的位置和上下文将呈现不同形状(显现字符)。蒙古文的书写方式也非常独特,以词为单位竖写,即同一个单词中所有字母按出现的先后顺序从上到下依次拼接起来,因此它的书序是从上到下,列序是从左到右。蒙古语属于黏着语,通过

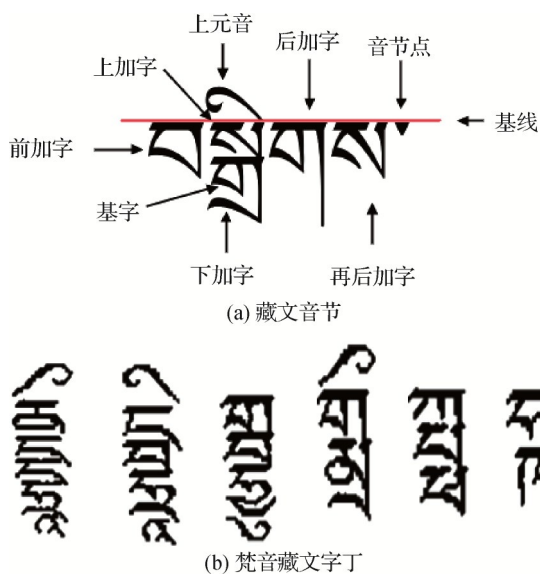


图1 藏文音节结构和梵音藏文示例

Fig. 1 Tibetan syllabic structure and Sanskrit examples
(a) Tibetan syllabic; (b) Sanskrit

在词根或词干后缀接一个或者多个词缀的方式产生新词,这使得蒙古语的词汇量相当庞大。根据统计,蒙古文词根(干)有6万左右,词缀有数百种,按照其构词规则,蒙古文词汇量可达百万以上。

维吾尔文由32个字母构成,其中8个元音字母,24个辅音字母。几个辅音字母和一个元音字母合在一起构成一个音节,音节合在一起构成词。词与词之间是分写的。维吾尔文是从右向左横写,字母没有大写小写之分,每个字母有不同的书写形式和规则。字母的不同形式可以单独出现,可以出现在词的不同位置,即词首、词中或者词尾。维吾尔文字母是由笔画和各种点、符号组成的,多一个点和符号或少一个点和符号,都会影响字母的形体,会造成无法拼读或改变词义的结果。

彝文分为规范彝文、通用彝文和传统彝文3种。规范彝文有四川规范彝文和云南规范彝文,四川规范彝文也就是“凉山规范彝文”,又称“新彝文”,是1980年国务院批准的一套由819个基本字结合次高调符号组成的音节文字。2010年云南开始试行的规范彝文方案包含表意字2 258个、表音字350个。通用彝文是四省区共同挑选出来的5 598个彝文字,以2016年出版的《通用彝文字典(第二版)》为代表,致力于向“书同文”迈进。传统彝文又称“古彝文”,以2004年发布的《滇川黔桂彝文字集》为代表,汇集四省区60余名古彝文专家根据传统经籍中文字的

实际使用情况提供了总计87 046(88 613)个未经整理的“原生态”字符。除819个凉山规范彝文字符外,其余彝文字符还没有获得预留的Unicode区段,在2006年提交的ISO规范化方案N3288中也还没有明晰的符义对应关系、跨地区形近字的区分,这是彝文古籍识别最大的瓶颈,也是为什么彝文识别要从最根本的原生态彝文着手的主要原因。

无论何种形态的彝文都遵循一字一音一义的形式,彝文的识别以字符为单位。在彝文的发展过程中,由于种种因素,导致异体字、变体字特别丰富,字符和释义“一对多、多对一”是常态。古彝文手写体的随意性、多样性等,都给古彝文的识别带来了极大的挑战。传统彝文此前没有被系统性地数字化编码,古彝文公开数据集稀少,标注困难,所以从最初语料的收集开始,就需要做大量的前置工作。再者,古彝文异体字繁多,每个字的异体写法少则两三种,多则几十种,且字体间风格差异大。因此,传统彝文的识别需要采取与字符量较少的表音文字不同的策略,建立一个专门的数据库,通过基础编号将不同样式归纳,才能“破解”古彝文“一对多”的关系,解决文字查询问题。

1.2 主要少数民族文字文本数据集

任何一项与文档识别相关的工作,都离不开数据集。汉文文本识别已经有各种各样的数据集,包括票据数据集及识别(乔梁等,2023)。针对识别的不同任务、数据及规模等,将目前部分藏文识别相关数据集汇总如表1所示、蒙文识别相关数据集如表2所示、维吾尔文手写识别的数据集情况如表3所示。

1.3 民族文字文本数据

少数民族文字文本分析与识别,也因为应用对象的不同而有区别,有些文字,如藏文、维文、蒙古文等不仅需要现代文档的分析与识别,还需要古籍文档分析与识别;而有些文字仅仅是古籍保护,如满文。图2所示分别为藏文、蒙古文、满文、东巴文和彝文等。其中,图2(a)左侧是藏文印刷体,两行是内容完全一样但分别是乌梅体和乌金体文本句子;同样古籍中也有乌梅体。不同字体的显形差异很大。图2(e)从左至右是彝文古籍文字、图文混排和螺旋状的文本格式。

表1 藏文识别相关数据集情况
Table 1 Dataset of relevant Tibetan recognition

数据集	面向任务	规模	构建方式
清华TH-OCR多字体印刷藏文(丁晓青等,2017)	印刷体藏文识别(白体、黑体、通用体、圆体、长体、竹体)	592类字符	不同字号的真实印刷字体
MRG-OHTC(Ma等,2011)	联机手写字符识别	910类字符;130书写人	人工标注
THCDB(Huang等,2014)	脱机手写字符识别	234类字符;314书写人	人工标注
NMU-OLTHWDB(Wang等,2019)	联机手写字符识别	7 240类字符;5 000样本每类170类部件	人工标注(部件)基于规则生成
藏文本刻版(仁青东主等,2021)	藏文古籍识别	600页文档;724 957个字符	人工标注,基于规则生成
THCU(Tibetan historical character of uchen)(Li等,2022b)	藏文古籍字符识别	238类字符;121 214样本 7 240类字符;5 000样本每类	人工标注 人工标注(部件)数据增强;基于规则生成
HUTD(historical uchen Tibetan documents)(Zhang等,2022a)	藏文古籍字符识别	610类字符;700样本每类	人工标注;数据增强
TxTAB(Zhi等,2021)	藏文古籍文本检测	202页文档	人工标注
NMU-THDTL(Northwest Minzu University-Tibetan historical document text line)(Li等,2019a)	藏文古籍文本行切分	210页文档	人工标注;数据增强
BiTCSD(郝玉胜等,2022)	藏汉双语场景文字检测识别	5 550幅真实图像; 87 680幅合成图像	人工标注;数据合成

表2 蒙文识别相关数据集
Table 2 Dataset of relevant Mongolian recognition

数据集名称	面向任务	规模	构建方式
大规模多字体蒙文单词数据集(Zhang等,2017)	印刷体蒙文识别	20万个词汇;4种字体(白体、报体、标题体、哈旺体);训练集:80万个样本(每种字体20万个样本);测试集:2万个样本(每种字体5 000个样本)	人工标注
MRG-OHMW(Ma等,2016)	联机手写蒙文识别	946个词汇、282 954个样本,300人书写 训练集:235 826个样本;测试集:47 128个样本	人工标注
大规模蒙文联机手写单词数据集(李敏,2019)	联机手写蒙文识别	132 406个词汇、287 800个样本,600人书写 训练集:257 800个样本;测试集:30 000个样本	人工标注
MHW(范道尔吉等,2018)	脱机手写蒙文识别	训练集:5 000个词汇,10万个样本(每个词汇20个样本),200人书写;测试集I:1 000个词汇,5 000个样本(每个词汇5个样本)200人书写(与训练集相同);测试集II:939个词汇,14 085个样本(每个词汇15个样本),50人书写(与训练集不同)	人工标注
蒙文古籍数据集(Wei和Gao,2019)	木刻印刷蒙文古籍识别	100页文档、20 948个单词图像、1 423个词汇	人工标注

2 文本分析与识别研究状况

2.1 印刷体文本识别

2.1.1 印刷体藏文文本识别

2000年,清华大学丁晓青教授团队联合西北民

族大学开展了藏文印刷体识别的研究,其成果已经进行商业转化,即文通科技公司“民族语言文字识别系统”中的藏文识别板块。印刷体藏文多字体包括6种字体:白体、黑体、通用体、圆体、长体和竹体,均属于藏文的乌金体不同风格。王华和丁晓青(2004a)对多字体的印刷藏文字符的归一化问题研

表 3 维吾尔文手写识别的分类器和数据集
Table 3 Classifier and dataset for Uyghur alphabets handwriting recognition

分类器/训练算法	面向任务	数据集	构建方式
分段线性排序器(Simayi 等,2019)	联机手写字符识别	532 页文档	人工标注
具有梯度下降的SVM(Xu 和 Du,2017)	脱机手写字符识别	133 页文档	人工标注
用异或计算进行模板匹配(Simayi 等,2019)	脱机手写字符识别	324 页文档	人工标注(部件)数据增强 基于规则生成
RBF 核 SVM,超越了修正二次判别函数(modified quadratic discriminant function, MQDF),判别学习二次判别函数(discriminative learning quadratic discriminant function, DLQDF),学习尺度量化分类器(learning vector quantization, LVQ)(Simayi 等,2013)	联机手写字符识别	32 类字符、1 200 样本每类	人工标注(部件)基于规则生成
加权朴素贝叶斯(许亚美 等,2013)	联机手写字符识别	32 类字符、3 412 样本每类	人工标注、数据增强
KNN(姜文 等,2013)	联机手写字符识别	544 页文档	人工标注
基于最近邻、最小生成树算法(minimum spanning tree, MST),动态时间规整(dynamic Time Warping, DTW),聚类的分类器组合(Li 等,2016)	脱机手写字符识别	32 类字符、1 253 样本每类	人工标注、数据增强
HMM 和统计语言模型(柳玲玲和赵晖,2012)	联机手写字符识别	210 页文档	人工标注(部件)、数据增强 基于规则生成

究,综合考虑了字符的重心和外接矩形框,对字符的位置和大小进行归一化;提出的藏文字符识别方案(王华和丁晓青,2004b),使用方向线素特征和最近邻分类器,其特点在于分类时做了级联划分,并对样本间的欧氏距离进行了修正。王维兰和丁晓青等人(2003)对藏文印刷文本 6 种字体风格进行了识别结果的分析。

其他的印刷体藏文识别工作与上述工作框架类似,改进之处主要在于字符的特征、分类器的设计等。欧珠等人(2009)使用的主要特征依然是方向线素特征、粗网格,还有诸如笔划密度特征和轮廓特征等,使用主成分分析(principal component analysis, PCA)方法降维,神经网络、支持向量机(support vector machines, SVM)等被应用到印刷体藏文识别。李永忠等人(2012)将分形矩和粗网格技术应用在藏文字符特征提取,一定程度上克服了藏文字丁多而造成的误识率高的缺点。

刘芳和尼玛扎西(2020)将藏文字符拆分为笔划,基于笔划提取特征并进行藏文识别研究。利用投影法和连通域对文档进行分割,并基于音节点切分后识别,能够提高识别性能,开发了相应的试用软件(才让当知 等,2022;王悦凝 等,2021)。针对乌金体和乌梅体多字体藏文文档开展了检测和识别研究

(侯闫 等,2023;高定国 等,2023),使用 DBNet + ResNet + CRNN 的框架实现了检测 +识别的端到端过程。

2011 年,美国弗吉尼亚大学开发了藏文光学字符识别(optical character recognition, OCR)系统,融合使用了 196 维的梯度方向特征、90 维的 Zernike 矩特征以及 14 维的穿线特征。采用了两步识别的策略,先使用基于径向基函数(radial basis function, RBF)核的 SVM 对所有字符进行识别,后使用维特比算法寻找隐马尔可夫(hidden Markov model, HMM)中最可能的隐藏状态序列,据此来调整 SVM 的识别结果(Rowinski 和 Keutzer,2016),并在加州伯克利分校持续开发。2014 年,成为藏传佛教资源中心(buddhist digital resource center, BDRC)的合作伙伴,为 BDRC 收藏的大量印刷体藏文文献进行了数字化存储与检索等。

总体而言,印刷体藏文识别随着深度学习技术的突破以及人工智能领域的加速崛起得到了广泛的应用研究。虽然有利用最新技术开发的藏文识别软件,相对于主流文种,可用于训练和优化藏文文字识别系统的数据有限。同时藏文乌梅体类更是笔画相互交织、有的笔画甚至前后跨越好几个字符,字母之间的连接方式也比较独特,需要解决较高复杂性和

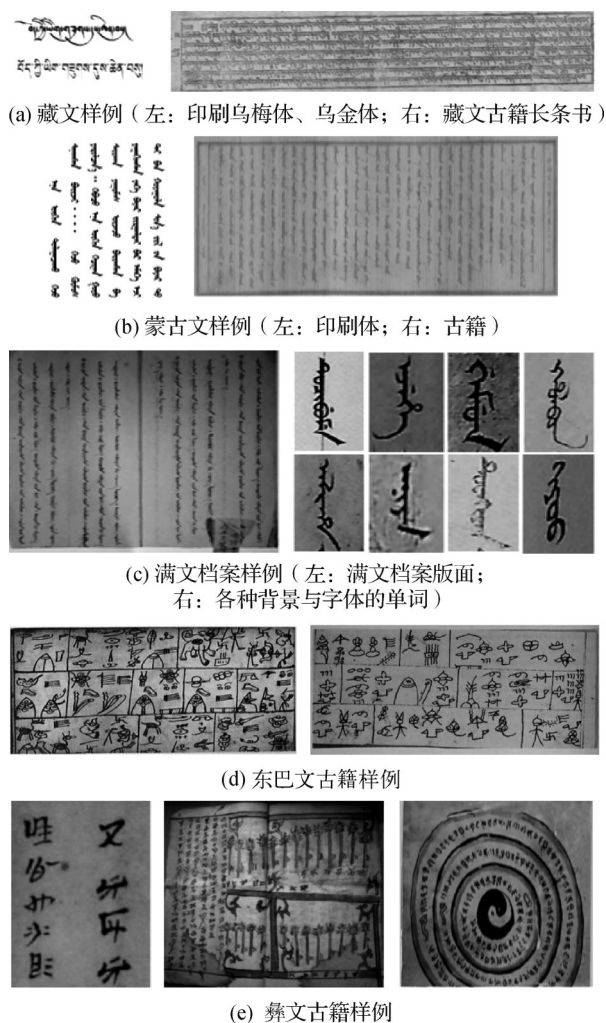


图2 部分民族文字文本样例

Fig. 2 Examples of texts in different ethnic scripts

((a) Tibetan; (b) Mongolian; (c) Manchu;
(d) Dongba pictographs; (e) Yi scripts)

难度的多风格藏文识别,实现真正的多字体印刷文本识别。

2.1.2 印刷体蒙文文本识别

研究者针对蒙古文印刷文档,考虑其版面元素类型多样、结构复杂、多种文字混排等情况,先后开展了版面分析、文种和文字识别等。魏宏喜和高光来(2007a)最先提出一种基于连通域的蒙古文文档图像版面分析方法,通过提取文档图像中所有连通域,并根据连通域的大小进行聚类,可获得文字连通域和非文字连通域,再根据文字连通域的上下位置关系归其所属列,最后按照从左到右的列序形成文字段落。该方法对于版面结构相对简单的印刷文档取得了较好效果,但对版面结构更为复杂的蒙古文期刊、报纸等印刷文档的效果有待提升。此外开展

基于连通域的自底向上和基于段落间距离的自顶向下相结合的图像版面分析方法,能够较好地处理期刊、报纸等版面结构更为复杂的印刷文档。

针对蒙汉英多种文字混排问题,进行了粗分类与细分类相结合的两阶段文种识别方法。在粗分类阶段,依据文字图像的宽度、高度等信息进行分类,将所有文字图像粗略地分为蒙古文、汉文和英文类。在细分类阶段,对汉文、英文中的文字图像分别使用卷积神经网络进一步确定其所属文种类别。该文种识别方法的识别准确率能达到98%以上。针对蒙古文印刷文档倾斜检测与矫正的研究,魏宏喜和高光来(2007b)提出了一种基于最小二乘法的倾斜检测方法。通过提取文档图像中文字连通域的质心,使用最小二乘法对同一列文字连通域的质心进行直线拟合,从而可以得到文字质心所在直线方程,该直线的斜率反映文档图像的倾斜情况。可以对整幅图像进行相应角度的旋转,实现倾斜矫正。

基于切分的识别方法需要先对输入的单词图像进行切分得到字形基元(字元),再对得到的字元进行特征提取,进而识别相应字元。内蒙古大学首次针对印刷体蒙古文字开展了基于切分的识别方法研究(李振宏和高光来,2003;李振宏等,2003;魏宏喜和高光来,2006),通过字元切分、特征选择、特征提取及特征匹配等完成识别,在常用的印刷字体上达到了90%以上的单词识别准确率。

使用深度神经网络提取待识别文字图像的特征,通常使用卷积神经网络进行特征提取,再使用循环神经网络进行序列建模,最后利用连接时序分类对输出的序列进行处理,得到最终整词识别结果。Zhang等人(2017)最先提出了一种序列到序列的整词识别方法,将文字识别看成是一个序列到序列的映射问题,即输入序列是一个蒙古文单词图像的片段序列,输出序列是蒙古文单词对应的文本标注(字符序列)。设计了一个带注意力机制的编—解码器作为识别模型,还构建了一个含4种常用蒙古文字体(白体、报体、标题体、哈旺体)的数据集。该方法在测试集上的单词识别准确率为89.6%。

Wei等人(2021b)将多任务学习的思想引入到该任务中,可以同时完成整词识别和字元切分两项任务。在提升单词识别性能的同时,还能提高字元分割的准确性。在Zhang等人(2017)构建的数据集上进行测试,多任务学习方法在4种字体上的识别

准确率均优于序列到序列的方法,单词平均识别准确率在99%以上。同时,集外词也能被较好地识别。

另外,Cui等人(2022)还围绕不规则印刷体蒙古文字识别,提出了一种基于三元注意力调节网络的整词识别模型。先通过空间变换网络对不规则文字图像进行矫正,再用所提出的整词识别模型进行识别。该方法能够处理文字图像存在的倾斜、扭曲等不规则问题。

2.1.3 印刷体维哈柯文文本识别

2000年以来,在印刷体文字、文本识别方面,新疆大学和清华大学开展合作研究。完成维吾尔、哈萨克、柯尔克孜三种基于阿拉伯字符集的民族文字单字识别、维吾尔(混排汉英)文本识别的研究和系统开发,使得一般文档达到自动识别的实用化程度。

研究者进一步针对维文文档图像和维文文字特点,提出模板特征、环特征、连通区域特征、笔划特征、笔划密度特征等多特征融合,并使用反向传播神经网络(back propagation neural network, BPNN, 简称BP)分类器,取得较高的文字识别率。对于复杂背景图像中维吾尔文字切分与识别,通过对字符边缘特征和笔画分布规律的研究,得到有效的切分方法;结合维吾尔文的特有规则及维吾尔文周边结构模式提取特征,通过最近距离输出识别结果(依米提等, 2014)。分别建立基于隐马尔可夫模型HMM的印刷体维吾尔文和阿拉伯文识别系统,其中提取分布密度特征和局部方向特征,利用HTK(hidden Markov model toolkit)工具建立维吾尔文和阿拉伯文统计语言模型,改进了识别系统的性能(喀迪尔等, 2015)。通过分析模型状态在图像层面的聚类意义,提出基于观测合理聚类的模型结构优化方法、结构与参数相结合的字符模型优化策略,将其应用于多字体维吾尔文文本行的无切分识别系统(姜志威等, 2015)。

在深度学习方面,针对维吾尔文属于粘连性文字难于切分和识别的问题,建立了维吾尔文图像识别的数据集,在卷积循环神经网络(convolutional recurrent neural networks, CRNN)的基础上提出了TRBGA(TPS + ReNet + BiLSTM + GRU + ATTN)模型,对比实验证明基于注意力的模型在提升识别性能方面要优于基于连接时序分类器模型(connectionist temporal classification, CTC)(汤敬等, 2021);选用CRNN + CTC的文本识别方案进行维吾尔文识

别,对CTC的输出部分进行改进,设计了符合维吾尔文书写顺序从右向左的字符排序规则,从而正确训练模型;使用循环卷积神经网络对印刷体维文整词进行识别,在循环层中使用双向门控循环单元(bi-directional gated recurrent unit, BiGRU)对双向短时记忆(bi-directional long short term memory, BiLSTM)进行替换,有效解决了单词图像不等长和维文字符切分的问题;使用空间变换网络对文本图像进行自适应的矫正,再将矫正结果输入到改进的CRNN识别网络以提高文本识别精度(熊黎剑等, 2021)。

2.1.4 印刷体多文种文本识别

多文种识别面临的主要问题是训练数据稀缺和不同语种独立部署带来的硬件资源开销过大。为解决这些问题,科大讯飞科研团队研发出编码统一的多语种联合建模技术。该技术通过一个模型就能识别多个语种,在大幅降低硬件资源开销的同时显著提升了多语种图文识别的效果和泛化性。为了实现从单一语种识别到多语种识别的技术突破,同时解决小语种数据资源不足导致的效果和鲁棒性差的问题,提出一种基于端到端的多语种图文识别统一建模方法(Wu等, 2023)。其基本思想是:首先将多个语种的文本图像进行编码,提取它们的图像特征。在编码阶段,提出方向特征强化模块,实现了多方向的特征提取。然后,利用带有注意力机制的解码模块得到识别结果。在解码模块中,解码单元为单字节,能避免多语种混合建模时的大词表问题。解码过程中,该方法融合了语种类别信息,相似语种的类别属性一致,使得模型能够学到相似语种间的共性信息,避免统一建模造成的语种串扰。

目前已经具备日、韩、俄、维文、传统蒙古文等18个文种或民族语言的图文识别能力,整体识别情况如表4所示。

2.2 手写民族文字识别

2.2.1 手写藏文识别

在OCR研究领域,联机字符识别难度要低于脱机字符的识别,当然,联机公式识别(杨晨等, 2023)除外。藏文联机字符识别的研究工作晚于脱机字符识别,这主要是硬件的发展和人们需求的变化所致。随着2007年第一代苹果手机的发布,触屏式的智能手机颠覆了人们的生活,广泛的应用需求刺激联机手写输入技术的发展。2000年左右,学者对联机手

表4 端到端多语种图文识别统一建模方案在
18个重点语种或民族语言上的结果
Table 4 Results in 18 key languages or ethnic languages
based on end-to-end unified multilingual modeling method

语种	识别率/%
英语	96.30
法语	94.20
德语	92.98
日语	96.64
俄语	95.66
韩语	96.18
西班牙语	93.49
葡萄牙语	93.23
意大利语	93.04
泰语	96.48
越南语	91.90
印地语	92.03
阿拉伯语	92.85
马来语	97.04
印尼语	95.87
哈萨克语	90.35
维吾尔语	88.68
传统蒙古语	75.48

写藏文识别开展了研究,对藏文字符结构进行划分,提取各划分结果块的4方向、6方向、8方向的笔迹特征,以及不同分块投影特征和几何特征,将这些特征进行融和,从而构成藏文字符的原始特征集。使用2DPCA(two-dimensional PCA)方法对原始特征进行降维(Wang等,2010)。在此基础上,基于MQDF(minimum quadratic distance function)分类器实现了对手写字符的分类,在上述识别过程中还加入了基于音节的联想输入功能,并在安卓平台上开发了联机手写识别输入法。此外,还对藏文大字符集进行了相关的研究,生成联机手写字符集样本库NMU-OLTHWDB(northwest minzu university-online Tibetan handwriting database)(王维兰等,2017),该库包含7 240类藏文字符,每个字符有5 000套样本。并开展了联机藏文相似字丁识别的相关研究(Cai和Wang,2018)。

中国科学院软件研究所的马龙龙团队构建了一个藏文联机手写数据集MRG-OHTC(multi-tech

research group-online handwritten Tibetan character)(Ma等,2011),提出了一个三级级联分类的方法,并对相似字的识别进行了研究,开发了联机手写藏文识别系统(Ma等,2013)。根据藏文字符上下叠加的特点,提出了基于部件分割的识别方法,分割后使用条件随机场对过分割的块进行重组,选择出合适的序列。对分割得到的部件进行识别,组合得到字符。如何对字符中部件的位置进行半自动的标注,以及如何利用置信度网络自动地提取部件特征,将识别的内容由字符扩展到了音节等,进一步提出了基于统计方法的无约束联机手写字符的识别方案(Ma和Wu等,2016)。

科大讯飞公司也开展了联机手写藏文识别的研究,为东噶藏文输入法提供了手写识别的技术支持。

由于缺乏需求的驱动,脱机手写藏文字符的研究较少。实际上藏文古籍文档分析与识别,也属于脱机手写文档识别的范畴,在此将通常脱机手写识别的内容予以总结。研究者开发了一个脱机手写藏文字符的数据集THCDB(Tibetan handwritten character database),在该数据集的基础上,对脱机手写字符的识别进行了深入研究,提出了小波变化和梯度方向的特征(Huang等,2014)。基于小波变换、矩的特征、基于稀疏表示、核主成分分析的特征,以及基于投影的符特征提取方法,使用MQDF、基于K-近邻和稀疏表示的级联分类等(Huang和Da,2014;蔡晓娟和黄鹤鸣,2016)。

2.2.2 手写蒙文识别

针对联机手写蒙古文字识别任务, Ma等人(2016)发布了一个联机手写蒙古文单词数据集MRG-OHMW(multi-tech research group, online handwritten Mongolian word)。他们将联机手写样例转为单词图像,并采用LeNet-5卷积神经网络作为识别模型,Top-1识别准确率为91.2%,但该模型仅能对946个单词进行识别;Liu等人(2016)又提出一种基于MWRCNN(Mongolian word region convolutional neural networks)和位置图相结合的整词识别方法,在MRG-OHMW数据集上进行测试,Top-1识别准确率提升至93.24%;Liu等人(2017)又提出基于循环神经网络的整词识别方法,先利用滑动窗口获取联机手写单词的序列特征,再送入双向LSTM(long short-term memory)与CTC相结合的识别模型。在MRG-OHMW数据集的一个子集上进行测试,Top-1识别

准确率为90.35%。虽然识别性能没有进一步提升,但是该方法能够处理集外词。

Zhang等人(2017)提出的序列到序列蒙古文整词识别方法,设计并实现了基于序列到序列的联机手写蒙古文整词识别模型。从现有整词识别模型的性能来看,联机手写识别任务仍有很大提升空间。

针对脱机手写蒙古文字识别任务,范道尔吉等人(2018)发布了一个公开的数据集MHW(Mongolian offline handwritten database),训练集与测试集I是由同一批200位书写人完成,而测试集II是由另外50位书写人完成。

Fan和Gao(2016)将语音识别相关方法成功移植到脱机手写蒙古文字识别任务中,提出了深度神经网络与隐马尔可夫模型相结合的整词识别模型,将脱机手写蒙古文单词图像沿书写方向分割成若干子词序列,通过隐马尔可夫模型描述这些子词序列的生成过程,并采用深度神经网络描述子词序列的概率分布。在MHW数据集的测试集I和测试集II上,单词的识别准确率分别为82.16%和73.16%。为进一步提升识别性能,Fan等人(2019)又提出了一种多维LSTM与CTC相结合的整词识别模型,并通过基于子词的语言模型处理集外词。在MHW数据集的测试集I和测试集II上,单词的识别准确率分别提升至85.22%和76.78%。

Wei等人(2019)则将带有注意力机制的序列到序列模型应用于脱机手写识别任务,提出了一个端到端脱机手写整词识别模型。在MHW数据集的测试集I和测试集II上,单词的识别准确率进一步提升至87.68%和81.12%。此外,针对样本不足的问题,采用传统的空间变换方法实现了数据增广,将MHW数据集的样本数量扩充至原来的3倍,并使测试集I和测试集II上的单词识别准确率分别提升至90.68%和84.16%。Zhang等人(2021)又提出了一种蒙古文手写单词的局部增广方法,通过移动笔划的端点和笔划外的控制点,使用Bezier曲线重构笔划,对笔划产生局部形变,实现数据增广、扩充样本的多样性。但该数据增广方法并未在MHW数据集上测试相关识别模型的性能。

考虑到蒙古文词汇量巨大的特点,人工收集一个包含全部词汇(或者常用词汇)的数据集代价高、效率低。为此,Sun和Wei(2022)提出了基于生成对抗网络的蒙古文脱机手写样例生成模型。经过充分训练之

后,该模型可根据给定的任意蒙古文单词(文本)和随机选取的书写人信息,生成与所选取书写人风格一致的该蒙古文单词对应的手写样例(图像),从而可以获取新词汇的手写样例。该手写样例生成方法能够为收集包含全部词汇的脱机手写数据集奠定基础,有利于脱机手写蒙古文字识别技术的进一步发展。

2.2.3 维哈柯手写文字识别

1)维吾尔手写文字识别

维吾尔手写文字在日常交流、电子商务和金融等方面均有着广泛的应用。在无切分的维吾尔文手写体文本行识别中,隐马尔可夫模型是常用的模型之一(Simayi等,2016),传统方法主要采用“先切分,再识别”的思路。维吾尔文很多字符存在连写和合并现象,故字符之间的边界很难准确切分(Simayi等,2019)。基于深度学习的方法,例如卷积神经网络(convolutional neural network, CNN)和长短时记忆网络(LSTM),可以更好地提取特征和建立字符模型。近年来,随着深度学习技术的不断发展,研究者提出了一种基于深度卷积神经网络的无切分维吾尔文本行识别方法,获得了较好的识别效果。此外,采用深度融合网络解决多角度、多字体等问题,进一步提高了维吾尔手写体识别的准确率。然而,深度学习技术需要大量的训练数据和计算资源。由于维吾尔文手写体的样本量较小,因此如何设计有效的训练集和测试集,以及如何避免过拟合等问题成为研究的重点之一。

许多实验的样本数据并非一个来源,无法比较识别结果,也不方便分析不同分类器的性能。通过不同分类器的集成应用,在实验中产生了最高的分类或识别率(Xu和Xue,2019)。

对新型通信技术和智能应用的巨大需求,要求手写识别技术具有较高的精度和速度。尽管维吾尔族手写识别的研究取得了一定成绩,但为了提高维吾尔族手写识别的准确率,未来在数据样本建设、应用系统研发等方面仍需要做大量的工作。

哈文仅有少量手写体相关研究,而柯文手写识别研究仍是空白。

2)维哈柯签名识别

维吾尔文的手写签名验证起步较晚,仅有一些代表性的研究成果。维吾尔文作为新疆维吾尔自治区的通用文字在日常的交易合同和协议上使用频率、安全性要求较高。库尔班·吾布力团队对维吾尔

签名识别开展了系列研究,Ubul等人(2012)提出一种基于改进的网格特征的特征提取算法,使用KNN(K-nearest neighbour)分类器分类决策后得到的识别率为93.53%;库尔班·吾布力等人(2016)对维吾尔文签名图像分别提取密度特征、方向特征和局部中心点特征等,并使用相似性度量和SVM分类器,得到最高96%的识别率;Ubul等人(2018)对多文种(英文、汉文、维吾尔文)签名提取128维局部中心点特征和112维特征,用距离度量法进行分类,两两混合的识别率达到91.50%、95.75%和97.50%;莫龙飞等人(2020)使用灰度共生矩阵等融合提取维吾尔文、柯尔克孜文的特征,用BP分类器和SVM分类器进行签名的识别,单个文种识别率为92.75%,维吾尔文和柯尔克孜文混合的识别率为96.75%。Rexit等人(2022)提出一种基于局部最大出现特征(local maximum occurrence features, LOMO)的手写签名识别方法,在公共数据集CEDAR(center of excellence of document analysis and recognition)上的识别率达到了98.4%,同时提出融合局部最大出现特征和定向梯度特征直方图的离线签名识别方法,在自建数据集上对单语言签名的识别率为96.8%,对混合多语言签名数据集的识别率为96.41%。另外,Li等人(2021)使用模型微调的方法,在融合SE(squeeze and excitation)注意力机制的ResNet网络上,分别使用交叉熵损失函数和焦点损失函数,在自建的汉文、维吾尔文、哈萨克文和柯尔克孜文签名数据集进行识别,在多语种混合数据集上的最高识别准确率为99.44%,在维吾尔文和哈萨克文数据集上的最高识别准确率为99.75%,在柯尔克孜文数据集上的最高识别准确率为99.00%。

Li等人(2022a)所提出的基于卷积神经网络的小样本多语种离线手写签名识别的方法(offline handwritten signature-neural network, OHS-Net),与经典的VGG(Visual Geometry Group)网络相比,该方法具有参数少、模型小、泛化性能好、准确率高等优点,在多语种数据集中Top-1的最高平均识别准确率为96.17%、Top-5的最高平均识别准确率为98.73%。

3) 维哈柯签名鉴别

祖丽皮亚·艾尼等人(2018)提取灰度共生矩阵并计算灰度共生矩阵的能量、惯性矩、局部平稳性和熵等算子,对其进行加权融合,实验数据包括15个书写者的600个维吾尔文手写签名,使用BP识别的

正确率为91.78%,在GPDS数据库的实验结果为91.10%。张淑婧等人(2022)使用图像分块理论的多尺度局部二值模式MBLBP(multiscale block local binary pattern)特征和局部相位量化LPQ(local phase quantization),提取维吾尔签名的纹理特征,用SVM分类器和随机森林进行签名的鉴别,正确率为96.35%,对CEDAR数据库中的签名进行分类鉴别,正确率为96.73%。Xamxidin等人(2022)针对签名图像稀疏问题,提出了一种基于改进的IDN(inverse discriminator network)网络,利用通道注意力机制和增强的空间注意力模块提取签名验证的有效特征信息,对多文种签名(汉文、维吾尔文、哈萨克文、BHsig-B、BHsig-H)数据库进行实验,单文种总识别率(overall recognition rate, ORR)为98.38%,两个文种混合的ORR为96.33%,3个文种混合的鉴别拒识率(false rejection rate, FRR)为3.49%、误识率(false acceptance rate, FAR)为4.58%、ORR为95.97%。

2.3 古籍文档识别

2.3.1 藏文古籍文档识别

藏文古籍文档识别的研究最早见于上个世纪80年代,甚至早于印刷体藏文识别的研究。1991年日本东北工业大学的Kojima等人(1995)就开始了对木刻版藏文文档识别的研究,分析了识别的过程和技术手段。使用了一种基于逻辑推断的方法对文档中的字符进行切分,并对文档中字符的特征提取、相似字符的识别开展研究,实现了一个刻版藏文文档识别系统。由于当时的技术条件所限,该系统需要人工将每一个字符分开,其本质上是一个字符识别系统。

在这之后的十多年,关于藏文古籍文档识别的研究一直没有相关的报道。研究者的兴趣转移到藏文印刷体文本识别、联机手写藏文识别上。到2011年,美国加州大学伯克利分校的Hedayati等人(2011)提出了一个刻版藏文古籍识别的完整方案,详细阐述了藏文古籍识别过程中的难点,包括基线检测、行字切分和识别中面临的各种问题,并给出了解决方案,其数据源于7页木刻德格版《甘珠尔》,其中有6555个标记的字符168类。

从2017年起,西北民族大学、北京工业大学、西藏大学等开始了藏文古籍文档识别的研究工作,研究对象主要是木刻版的乌金体藏文文档。文档分析与识别流程大体上可以分为图像预处理、版面分析、行字切分以及字符识别等几个方面。

1)古籍文档图像预处理。不同的文档图像需要不同的方法,几何矫正是一个方面(秦海等,2023)。对于藏文古籍来说,由于其存在严重的退化,此外藏文古籍长条书,图像获取时拍摄光线不均、模糊和低光照环境获取等,通常的二值化难以奏效。针对这类图像,研究者改进局部阈值方法(Li等,2019b),在不同的颜色空间进行处理(Han等,2019),以及基于U-Net的深度学习方法(Zhao等,2021a)等进行处理;针对低光照的文档图像采取一系列的措施进行增强(Zhao等,2022;Zhao和Wang,2023)。针对乌金体藏文文本行存在显著基线的特点,使用相关的基线检测算法得到基线,并基于基线对扭曲的文档进行矫正(Li等,2019b);针对藏文文档中存在遮挡情况,在U-Net模型的基础上进行改进,通过提取侧面特征恢复文字的边缘(Liu等,2020)。

2)版面分析。版面分析旨在对文档图像中的不同区域进行检测与分类,以明确其归属性质(文本、图像、边框等),以便后续识别或版面复原等处理。针对藏文长条书版型,Liu等人(2019)通过连通域分析的方法获取其边框、内容以及注释内容;Wang等人(2022)对文档中的笔画宽度进行估计,利用该特征对页面元素进行一次初步分类,之后通过启发式的搜索方法获得精确的结果。除了上述基于图像处理的方法外,也有无监督的聚类方法(张西群等,2018)、深度神经网络细粒度的版面分割(Zhao等,2021b),以得到更加精细的文本行或行段。

3)行、字切分。在获取了藏文文本区域后,需要对该区域进一步划分以得到单个的字符,直观的做法就是先切出文本行,再对文本行中的每一个字符进行切分,基本的切分思路为:一是从文本行本身出发,二是从行之间的空白入手。在具体切分时,研究者所提出的方法包括基于文本行重心和轮廓跟踪的方法(Zhou等,2018)、基于连通区域分析和局部基线信息的方法(Hu等,2021)、基于藏文字符结构属性的方法(张策和王维兰,2021;Zhang等,2022b)、基于区域增长的方法(李金成等,2021)、基于路径规划的方法(Liu等,2022)等。其算法设计均从藏文文本本身的形制特征出发,包括藏文的基线、字符的重心、藏文音节符号、上下元音等,解决的主要问题包括文本行之间的粘连、行内字符间的粘连及笔划断裂等,算法的核心思路还是基于传统的图像处理过程完成行切分和字丁切分。

4)文字、文本识别。目前藏文识别大多以字丁为单位,藏文古籍中存在大量的梵音藏文,导致待识别的符号类别较多,且由于藏文构字特征,其相似字符也较多。有研究者的动机是将单个字丁拆分多个部件,对部件进行识别,然后再合并以得到最终的识别结果(Li等,2019a)。由于藏文构字部件类别较少,进行部件的识别工作极大地降低了待识别的符号种类。在特征选择上,也是从传统的图像特征(Dey等,2020)全面转移到深度学习上(Kaldan和Vijayalakshmi,2021;侯闫等,2023;Zhang等,2023)。通过设计各类不同的网络模型,自动化地提取特征并实现字符识别。此外就是行切分后的端到端识别。

2.3.2 蒙文古籍文档识别

对于蒙古文脱机手写文档和蒙古文古籍文档,通常仅包含文本且版面结构简单,因此这两类文档的版面分析任务只涉及文本的列分割和单词分割。针对木刻印刷蒙古文古籍文字,开展了基于切分的识别方法研究。Gao等人(2011)提出了蒙古文古籍单词图像的主干线定位和字元切分方法,确定了用于识别的蒙古文古籍字元集合,通过BP神经网络识别切分后的字元。在由5000余个图像质量较好的古籍单词构成的数据集上,测试了该方法的性能,单词识别准确率仅为71%。而通过单词显著轮廓点检测、基于Logistic回归模型的基线定位以及基于启发式规则的候选分段路径生成等3个步骤,改进字元切分方法,从欠分段与过分段片段的识别、字元分组以及基线信息合并等有效融合策略,使单词的识别准确率提升至81%(Su等,2015,2016)。

基于切分的识别方法先将待识别单词图像分割成对应字元,这种方法适用规范的文字。对于字形变化大或者书写自由度大、规范性差的手写文字,因字体或书写变形将导致很难正确分割出相应字元,造成切分效果不佳甚至无法切分的情况。故以蒙古文单词为识别单位,开展整词识别研究。

针对木刻印刷蒙古文古籍识别任务,Wei和Gao(2019)提出了一种基于卷积神经网络的整词识别方法,同时为了解决数据集上样本分布不平衡问题,使用了SMOTE(synthetic minority over-sampling technique)技术生成样本。在100页蒙古文《甘珠尔》扫描图像上,构建了包含20948个单词图像的(词汇量为1423)的蒙古文古籍数据集。在该数据集上,如果采用基于切分的识别方法,字元切分的准确率仅为

63.77%, 单词的识别准确率则更低。而所提出的基于卷积神经网络的整词识别方法, 未进行样例扩充的情况下单词的识别准确率为72.65%, 而进行样例扩充的情况下单词的识别准确率高达98.75%。但是, 这种基于卷积神经网络的整词识别方法无法处理集外词, 只能识别数据集中的词汇。

为解决集外词问题, Kang等人(2019)提出了一种带有注意力机制的序列到序列整词识别模型。该模型的编码器由一个深度神经网络与一个双向长短期记忆网络组成, 输入的图像帧序列通过深度神经网络进行特征提取, 之后送入双向长短期记忆网络获取帧序列之间的上下文关系, 并生成帧序列对应的特征向量序列。该模型的解码器由一个长短期记忆网络和softmax分类层组成, 它对编码器生成的特征向量序列进行解码, 生成相应的识别结果。该模型的注意力网络用于连接编码器和解码器, 使解码器在不同时刻关注到与目标字符最相关的一个或者多个图像帧, 从而提高解码器的准确性。在蒙文古籍数据集上进行四折交叉验证, 单词的识别准确率可达88.29%。而且该方法能够处理集外词。

由于蒙古文古籍单词样本的稀缺性, 在一定程度上限制了相关识别工作的进展。数据增广是一个很好的解决思路, Wei等人(2021a)提出的一种基于循环一致生成对抗网络(cycle-consistent generative adversarial networks, CycleGAN)的数据增广方法, 能够在训练数据不成对的情况下, 实现两个样本空间的相互转换。为了获取新样本, 可将蒙古文古籍单词图像送入训练后的CycleGAN, 先将其转为目标空间的样本, 然后转换为原始空间的样本, 以此获取同一个单词的新样本, 实现数据增广。在蒙文古籍数据集上进行四折交叉验证, 该数据增广方法能够将

单词的识别准确率由88.29%提升至92.49%。

2.3.3 彝文古籍文档识别

彝文古籍文档的识别分为样本的采集与扩增、图像预处理与字符检测、变/异体字的聚类与输出、校对系统的建立等4个步骤。

1) 样本的采集与扩增

由于通晓古彝文的专家人数本来就很少, 能够直接无任何参考标识的古籍上进行标注的更是寥寥无几, 无法满足智能识别对大量样本的需求。为此中央民族大学采用基于密度的无监督聚类算法进行自动类别标注。实验证明, 这种方法不仅提高了标注效率, 而且提升了标注的质量。研究者以《指路经》、《阿诗玛》等古籍为语料库, 使用聚类算法构造了包含100个彝文文字、10 530个样本的手写彝文公开数据集(贾晓栋, 2017)。

西南大学联合贵州工程应用技术学院, 用当代彝族师生手写样本模拟古籍抄写者风格的多样性, 采集了2 142个常用字的151 200个样本(陈善雄等, 2019), 又通过腐蚀、膨胀、仿射变换和旋转等形态学变换等, 实现了样本库的自动扩充, 得到3 786个常用字的297 414个样本(韩旭, 2020)。

2023年, 合合信息公司联合上海大学、华南理工大学的合作项目也采取了相似的策略, 完成1 135个常用字及其异体字的共计78 042样本的收集, 并利用光学原理对样本库进行扩增, 相应的行列页码已经公布(<https://github.com/wnfrdshao/yi-scan-corpus>)。与前者略有不同的是, 后者的样本中还包含了约20 000个译著截图样本。即通过汉字反查彝文的方式, 对包括《西南彝志》、《彝族源流》等已出版的彝汉文对照古籍译著中的某个用字情况进行截图保存, 确保样本库的多源性与真实性, 如图3所示。该数



图3 古彝文手写样本与译著彝文样本

Fig. 3 Handwritten samples and translated samples of Yi scripts

据集的另一特色是制作了1150行的版式数据,有助于提高后续文字定位的准确性。

2) 图像预处理与字符检测

构建彝文古籍数据库需要解决的首要问题是低质量的图像资料处理。传统的古彝文大多记录在纸书、岩石、布书、竹简等上,在潮湿的自然环境下难以完整保存,往往会变得模糊不清或残缺不全。在资料文档数字化的过程中,采用智能图像处理技术提升图像质量是关键的一个环节。古彝文有横、竖、螺旋形、放射状等多种排列方式,所以字符检测识别是第2个关键的环节。一些经书中有大量的插图,这种图文混排的复杂情况也是研究的重点。

西南大学与贵州工程学院的合作项目将彝文古籍分为白板经书和复杂背景下的彝文古籍两类。针对版面结构比较简单、噪声较少的彝文经书,采用连通域分析及回归式字切分的方法取得了较好的检测效果(韩旭,2020)。基于连通域的算法在印刷体四川规范彝文的识别中多有应用(刘赛等,2009)。针对版面结构复杂、排版缺乏规范、存在图文混排等情况的彝文古籍,提出了一种基于最大极值稳定区域(maximally stable extremal regions, MSER)和卷积神经网络(CNN)的字符检测方法。该方法对古籍中文本和非文本区域进行了有效的分离,并在单字符检测实验中取得了较高的准确率和召回率,能有效地解决复杂排版的古籍文献的字符检测问题(陈善雄等,2020)。

为了提高古籍彝文识别精度,还可采用曲面矫正等技术。采用基于深度学习的边界去除、内容迭代矫正曲面检测和矫正模型(Zhang和Luo等,2022),对形变文本图像进行像素级的矫正。该方法不仅可以处理各种复杂的曲面形状,而且能够保存文档的原始信息,不会出现传统方法中的模糊和扭曲问题。此外,还可以采用一些可公开访问的文档图像处理工具提升古籍文档图像的质量,例如TextIn平台的“智能图像”处理技术(https://www.textin.com/product/textin_image)应用在古彝文语料收集与字符集的整理过程中,可自动检测书籍影印图像中存在的问题或字符样本的高清整理,一键处理模糊、阴影、手指、屏幕纹和水迹等干扰因素,以便后续图像处理,缩短内容识别、编码的操作周期。

总之,彝文古籍因年代久远纸质泛黄、变脆,甚至出现残边;一些碑刻的彝文由于长期的侵蚀也变

得模糊。这就需要采用图像处理技术进行增强、滤波,提升其识别度。现有的图像处理技术虽然能很大程度上提高古籍的成像质量,但对于出现的残边甚至污损,则需要利用概率统计模型以及结合古彝文语义语境判断受损内容。

3) 变/异体字的聚类与输出

由于历史原因,古彝文没有标准的书写体例,即便采用同一种写法,不同的人撰写的文献在格式和字形字体上大相径庭,传统的统计特征提取方式无法准确获取变化较多的古彝文字符的特征表达。综合考虑文字的多样性以及实际使用场景,可以采用卷积神经网络处理单字符图像的识别。

在样本收集阶段,研究人员对字符进行了人工切分,目的是将复杂的文字图像分割成单字符样本,为模型提供训练样本。基于卷积神经网络的单字符彝文检测识别系统以细致的数据处理、轻量级网络设计和有效的损失函数选择,能够在多样性的文本和各种应用场景下实现高效的单字符识别。由于彝文古籍的抄写认真、一笔一划,因此,前人针对印刷/手写彝文特征提取的方法仍有一定参考价值,包括网格和交叉点特征(王嘉梅等,2008)、笔画密度、投影、笔画粗切割等特征(朱龙华和王嘉梅,2010)以及1024维周边方向贡献度(朱宗晓和吴显礼,2012)等。

不同于中文的简体繁体仅两种字形,历史上没有设立过专门的机构主管彝文规范,各地方或个人为了经文传承和保密的需要,刻意增减笔划,致使彝文变体字、异体字特别多。因此,需要通过提取同一字符不同异体字的特征进行比对,利用特征演化和序列匹配的模式,对其变化特征进行判断,最终实现异体字的同化处理。

4) 校对系统的建立

彝文古籍图像识别的远景是赋能彝文到汉文的机器翻译技术的开发。然而,古彝文不像拼音文字那样可以直接输出为通用表音的拉丁字母,目前已有的彝文机器翻译的输出语言是四川规范彝文,针对传统彝文的识别输出还只能是借用韩文的Unicode码段AC01-D456(韩旭,2020),或是用户自定义区的Unicode码段E000-FFFF(陈敏等,2006),不利于阅读、印刷传播和后续校对等工作。

为了降低彝文古籍校对的准入门槛,Shao(2021)提出设想,整理了业内首个形义相关的古彝

文基础编码数据库(<https://github.com/wnfrdshao/yi-scan-corpus>),该数据库以古彝文在字集中出现的位置及跨地区字形整合为线索建立了两套形义相关的基础编码。实验表明,即使没有任何彝语基础的实习生也可完成指定的编码任务并达到较高的准确率。

西南大学还开发了一个基于Web的彝文联机手写识别系统,使用户能够便捷地进行彝文文字检索。该系统通过接入彝文识别服务,实现对手写彝文的在线识别(王定旺,2021)。该识别系统目前只针对四川规范彝文。

2.3.4 东巴文古籍文档识别

与东巴文文本分析和识别的相关研究主要分为东巴文单字识别和东巴文古籍机器翻译。东巴文单字识别目前研究较多,而且基于深度学习的方法已经在识别率和可识别字数上超过了传统算法,具有代表性的方法总结如下。

早期的东巴文识别方法大多基于传统的机器学习方法,通常有3个独立的步骤:预处理、特征提取和分类器识别。图像预处理是东巴文识别的一个重要步骤,大津法二值化(Guo和Zhao,2010)、边缘检测(Guo等,2010a)、小波变换(Guo等,2010b)等预处理方法已经广泛应用。特征提取方法例如粗网格、定向元素(Da等,2011)、连接域优先标记算法(Yang和Kang,2018)、特征曲线提取(杨玉婷和康厚良,2019a)、HOG(histogram of oriented gradient)(申彤等,2020)等方法在东巴文识别中都取得了显著的效果。支持向量机(王海燕等,2016)是东巴文识别中应用最多的分类器。具有代表性的研究有Guo等人(2012)基于线性判别分析(linear discriminant analysis, LDA)的纳西文(即东巴文)降维方法,实现对东巴文的较好识别;徐小力等人(2017)提出一种先拓扑特征处理后投影法特征提取的分步骤信息处理方法,并采用模板匹配法进行东巴文识别;杨玉婷和康厚良(2019b)提出东巴文的预处理及基于网格分解的分类识别算法。上述研究虽然取得了一定的成果,然而,这些传统方法是基于独立元素的组合,过程复杂且效率较低,并且针对不同风格的东巴文中具有不同最优值的超参数,仅能够实现部分东巴象形文字的识别,算法的通用性较差且识别准确率有待提高。

近年来,深度学习在图像识别领域取得了突破

性进展,目前已基本替代了基于传统数字图像处理的单字识别方法。但是,在东巴文单字识别方面,直到2019年才有基于深度学习识别方法的论文发表。Hua和Xu(2019)和Wu等人(2019)设计了普通的卷积神经网络,均在200个东巴文上实现了95.20%的识别率;Luo等人(2020)采用ResNet-34(residual network)作为基础网络,同时联合HOG特征提取方法,利用3个分支同时提取东巴文的多尺度特征,在1424个东巴文上取得了89.71%的识别率;谢裕睿和董建娥(2021)通过构造恒等残差块和卷积残差块搭建了ResNet-20模型,在536个东巴文上达到了93.58%的识别率;骆彦龙等人(2022)设计改进的ResNet模型,在包含22万幅图像、1387个东巴文上取得了98.65%的识别率;Luo等人(2023)设计多尺度特征提取网络,并且通过加入注意力机制获得东巴文更多的细节特征,在提出的当前图像规模最大(440 K)、东巴文字种类最多(2546个)的数据集上识别率为99.34%。

2.3.5 满文古籍文档识别

清代中央和地方各级机构中形成了大量的满文古籍档案,以满文历史档案为例,虽然由于战乱、灾害等原因没有完全保存下来,然而保存至今的数量仍然十分巨大,仅中国第一历史档案馆目前保存的满文历史档案就有200余万册。此外,辽宁省档案馆保存着6.5万件20余万册,吉林省档案馆、黑龙江省档案馆、内蒙古自治区档案馆、大连市图书馆等很多其他省、市、县各级档案馆都保存着一定数量的满文历史档案(吴元丰,2015)。这些古籍档案不仅具有重要的历史、文化研究价值,而且对于当代社会经济建设具有重要的参考和凭证作用(赵彦昌,2012)。然而,目前对满文古籍档案的开发利用工作却严重滞后,以中国第一历史档案馆为例,虽然馆藏满文历史档案约占整体馆藏档案的1/5,但是其提调量却只占总提调量的不到4%(赵彦昌和王红娟,2010;赵彦昌和苏亚云,2017)。主要原因是目前满文古籍档案整理加工和利用方式的信息化程度较低,而且由于受满文OCR技术的限制,目前的满文数字化古籍档案几乎都没有提供全文检索和查询功能,从而严重制约了其利用(李健民,2019)。

2009年满语已经被联合国教科文组织列为极度濒临灭绝的语言,目前掌握满文的人员非常稀少,没有看到国外研究的相关报道,国内只有东北大学、

大连民族大学等少数几个团队开展了满文识别相关研究。概括起来人们主要对满文单词识别问题从识别算法构建、训练数据集建设和应用开发等方面发展了研究。

早期主要是对基于基元(字母)切分的满文识别算法进行了研究,即先将满文单词切分成基元序列,然后对基元进行识别,最后根据得到的基元序列识别整体满文单词,建立了基于模糊串匹配策略的识别算法(Zhang等,2004;赵骥等,2006a);考虑到基元的切分与识别对整体精度有较大影响,Zhang等人(2006)建立了一种基于小波神经网络的基元识别算法;赵骥等人(2006b)通过加入后处理机制对拒识词和错误词进行了处理;Xu等人(2017)分别研究了基于多特征集成和多分类器融合的识别算法。尽管满文也是一种拼音文字,每个单词由多个满文字母自上而下拼接而成,但与英文不同,满文单词的字母之间紧密相连没有空隙,而且同一个字母在不同位置有不同写法,这就使得满文单词的字母切分非常困难,从而导致基于基元切分的识别算法的精度通常较低。鉴于此,2017年开始进行了整词的满文识别算法研究,Huang等人(2017)建立了基于有向无环图支持向量机,且空间金字塔池化深度卷积神经网络(Li等,2018;Zheng等,2018)的识别算法,比基元切分方法取得了更高的精度,不足之处是该类算法需要每类单词都有大量的训练样本,目前只能处理满文日常用语的识别问题。受该情况启示,郑蕊蕊等人(2021)对小样本情况下的满文识别,提出了一种基于 N 元ECOC的大类别 K -shot满文识别方法。由于基于整词方法的基本思想是将整个文字看成一个不可分割的整体,从而转换文字识别问题为大类别图像分类问题,而满文是一种拼音文字,基于整词的方法很难覆盖所有满文单词,因此,近年来人们对基于隐式字符(片段)序列模型(如CNN+RNN+CTC或Attention)的满文识别方法进行了研究(Wang等,2023),但相关成果还较少。

1)训练数据集建设方面。目前尚无公开的满文数据集。中国第一历史档案馆项目组(张晶等,2019;郑蕊蕊等,2021)建立的一个包含666类满文单词的涵盖文本行、单词和基元三级信息的数据集,包含21245个样本图像的数据集,但是没有公开。

2)满文识别系统应用开发方面。鉴于满文档案全文数据库建设的需要,中国第一历史档案馆、汉王

科技股份有限公司以及潍坊北大青鸟华光照排有限公司,从2015年开始合作开展了国家档案馆科技项目“满文档案图像识别软件的开发与应用”研究,并于2017年基于多特征过切分、多模型集成、音节统计识别和深度神经网络等技术成功研发了一套满文档案图像识别软件系统(张晶等,2019)。但是该系统对满文档案中使用最为广泛的一种字体的单词识别精度只能达到86.6%,远低于目前汉字等通用文字的识别精度。单词识别仍然是满文档案信息化的一个瓶颈问题。

除单词识别以外,满文古籍档案版面分析等相关研究还几乎没有,只能查到满文文档行分割、单词提取(张晶等,2019;赫欣等,2022)和文档印章检测与去除(卢海涛等,2018)两项相关成果。

由于满文古籍档案的版面包括公文原件、编年体汇抄的档簿、修书馆档案和满文舆图等多种,有楷书、行书、草书和篆书等多种字体,书写载体有纸张、木牌和丝织品等多种,而且长期保存会出现墨水渗透、膨胀、变色和污渍等劣变现象,目前还很少有考虑这些问题的研究。此外,训练数据通常需要满文专业人员人工标注,如今满文专业人员极少,而且满文档案中的常见满文单词数量巨大、背景和书写形式各异,很难建立涵盖所有单词并具有一定多样性的数据集,因此数据严重缺乏与获取困难是满文古籍档案识别与分析面临的一大挑战。

2.4 场景文字检测与识别

2.4.1 场景藏文检测与识别

场景文本检测是场景理解和文字识别领域的重要任务之一(王紫霄等,2023),是目前OCR领域的一个热点研究问题,旨在开放式的环境下对各类不同场景中的文字实现检测和识别。相比于文档识别工作,通常场景中的文本数量不多,但存在多字体、多方向以及多语言的情况,且文本的背景也较为复杂,其技术挑战性更高。基于滑动窗口对文本进行检测,分别使用DBNet(differentiable binarization network)进行文本检测、预训练的MobileNetV3作为特征提取器、CRNN实现场景藏文文本识别(仁青东主等,2021)。针对场景中藏汉文本同时出现的情况,基于CNN对不同类型的文本区域进行分类,以降低后续识别的复杂度;基于区域扩展的方法对场景图像中文本区域进行检测(李金成等,2021);发布具有相当规模的藏汉双语场景图像数据集(bilingual

Tibetan-Chinese scene image dataset, BiTCSD), 并就如何生成更加逼真的场景文字图像进行了深入探讨(Li等, 2021a; 郝玉胜等, 2022)。

2.4.2 场景维吾尔文识别

针对具有粘性、基线和斜透视形态等特征的场
景维吾尔族文本图像, 提出了一种改进注意力机制
的端到端场景维吾尔族文本识别模型, 通过引入协
调注意机制, 按通道关系沿着宽度和高度方向聚集
空间特征, 用于预测文本特征在行、列中的差异, 强
化特征提取能力, 在自建的自然场景文本图像数据
上达到92.78%的识别准确率(Wang和Ao, 2022)。

2.4.3 场景多文种识别

科大讯飞科研团队提出了场景风格可迁移的图
文生成技术、感受野可控的对比式预训练技术, 实现
低资源下场景多文种的有效建模。

1) 场景风格可迁移的图文生成方法

场景风格可迁移的图文生成方法, 其核心技术
点是通过自适应示例规范化(adaptive instance nor-
malization, AdaIN)的方式来编码和控制风格(Huang
等, 2017)。研究者的实验发现通过改变归一化层的
缩放和平移参数可以调节风格, 并且由于归一化层
的仿射变换参数不需要预先设置嵌入空间, 而是直
接从风格图像中学习, 从而实现任意风格转换。然
而, 将这一方案应用于文本图像的风格迁移时, 面临
3大问题: (1)文字结构复杂, 变化灵活, 不能出现结
构扭曲失真的问题; (2)文本图像分为文字与背景区
域, 合成后的图像风格需要语义对齐; (3)需要适应
不同语种、不同场景下的风格转换, 需要很强的泛化
性。于是提出了一种适合文本图像的风格迁移算
法, 有4点原创性改进: (1)采用匹配对图像C-GAN
的训练方式。采用语义分割对现有的文本行图像进
行分割获取图像对, 将分割图作为风格迁移的内容
输入, 文本行图像作为目标输出, 用于约束网络的训
练, 保证内容结构不扭曲丢失, 同时对齐不同类别风
格。(2)不重合采样文本行图像分别用于内容输入与
风格输入, 完全解耦内容与风格信息。(3)计算合成
图像与目标图像在特征层的距离来替换像素间的距
离损失, 防止过拟合, 增强模型在不同语种场景下的
泛化性。(4)在GAN网络结构设计上, 采用Patch-
GAN(Isola等, 2017)+多尺度判别器的结构, 增强文
字生成细节。

上述多场景多语种数据合成系统合成效果真

实、性能稳定, 如图4所示。

将系统合成的数据和真实数据一起组合成弱监
督数据集, 用于多语种各类场景文本识别(scene
text recognition, STR), 效果提升显著, 如表5所示。
基线系统的训练数据由每个语种50万行真实数据
构成, 合成系统的训练数据包括与基线系统相同的
真实数据以及合成50万行数据。从结果可见, 本方
案在充分利用分割图做内容对齐后, 最大程度保证
了风格抽取和合成内容的完整性, 同时在解耦内容
信息与风格信息后, 引入多尺度多区块判别器, 显著
提升了复杂风格和背景样例的合成稳定性, 从而显
著提升了多场景多语种自适应识别能力。



图4 条件归一化风格嵌入系统合成效果

Fig. 4 Synthesis effect of conditional normalization style embedding system

表5 基于多模块数据合成的系统的识别准确率

Table 5 System recognition accuracy based on multi-patch data synthesis

	/%			
方案	维吾尔语	哈萨克语	阿拉伯语	马来语
基线系统	97.18	95.85	96.02	96.24
合成系统	98.39	97.07	97.72	97.89

2) 感受野可控的对比式预训练方法

在场景文字识别领域, 目前主流的研究工作大
多是基于深度神经网络的识别模型。数据驱动是深
度神经网络模型的特点, 海量有标注的文本行图像
是实现性能可靠的文字识别模型的前提。现有的主
流方法都是利用海量的合成文本图像数据进行有监
督训练。合成文本图像的显著优势是获取成本低
廉; 缺陷是字体和图像背景风格单一, 缺乏多样性,
与真实数据差异性大。真实数据虽然能带来更丰富
的多样性, 但是其获取代价昂贵且耗时。

虽然真实场景下的有标注文本图像数据难以获
取, 但真实场景下的无标注文本图像却很容易获得。

只需利用已有的文本检测模型在图像中检测文本框,便能抠取得到文本行图像,所得文本行图像是未标注的。因此,希望以自监督学习的方式获取有助于文字识别的丰富图像表征。为了以自监督预训练的方式学习得到有助于文字识别的图像表征,核心技术点之一是设计了文本行图像内的对比预测编码(contrastive predictive coding, CPC)(Van den Oord 等,2019)作为代理任务,由此为图像表征引入了序列相关性。

所提出的自监督 STR-CPC 预训练的模型结构包括编码器网络和上下文网络。文本图像中的左右顺序在时间上对应过去和未来,STR-CPC 引导模型预测未来时间的潜在空间(latent space)特征。将文本行图像视做从左到右的序列进行处理,使用 InfoNCE(noise contrastive estimation)损失函数,正样本自预测区间内的特征,而负样本则是由均匀分布随机采样获得。有别于语音等形式的序列信号,文本行图像内的字符位置紧凑,特征图上某一点的感受野范围重叠很大,使得预测区间内正样本相似度很高,模型可以通过捷径完成对比预测编码任务,因此无法获取有用表征。该方案核心技术点之二是设计了宽度因果卷积(widthwise causal convolution, WCC)并解决了感受野重叠过大导致信息泄露的问题。

考虑到宽度因果卷积编码器和普通卷积编码器的差异,模型预训练和微调之间存在结构不匹配的问题。微调过程中会降低文字识别模型的性能。本方案核心技术点之三是设计了渐进式恢复训练策略(progressive recovery training strategy, PRTS)。模型包含 4 个模块:矫正、特征提取、序列建模和结果预测。

在微调阶段,渐进将预训练阶段中的宽度因果卷积恢复为普通卷积,以减少预训练与微调之间的不匹配问题。这种渐进式恢复训练策略不仅保留了预训练阶段的视觉序列相关性,而且能提高下游文字识别模型的性能。

上述自监督对比预测编码预训练方案在场景文字识别的基准测试结果如表 6 所示。从实验结果可以看出,在 10% 训练数据量下,所设计方案使 CRNN 和 TRBA(TPS-Resnet-BiLSTM-Attn)基线模型分别获得 17.8% 和 25.6% 的性能提升。一系列的消融实验也证明了所设计的宽度因果卷积和渐进式恢复训练策略的有效性与必要性。

表 6 经 STR-CPC 预训练后下游文字识别模型的微调效果
Table 6 Fine-tuning recognition results after pre-training with STR-CPC

		/%	
方法	预训练方法	总准确率	总准确率 (10% 训练量)
CRNN(Shi 等, 2017)	None	74.8	40.9
	Vanilla CPC	75.0	54.3
	STR-CPC w/o PRTS	74.2	56.1
	STR-CPC	77.0	58.7
TRBA(Baek 等, 2021)	None	86.6	41.3
	Vanilla CPC	86.1	61.9
	STR-CPC w/o PRTS	86.5	64.7

3 发展趋势与展望

少数民族文字文本识别研究的发展总是广泛地受到外部环境的影响,从所研究对象的角度来说,其研究的热点大致可依次归纳为印刷体识别、联机/脱机手写识别、古籍识别以及场景文本识别。印刷体文档版面简洁,字符的获取相对简单,大部分的研究工作都集中在字符的特征选择以及分类器设计上,其技术目前已经较为成熟,并且得到实际应用。但类似少数民族报刊、杂志等复杂版面的文档分析、识别和版面复原还有待开发。联机手写字符识别是一个有约束的问题,在获取了一定的样本后,主流技术都可以为该问题提供一个较好的解决方案,而不同的文种也具有不同的挑战,如藏文主要是相似字丁多、字体差异大和藏文字符集规模较大等,使得任意书写的联机识别输入还有待后续开发。脱机手写字符识别随着经济文化的持续发展,这方面的需求也势必增加,无论在技术上还是应用上都有更多的拓展空间。对古籍文档智能分析与识别的任务,更多的工作集中在前期的图像预处理、版面分析、文本行的切分、字符切分以及相应的数据集建设上。继续开展少数民族文献和古籍数字化保护是“中华古籍保护计划”的主要内容,其开发利用是服务当代、面向未来的工作。场景文字识别是目前最热门的研究领域之一,在民族地区,有大量的双语场景,因此在进行场景文字识别时,一方面要提高文字的检测和识别性能,另一方面可以考虑辅助使用场景中出现

的汉字信息,提高识别准确率。

从文档的智能分析与识别技术角度来讲,少数民族文字文本识别作为OCR领域的一个应用分支,经历了从经典机器学习到深度学习的转变。经典的图像处理、特征提取、分类模型以及各种深度学习框架都自然而然地从汉文、英文主流文字文本识别应用到了少数民族文本识别,除了根据各文种自身的特点进行一定的调整外,不同文种版面特征、行或者列的特点、行或者列内字符或者词之间的粘连等,以及同一文种不同字体之间较大的差异,都影响少数民族文字文本识别性能。当然无论哪种文字的文本识别,其技术发展的路径大致是一致的。总体都是图像理解任务所涉及的数据集、评价指标和方法(张言等,2023)等。特别对于古籍智能分析与识别来说,如同敦煌古籍残片的处理(郑玉彤等,2023),不同文种对残片的处理中除了几何特征外,还有复杂的文字语义特征有待研究。

从对新型通信技术和智能应用的需求而言,要求手写识别技术具有较高的精度和可接受的速度。尽管各文种手写识别的研究取得了一定成果,为了更好地提高相应手写识别的在线和离线准确性,仍需要做大量的研究。如维哈柯签名识别与鉴别技术的难点主要有:1)样本类型的多样化。目前普遍使用的公开数据集的样本来自扫描,而现实应用中,利用手机等设备对签名拍照也是数据采集过程中需要考虑的。2)样本的多语种性。现有的公共数据集是以单文种为主,对于多文种的离线手写签名研究较少,该技术在未来仍然有很大的开发潜力,尤其是一些小语种的文字。3)不同语种样本之间的相似性。我国新疆地区的部分少数民族(维吾尔族、哈萨克族和柯尔克孜族)文字与不少中亚国家的文字具有很多相似之处,这对于多语种的数据集和识别都有一定的挑战。

综上所述,针对不同文种的民族文字文本应用需求和现有研究方面的不足,首先都是低资源条件下的不同民族文字文本识别问题,类别数巨大而有些类别训练样本稀少的长尾问题,以及有些文种字符的开放集识别问题。样本不均衡问题、未来的少数民族文字文本识别至少需要在如下几个方面重点开展工作:1)结合文字自身特点和应用场景开展研究和应用系统开发。不同少数民族文字本身特征鲜明,例如乌金体藏文的基线以及字头特征、藏文音节

的构字规则、藏文古籍长条书的版面形制、场景中藏汉双语的大规模出现等,都是可以进一步利用的信息;藏文古籍木刻版、各种手写草书版的智能分析与识别,可以进行定制性的开发,以便逐渐满足应用需求。2)积极拓展新的应用方向。大部分民族文字文本识别和汉文文本识别的发展相比具有一定的差距,在教育、安全、民生等方面的落地应用还有很大的拓展空间。3)大语言模型向少数民族文字文本识别的迁移。近两年来,基于语言模型的超大规模多模态预训练模型在各类文本相关下游任务中取得了瞩目的成效,语言模型和OCR技术结合顺理成章。4)多语种联合建模。训练数据稀缺和不同语种独立部署使得硬件资源开销过大,进一步研究多文种联合建模实用技术,是大幅降低资源开销、显著提高多语种识别效果与泛化性的有效途径。

4 结 语

对民族文字、文本的分析与识别研究,主要包括藏文、蒙文、维文、彝文、满文、东巴等,在印刷体文本识别、联机手写识别、古籍文档识别和场景文本识别方面进行了梳理和总结,自2000年以来,紧随主流语言文字如英文、汉文的相关研究,中国的少数民族文字文本分析和识别研究得到高度重视并取得了一系列的研究成果,有些已达到实用的目的,获得了良好的社会效益。随着新技术的发展,在端到端多语种图文识别统一建模方案研究中,涉及到更多语种,如英、法、德、日、俄、韩、阿拉伯、马来、印尼、印地等。然而,少数民族文字文本的分析与识别,均属于低资源条件下的研究,而且各语种相关方法和技术研究都缺乏相应的公开数据集,还有待进一步完善。此外,如何基于大模型、多模态和零样本少样本学习等开展少数民族文字文本分析与识别的研究和应用,也具有重要意义。

致 谢 本文由中国图象图形学学会文档图像分析与识别专委会组织撰写,该专委会链接为<https://www.csig.org.cn/16/201801/49333.html>。

参考文献(References)

Baek J, Matsui Y and Aizawa K. 2021. What if we only use real datasets

- for scene text recognition? Toward scene text recognition with fewer labels [EB/OL]. [2024-01-25].
<https://arxiv.org/pdf/2103.04400.pdf>
- Cai R D Z, Hua Q C R and Huang H M. 2022. Tibetan printed recognition based on syllable segmentation. *Computer Engineering and Design*, 43(9): 2594-2600 (才让当知, 华却才让, 黄鹤鸣. 2022. 基于音节切分的藏文印刷体识别. *计算机工程与设计*, 43(9): 2594-2600) [DOI: 10.16208/j.issn1000-7024.2022.09.025]
- Cai X J and Huang H M. 2016. Feature extraction method of off-line handwritten Tibetan character based on multiple projection. *Computer Technology and Development*, 26(3): 93-96 (蔡晓娟, 黄鹤鸣. 2016. 基于多投影的脱机手写藏文字符特征提取方法. *计算机技术与发展*, 26(3): 93-96) [DOI: 10.3969/j.issn.1673-629X.2016.03.022]
- Cai Z Q and Wang W L. 2018. Online handwriting Tibetan character recognition based on two-dimensional discriminant locality alignment// *Proceedings of the 1st Chinese Conference on Pattern Recognition and Computer Vision*. Guangzhou, China: Springer: 88-98 [DOI: 10.1007/978-3-030-03338-5_8]
- Chen M, Wu X and Ma D J. 2006. Design of character set encoding for ancient Yi script in Guizhou. *Science and Technology Market*, (7): #222 (陈敏, 吴颢, 马德江. 2006. 贵州古彝文字符集编码设计. *科技经济市场*, (7): #222) [DOI: 10.3969/j.issn.1009-3788.2006.07.216]
- Chen S X, Han X, Lin X Y, Liu Y and Wang M G. 2020. MSER and CNN-based method for character detection in ancient Yi books. *Journal of South China University of Technology (Natural Science Edition)*, 48(6): 123-133 (陈善雄, 韩旭, 林小渝, 刘云, 王明贵. 2020. 基于MSER和CNN的彝文古籍文献的字符检测方法. *华南理工大学学报(自然科学版)*, 48(6): 123-133) [DOI: 10.12141/j.issn.1000-565X.190812]
- Chen S X, Wang X L, Han X, Liu Y and Wang M G. 2019. A recognition method of ancient Yi character based on deep learning. *Journal of Zhejiang University (Science Edition)*, 46(3): 261-269 (陈善雄, 王小龙, 韩旭, 刘云, 王明贵. 2019. 一种基于深度学习的古彝文识别方法. *浙江大学学报(理学版)*, 46(3): 261-269) [DOI: 10.3785/j.issn.1008-9497.2019.03.001]
- Cui S D, Su Y L, Ren Q D E J, and Ji Y T. 2022. An end-to-end network for irregular printed Mongolian recognition. *International Journal on Document Analysis and Recognition*, 25(1): 41-50 [DOI: 10.1007/s10032-021-00388-y]
- Da M J, Zhao J Y, Suo G J and Guo H. 2011. Online handwritten Naxi pictograph digits recognition system using coarse grid//*Computer Science for Environmental Engineering and EcoInformatics*. Kunming, China: Springer: 390-396 [DOI: 10.1007/978-3-642-22694-6_55]
- Dey A, Mitra S and Das N. 2020. Handwritten Tibetan character recognition based on ELM using modified HOG features//*Proceedings of 2020 IEEE Calcutta Conference (CALCON)*. Kolkata, India: IEEE: 451-456 [DOI: 10.1109/CALCON49167.2020.9106541]
- Dhondrub R, Tsering T and Tashi N. 2021. Study on a synthesis method for training data of ancient Tibetan book character recognition. *Plateau Science Research*, 5(3): 84-91 (仁青东主, 头旦才让, 尼玛扎西. 2021. 藏文古籍文字识别训练数据的合成方法研究. *高原科学研究*, 5(3): 84-91) [DOI: 10.16249/j.cnki.2096-4617.2021.03.011]
- Ding X Q, Wang Y W. 2017. *Character Recognition: Principles, Methods and Practice*. Beijing: Tsinghua University Press (丁晓青, 王言伟. 2017. *文字识别: 原理、方法和实践*. 北京: 清华大学出版社)
- Fan D E J and Gao G L. 2016. DNN-HMM for large vocabulary Mongolian offline handwriting recognition//*Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition*. Shenzhen, China: IEEE: 72-77 [DOI: 10.1109/ICFHR.2016.0026]
- Fan D E J, Gao G L and Wu H J. 2018. MHW Mongolian offline handwritten dataset and its application. *Journal of Chinese Information Processing*, 32(1): 89-95 (范道尔吉, 高光来, 武慧娟. 2018. MHW 蒙古文脱机手写数据库及其应用. *中文信息学报*, 32(1): 89-95) [DOI: 10.3969/j.issn.1003-0077.2018.01.012]
- Fan D E J, Gao G L and Wu H J. 2019. Sub-word based Mongolian offline handwriting recognition//*Proceedings of 2019 International Conference on Document Analysis and Recognition*. Sydney, Australia: Springer: 246-253 [DOI: 10.1109/ICDAR.2019.00048]
- Gao G L, Su X D, Wei H X and Gong Y Y. 2011. Classical Mongolian words recognition in historical document//*Proceedings of the 11th International Conference on Document Analysis and Recognition*. Beijing, China: Springer: 692-697 [DOI: 10.1109/ICDAR.2011.145]
- Gao D G, Hou Y, Gao H M and Suolang Q Z. 2023. Study on detection and recognition of Tibetan Wumei printing multi-fonts text. *Plateau Science Research*, 7(1): 92-100 (高定国, 侯闫, 高红梅, 索朗曲珍. 2023. 乌梅印刷多字体藏文文本的检测与识别. *高原科学研究*, 7(1): 92-100) [DOI: 10.16249/j.cnki.2096-4617.2023.01.011]
- Gao L C, Li Y B, Du L, Zhang X P, Zhu Z Y, Lu N, Jin L W, Huang Y S and Tang Z. 2022. A survey on table recognition technology. *Journal of Image and Graphics*, 27(6): 1898-1917 (高良才, 李一博, 都林, 张新鹏, 朱子仪, 卢宁, 金连文, 黄永帅, 汤帆. 2022. 表格识别技术研究进展. *中国图象图形学报*, 27(6): 1898-1917) [DOI: 10.11834/jig.220152]
- Gheni Z, Mahpirat, Yadikar N, Aysa Y and Ubul K. 2018. Gray level co-occurrence matrix feature weighting fusion for Uyghur signature verification. *Computer Engineering and Design*, 39(4): 1195-1201 (祖丽皮亚·艾尼, 麦合甫热提, 努尔毕亚·亚地卡尔, 尤努斯·艾沙, 库尔班·吾布力. 2018. 灰度共生矩阵特征加权融合的维文签名鉴别. *计算机工程与设计*, 39(4): 1195-1201) [DOI: 10.16208/j.issn1000-7024.2018.04.052]
- Guo H, Yin J H and Zhao J Y. 2012. Feature dimension reduction of

- NaXi pictographs characters recognition based on LDA. *International Journal of Computer Science Issues*, 9(6): 90-96
- Guo H and Zhao J Y. 2010. Segmentation method for NaXi pictograph character recognition. *Journal of Convergence Information Technology*, 5(6): 87-98 [DOI: 10.4156/jcit.vol5.issue6.9]
- Guo H, Zhao J Y and Da M J. 2010a. A preprocessing method for NaXi pictograph character recognition. *Journal of Convergence Information Technology*, 5(2): 59-66 [DOI: 10.4156/jcit.vol5.issue2.7]
- Guo H, Zhao J Y, Da M J and Li X N. 2010b. NaXi pictographs edge detection using lifting wavelet transform. *Journal of Convergence Information Technology*, 5(5): 203-210 [DOI: 10.4156/jcit.vol5.issue5.23]
- Han X. 2020. *Research and Implementation of Character Detection and Recognition in Ancient Yi Books*. Chongqing: Southwest University (韩旭. 2020. 彝文古籍字符检测和识别的研究与实现. 重庆: 西南大学) [DOI: 10.27684/d.cnki.gxndx. 2020. 001060]
- Han Y H, Wang W L, Liu H M and Wang Y Q. 2019. A combined approach for the binarization of historical Tibetan document images. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(14): #1954038 [DOI: 10.1142/S0218001419540387]
- Hao Y S, Wang W L, Li J C and Lin Q. 2022. A method for bilingual Tibetan-Chinese scene image dataset synthesis and text detection. *Journal of Computer-Aided Design and Computer Graphics*, 34(4): 592-604 (郝玉胜, 王维兰, 李金成, 林强. 2022. 藏汉双语场景图像数据集合成及文本检测方法. 计算机辅助设计与图形学学报, 34(4): 592-604) [DOI: 10.3724/SP.J.1089.2022.18954]
- He X, Li H J, Zhou Y, Zheng R R and He J J. 2022. APSENet: a text line detection method of Manchu archives based on instance segmentation network. *Journal of Minzu University of China (Natural Sciences Edition)*, 31(1): 19-27 (赫欣, 李厚杰, 周瑜, 郑蕊蕊, 贺建军. 2022. APSENet: 一种基于实例分割网络的满文档案文本行检测方法. 中央民族大学学报(自然科学版), 31(1): 19-27) [DOI: 10.3969/j.issn.1005-8036.2022.01.003]
- Hedayati F, Chong J K and Keutzer K. 2011. Recognition of Tibetan Wood Block Prints with generalized hidden Markov and kernelized modified quadratic distance function//*Proceedings of 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*. Beijing, China: ACM: #12 [DOI: 10.1145/2034617.2034631]
- Hou Y, Gao D G and Gao H M. 2023. Wujin printing multi-fonts text detection and recognition in Tibetan. *Computer Engineering and Design*, 44(4): 1058-1065 (侯同, 高定国, 高红梅. 2023. 乌金印刷多字体藏文的文本检测与识别. 计算机工程与设计, 44(4): 1058-1065) [DOI: 10.16208/j.issn1000-7024.2023.04.014]
- Hu P F, Wang W L, Li Q Q and Wang T J. 2021. Touching text line segmentation combined local baseline and connected component for Uchen Tibetan historical documents. *Information Processing and Management*, 58(6): #102689 [DOI: 10.1016/j.ipm.2021.102689]
- Hua R X and Xu X L. 2019. Intelligent classification on images of Dongba ancient books. *The Journal of Engineering*, 2019 (23): 9039-9042 [DOI: 10.1049/joe.2018.9177]
- Huang D, Li M, Zheng R R, Xu S and Bi J J. 2017. Synthetic data and DAG-SVM classifier for segmentation-free Manchu word recognition//*Proceedings of 2017 International Conference on Computing Intelligence and Information System*. Nanjing, China: IEEE: 46-50 [DOI: 10.1109/CIIS.2017.15]
- Huang H M and Da F P. 2014. Sparse representation-based classification algorithm for optical Tibetan character recognition. *Optik*, 125(3): 1034-1037 [DOI: 10.1016/j.ijleo.2013.07.101]
- Huang H M, Da F P and Han X X. 2014. Wavelet transform and gradient direction based feature extraction method for off-line handwritten Tibetan letter recognition. *Journal of Southeast University (English Edition)*, 30(1): 27-31 [DOI: 10.3969/j.issn.1003-7985.2014.01.006]
- Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/ICCV.2017.167]
- Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]
- Jia X D. 2017. *Application Research of Handwritten Yi Character Recognition Technology Based on Deep Learning*. Beijing: Minzu University of China (贾晓东. 2017. 基于深度学习的手写彝文识别技术应用研究. 北京: 中央民族大学)
- Jiang W, Lu Z Y and Li J. 2013. Recognition of handwritten Uyghur character based on directional element feature. *Microelectronics and Computer*, 30(10): 97-100 (姜文, 卢朝阳, 李静. 2013. 基于方向线索特征的手写体维文字符识别. 微电子学与计算机, 30(10): 97-100) [DOI: 10.19304/j.cnki.issn1000-7180.2013.10.025]
- Jiang Z W, Ding X Q and Peng L R. 2015. Character model optimization for segmentation-free Uyghur text line recognition. *Journal of Tsinghua University (Science and Technology)*, 55(8): 873-877, 883 (姜志威, 丁晓青, 彭良瑞. 2015. 针对无切分维吾尔语文本行识别的字符模型优化. 清华大学学报(自然科学版), 55(8): 873-877, 883) [DOI: 10.16511/j.cnki.qhdxxb.2015.08.010]
- Kadier N, Peng L R and Halimulati. 2015. Uyghur and Arabic recognition methods based on hmm and statistical language model. *Computer Applications and Software*, 32(1): 171-174 (努尔艾力·喀迪尔, 彭良瑞, 哈力木拉提. 2015. 一种基于HMM和统计语言模型的维吾尔文及阿拉伯文识别方法. 计算机应用与软件, 32(1): 171-174) [DOI: 10.3969/j.issn.1000-386x.2015.01.044]
- Kaldan T and Vijayalakshmi A. 2021. TenzinNet for handwritten Tibetan numeral recognition. *International Journal of Information Technology*, 13(4): 1679-1682 [DOI: 10.1007/s41870-021-00711-0]
- Kang Y K, Wei H X, Zhang H and Gao G L. 2019. Woodblock-printing

- Mongolian words recognition by bi-LSTM with attention mechanism//Proceedings of 2019 International Conference on Document Analysis and Recognition. Sydney, Australia: Springer: 910-915 [DOI: 10.1109/ICDAR.2019.00150]
- Kojima M, Nunomiya C, Kawamura T, Akiyama Y, Kawazoe Y and Klmura M. 1995. Recognition of printed tibetan characters by object oriented designing//Proceedings of Annual Conference Japan Society of Information and Knowledge, 3(78): 53-60 [DOI: 10.2964/jsikproc.3.0_53]
- Li J C, Hao Y S, Wang W L, Wang T J and Li Q Q. 2021a. Scene text detection based on expanding the text center region for bilingual Tibetan-Chinese. International Journal of Pattern Recognition and Artificial Intelligence, 2021, 35(13): #2153007 [DOI: 10.1142/S0218001421530074]
- Li J C, Wang X J, Wang W L, Lin Q and Hu P F. 2021. Text line segmentation of Tibetan historical documents based on text core regions combined with expansion growth. Laser and Optoelectronics Progress, 58(2): #0210008 (李金成, 王筱娟, 王维兰, 林强, 胡鹏飞. 2021. 结合文字核心区域和扩展生长的藏文古籍文本行切分. 激光与光电子学进展, 58(2): #0210008) [DOI: 10.3788/LOP202158.0210008]
- Li J M. 2019. Exploration of Manchu archives informatization. China Archives, (3): 44-45 (李健民. 2019. 满文档案信息化工作探索. 中国档案, (3): 44-45)
- Li M. 2019. Research on Online Mongolian Handwriting Recognition System based on Deep Learning. Hohhot: Inner Mongolia University (李敏. 2019. 基于深度学习的联机蒙古文手写识别系统研究. 呼和浩特: 内蒙古大学)
- Li M, Zheng R R, Xu S, Fu Y and Huang D. 2018. Manchu word recognition based on convolutional neural network with spatial pyramid pooling//Proceedings of the 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Beijing, China: IEEE: 1-6 [DOI: 10.1109/CISP-BMEI. 2018. 8633131]
- Li P C, Zhu J D, Peng L R and Guo Y B. 2016. RNN based Uyghur text line recognition and its training strategy//2016 12th IAPR Workshop on Document Analysis Systems (DAS). Santorini, Greece: IEEE: 19-24 [DOI: 10.1109/DAS.2016.20]
- Li W Y, Mahpirat, Kang W X, Aysa A and Ubul K. 2021b. Multi-lingual hybrid handwritten signature recognition based on deep residual attention network//Proceedings of the 15th Chinese Conference on Biometric Recognition. Shanghai, China: Springer: 148-156 [DOI: 10.1007/978-3-030-86608-2_17]
- Li W Y, Mahpirat, Xu X B, Aysa A and Ubul K. 2022a. A simple convolutional neural network for small sample multi-lingual offline handwritten signature recognition//Proceedings of the 16th Chinese Conference on Biometric Recognition. Beijing, China: Springer: 393-403 [DOI: 10.1007/978-3-031-20233-9_40]
- Li Y Z, Wang Y L and Liu Z Z. 2012. Study on printed Tibetan character recognition technology. Journal of Nanjing University (Natural Sciences), 48(1): 55-62 (李永忠, 王玉雷, 刘真真. 2012. 藏文印刷体字符识别技术研究. 南京大学学报(自然科学), 48(1): 55-62) [DOI: 10.13232/j.cnki.jnju.2012.01.006]
- Li Z H and Gao G L. 2003. Extraction of features of Mongolian printed character recognition. Computer Technology and Development, 13(11): 117-119 (李振宏, 高光来. 2003. 印刷体蒙古文文字识别中常用特征的获取. 微机发展, 13(11): 117-119) [DOI: 10.3969/j.issn.1673-629X.2003.11.042]
- Li Z H, Gao G L, Hou H X and Li W. 2003. The research of printed Mongolian character recognition. Acta Scientiarum Naturalium Universitatis Neimongol, 34(4): 454-457 (李振宏, 高光来, 侯宏旭, 李伟. 2003. 印刷体蒙古文文字识别的研究. 内蒙古大学学报(自然科学版), 34(4): 454-457) [DOI: 10.3969/j.issn.1000-1638.2003.04.020]
- Li Z J, Wang W L and Cai Z Q. 2019a. Historical document image binarization based on edge contrast information//Proceedings of 2019 Computer Vision Conference on Advances in Computer Vision. Las Vegas, USA: Springer: 614-628 [DOI: 10.1007/978-3-030-17795-9_44]
- Li Z J, Wang W L, Chen Y and Hao Y S. 2019b. A novel method of text line segmentation for historical document image of the uchen Tibetan. Journal of Visual Communication and Image Representation, 61: 23-32 [DOI: 10.1016/j.jvcir. 2019.01.021]
- Li Z J, Wang W L, Wang Y Q and Zhang Q X. 2022b. Character recognition of Tibetan historical document in Uchen font: dataset and bench mark. Journal of Computational Methods in Sciences and Engineering, 22(5): 1779-1794 [DOI: 10.3233/JCM-226167]
- Liu C L, Jin L W, Bai X, Li X H and Yin F. 2023. Frontiers of intelligent document analysis and recognition: review and prospects. Journal of Image and Graphics, 28(8): 2223-2252 (刘成林, 金连文, 白翔, 李晓辉, 殷飞. 2023. 文档智能分析与识别前沿: 回顾与展望. 中国图象图形学报, 28(8): 2223-2252) [DOI: 10.11834/jig. 221112]
- Liu F and Tashi N. 2020. Study on the extraction method on characteristics of Tibetan characters based on strokes. Plateau Science Research, 4(3): 105-110 (刘芳, 尼玛扎西. 2020. 基于笔划的藏文字符特征提取方法研究. 高原科学研究, 4(3): 105-110) [DOI: 10.16249/j.cnki.2096-4617.2020.03.014]
- Liu H M, Bi X H and Wang W L. 2019. Layout analysis of historical Tibetan documents//Proceedings of the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). Chengdu, China: IEEE: 74-78 [DOI: 10.1109/ICAIBD.2019.8837040]
- Liu H M, Shi R M, Bi X H, Wang X Y and Wang W L. 2022. Line segmentation of Tibetan ancient books based on A* algorithm. Journal of Physics: Conference Series, 2356(1): #012046 [DOI: 10.1088/1742-6596/2356/1/012046]
- Liu J, Ma L L and Wu J. 2016. Online handwritten Mongolian word recognition using MWRCNN and position maps//Proceedings of the

- 15th International Conference on Frontiers in Handwriting Recognition. Shenzhen, China: IEEE: 60-65 [DOI: 10.1109/ICFHR.2016.0024]
- Liu J, Ma L L and Wu J. 2017. Online handwritten Mongolian word recognition using a novel sliding window method with recurrent neural networks//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 189-194 [DOI: 10.1109/ICDAR.2017.39]
- Liu L L and Zhao H. 2012. Comparing study of two language models of online Uyghur handwritten recognition. *Computer Applications and Software*, 29(9): 151-153 (柳玲玲, 赵晖. 2012. 联机手写维吾尔文单词识别中两种语言模型的比较研究. *计算机应用与软件*, 29(9): 151-153) [DOI: 10.3969/j.issn.1000-386x.2012.09.040]
- Liu S, Zhu Z X, Ma Z Q, Tong F and Xie R. 2009. Design and realization on character segmentation method for Yi language based on connected components. *Journal of South-Central University for Nationalities (Natural Science Edition)*, 28(2): 86-89 (刘赛, 朱宗晓, 马志强, 童飞, 谢蓉. 2009. 基于连通域的彝文文字切分算法的设计与实现. *中南民族大学学报(自然科学版)*, 28(2): 86-89) [DOI: 10.3969/j.issn.1672-4321.2009.02.021]
- Liu S Q, Jin L B and Miao F. 2020. Textual restoration of occluded Tibetan document pages based on side-enhanced U-Net. *Journal of Electronic Imaging*, 29(6): #063006 [DOI: 10.1117/1.JEI.29.6.063006]
- Lu H T, Wu L, Zhou J Y, Zheng R R and He J J. 2018. Seal detection on Manchu document based on faster R-CNN and data augmentation. *Journal of Dalian Minzu University*, 20(5): 455-459 (卢海涛, 吴磊, 周建云, 郑蕊蕊, 贺建军. 2018. 基于 Faster R-CNN 及数据增广的满文文档印章检测. *大连民族大学学报*, 20(5): 455-459) [DOI: 10.3969/j.issn.1009-315X.2018.05.016]
- Luo H N, Xu D, Yang B and Zhang H Y. 2020. Multi-scale feature fusion based Dongba character recognition//Proceedings of the 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). Harbin, China: IEEE: 1571-1575 [DOI: 10.1109/ICMCCE51767.2020.00344]
- Luo Y L, Bi X J, Wu L C and Li X L. 2022. Dongba pictographs recognition based on improved residual learning. *CAAI Transactions on Intelligent Systems*, 17(1): 79-87 (骆彦龙, 毕晓君, 吴立成, 李霞丽. 2022. 基于改进残差学习的东巴象形文字识别. *智能系统学报*, 17(1): 79-87) [DOI: 10.11992/tis.202112009]
- Luo Y L, Sun Y W and Bi X J. 2023. Multiple attentional aggregation network for handwritten Dongba character recognition. *Expert Systems with Applications*, 213: #118865 [DOI: 10.1016/j.eswa.2022.118865]
- Ma L L, Liu H D and Wu J. 2011. MRG-OHTC database for online handwritten Tibetan character recognition//Proceedings of 2011 International Conference on Document Analysis and Recognition. Beijing, China: IEEE: 207-211 [DOI: 10.1109/ICDAR.2011.50]
- Ma L L, Liu J and Wu J. 2016. A new database for online handwritten Mongolian word recognition//Proceedings of the 23rd International Conference on Pattern Recognition. Cancun, Mexico: IEEE: 1131-1136 [DOI: 10.1109/ICPR.2016.7899788]
- Ma L L and Wu J. 2013. A recognition system for online handwritten Tibetan characters//Graphics Recognition. New Trends and Challenges. Seoul, Korea (South): Springer: 99-107 [DOI: 10.1007/978-3-642-36824-0_10]
- Ma L L and Wu J. 2016. Online unconstrained handwritten Tibetan character recognition using statistical recognition method. *Himalayan Linguistics*, 15(1): 31-40 [DOI: 10.5070/H915130066]
- Mo L F, Mahpirat, Zhu Y L, Mamat H and Ubul K. 2020. Off-line signature recognition based on non-downsampled contourlet transform. *Computer Engineering and Design*, 41(12): 3472-3478 (莫龙飞, 麦合甫热提, 朱亚俐, 吾尔尼沙·买买提, 库尔班·吾布力. 2020. 基于非下采样轮廓波变换的离线签名识别. *计算机工程与设计*, 41(12): 3472-3478) [DOI: 10.16208/j.issn1000-7024.2020.12.026]
- Ngodrup, Putsesern, Daluosanglangjie, Zhao D C, Liu F and Bianbawangdui. 2009. Study on printed Tibetan character recognition. *Computer Engineering and Applications*, 45(24): 165-169, 172 (欧珠, 普次仁, 大罗桑朗杰, 赵栋才, 刘芳, 边巴旺堆. 2009. 印刷体藏文文字识别技术研究. *计算机工程与应用*, 45(24): 165-169, 172) [DOI: 10.3778/j.issn.1002-8331.2009.24.049]
- Qiao L, Li Z S, Cheng Z Z and Li X. 2023. SCID: a Chinese characters invoice-scanned dataset in relevant to key information extraction derived of visually-rich document images. *Journal of Image and Graphics*, 28(8): 2298-2313 (乔梁, 李再升, 程战战, 李玺. 2023. SCID: 用于富含视觉信息文档图像中信息提取任务的扫描中文票据数据集. *中国图象图形学报*, 28(8): 2298-2313) [DOI: 10.11834/jig.220911]
- Qin H, Li Y J, Liang Q K and Wang Y N. 2023. AsymcNet: a document images-relevant asymmetric geometry correction network. *Journal of Image and Graphics*, 28(8): 2314-2329 (秦海, 李艺杰, 梁桥康, 王耀南. 2023. 针对文档图像的非对称式几何校正网络. *中国图象图形学报*, 28(8): 2314-2329) [DOI: 10.11834/jig.220426]
- Rexit A, Muhammad M, Xu X B, Kang W X, Aysa A and Ubul K. 2022. Multilingual handwritten signature recognition based on high-dimensional feature fusion. *Information*, 13(10): #496 [DOI: 10.3390/info13100496]
- Rowinski Z and Keutzer K. 2016. Namsel: an optical character recognition system for Tibetan text. *Himalayan Linguistics*, 15(1): 12-30 [DOI: 10.5070/H915129937]
- Shao W Y. 2021. Unheard Voices and Alternative Pasts: Deciphering Chronicles of Southwest Yi and Its Layered Ranges of Signification. Columbus: Ohio State University
- Shen T, Zhuang J J, Li W S, Wang Y M, Xia Y F, Zhang Z J, Zhang X and Yang J Q. 2020. Research on recognition of Dongba script by

- a combination of HOG feature extraction and support vector machine. *Journal of Nanjing University (Natural Science)*, 56(6): 870-876 (申彤, 庄建军, 黎文斯, 王昀牧, 夏一飞, 张志俭, 张鑫, 杨继琼. 2020. 基于HOG特征提取和支持向量机的东巴文识别. *南京大学学报(自然科学)*, 56(6): 870-876 [DOI: 10.13232/j.cnki.jnju.2020.06.009])
- Shi B G, Bai X and Yao C. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (11) : 2298-2304 [DOI: 10.1109/TPAMI.2016.2646371]
- Simayi W, Ibrayim M and Hamdulla A. 2019. A study of RNN based online handwritten Uyghur word recognition using different word transcriptions//*Proceedings of the 11th International Conference on Simulation Tools and Techniques*. Chengdu, China: Springer: 518-527 [DOI: 10.1007/978-3-030-32216-8_50]
- Simayi W, Ibrayim M, Tursun D and Hamdulla A. 2013. Research on on-line Uyghur character recognition technology based on center distance feature//*IEEE International Symposium on Signal Processing and Information Technology*. Athens, Greece: IEEE: 293-298 [DOI: 10.1109/ISSPIT.2013.6781896]
- Simayi W, Ibrayim M, Tursun D and Hamdulla A. 2016. A survey on the classifiers in on-line handwritten Uyghur character recognition system. *International Journal of Hybrid Information Technology*, 9(3): 189-198 [DOI: 10.14257/ijhit.2016.9.3.18]
- Su X D, Gao G L, Wei H X and Bao F L. 2015. Enhancing the Mongolian historical document recognition system with multiple knowledge-based strategies//*Proceedings of the 22nd International Conference on Neural Information Processing*. Istanbul, Turkey: Springer: 536-544 [DOI: 10.1007/978-3-319-26535-3_61]
- Su X D, Gao G L, Wei H X and Bao F L. 2016. A knowledge-based recognition system for historical Mongolian documents. *International Journal on Document Analysis and Recognition*, 19(3): 221-235 [DOI: 10.1007/s10032-016-0267-1]
- Sun S W and Wei H X. 2022. A Mongolian handwritten word images generation approach based on generative adversarial networks//*Proceedings of 2022 International Joint Conference on Neural Networks*. Padua, Italy: IEEE: 1-8 [DOI: 10.1109/IJCNN5064.2022.9892917]
- Tang J, Silamu W, Xu M M, Xiong L J and Wang M H. 2021. Uyghur scanning body recognition based on deep learning. *Journal of Northeast Normal University (Natural Science Edition)*, 53(1): 71-76 (汤敬, 吾守尔·斯拉木, 许苗苗, 熊黎剑, 王明辉. 2021. 基于深度学习的维吾尔文扫描体识别. *东北师大学报(自然科学版)*, 53(1): 71-76) [DOI: 10.16163/j.cnki.22-1123/n.2021.01.012]
- Ubul K, Ablikim R, Yadikar N, Aysa A and Yibulayin T. 2016. Off-line Uyghur signature recognition technology based on density feature. *Computer Engineering and Design*, 37(8): 2200-2205 (库尔班·吾布力, 热依买·阿不力克木, 努尔毕亚·亚地卡尔, 阿力木江·艾沙, 吐尔根·依布拉克. 2016. 基于密度特征的维吾尔文离线签名识别. *计算机工程与设计*, 37(8): 2200-2205) [DOI: 10.16208/j.issn1000-7024.2016.08.041]
- Ubul K, Adler A, Abliz G, Yasheng M and Hamdulla A. 2012. Off-line Uyghur signature recognition based on modified grid information features//*Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. Montreal, Canada: IEEE: 1056-1061 [DOI: 10.1109/ISSPA.2012.6310446]
- Ubul K, Wang X L, Yimin A, Zhang S J and Yibulayin T. 2018. Multilingual offline handwritten signature recognition based on statistical features//*Proceedings of the 13th Chinese Conference on Biometric Recognition*. Urumqi, China: Springer: 727-735 [DOI: 10.1007/978-3-319-97909-0_77]
- Van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2024-01-15]. <https://arxiv.org/pdf/1807.03748.pdf>
- Wang D H, Wang W L and Qian J J. 2010. 2DPCA and IMLDA method of feature extraction for online handwritten Tibetan recognition//*Proceedings of 2010 International Conference on Networking and Digital Society*. Wenzhou, China: IEEE: 563-566 [DOI: 10.1109/ICNDS.2010.5479269]
- Wang D W. 2021. Research and Application of Online Handwriting Recognition in Yi. Chongqing: Southwest University (王定旺. 2021. 彝文联机手写体识别的研究与应用. 重庆: 西南大学) [DOI: 10.27684/d.cnki.gxndx.2021.003069]
- Wang H and Ding X Q. 2004a. A normalization method of multi-font printed Tibetan characters. *Application Research of Computers*, 21(6): 41-43 (王华, 丁晓青. 2004a. 一种多字体印刷藏文字的归一化方法. *计算机应用研究*, 21(6): 41-43) [DOI: 10.3969/j.issn.1001-3695.2004.06.014]
- Wang H and Ding X Q. 2004b. An algorithm for multi-font printed Tibetan character recognition. *Computer Engineering*, 30(13): 18-20 (王华, 丁晓青. 2004b. 一种多字体印刷藏文字符识别方法. *计算机工程*, 30(13): 18-20) [DOI: 10.3969/j.issn.1000-3428.2004.13.008]
- Wang H Y, Wang H J and Xu X L. 2016. Recognition of Naxi Dongba pictographs based on support vector machine. *Journal of Yunnan University(Natural Sciences Edition)*, 38(5): 730-736 (王海燕, 王红军, 徐小力. 2016. 基于支持向量机的纳西东巴象形文字识别. *云南大学学报(自然科学版)*, 38(5): 730-736) [DOI: 10.7540/j.ynu.20150757]
- Wang J M, Wen Y H, Li Y Q and Gao Y L. 2008. The recognition system of Old-Yi character based on the image segmentation. *Journal of Yunnan Nationalities University (Natural Sciences Edition)*, 17(1): 76-79 (王嘉梅, 文永华, 李燕青, 高雅莉. 2008. 基于图像分割的古彝文字识别系统研究. *云南民族大学学报(自然科学版)*, 17(1): 76-79) [DOI: 10.3969/j.issn.1672-8513.2008.

- 01.019]
- Wang W L, Ding X Q, Chen L and Wang H. 2003. Study on printed Tibetan character recognition. *Computer Engineering*, 29(3): 37-38, 94 (王维兰, 丁晓青, 陈力, 王华. 2003. 印刷体现代藏文识别研究. *计算机工程*, 29(3): 37-38, 94) [DOI: 10.3969/j.issn.1003028.2003.03.014]
- Wang W L, Li Z J, Cai Z Q, Lv X B, Zhaxi C and Han Y H. 2019. Online Tibetan handwriting recognition for large character set on new databases. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(10): #1953003 [DOI: 10.1142/S0218001419530033]
- Wang W L, Lu X B, Cai Z Q, Shen W T, Fu J and Caike Z X. 2017. Online Handwritten sample generated based on component combination for Tibetan-Sanskrit. *Journal of Chinese Information Processing*, 31(5): 64-73 (王维兰, 卢小宝, 蔡正琦, 沈文韬, 付吉, 才科扎西. 2017. 基于部件组合的联机手写“藏文—梵文”样本生成. *中文信息学报*, 31(5): 64-73) [DOI: 10.3969/j.issn.1003-0077.2017.05.009]
- Wang Y N, Huaque C R, Cairang D Z and Huan K Y. 2021. Application of Tibetan image text recognition system in android—Based on mixed attention mechanism neural network model. *Journal of Qinghai Normal University (Natural Science)*, 37(4): 26-33 (王悦凝, 华却才让, 才让当知, 环科尤. 2021. 藏文图像文本识别在安卓系统中的应用——基于混合注意力机制神经网络模型. *青海师范大学学报(自然科学版)*, 37(4): 26-33) [DOI: j.cnki.ssn1001-7542.2021.04.013]
- Wang Y Q, Wang W L and Cai Z Q. 2022. Text region extraction method for historical Tibetan document based on border detection// *Proceedings Volume 12172, International Conference on Electronic Information Engineering and Computer Communication (EIECC 2021)*. Nanchang, China: SPIE: 65-72 [DOI: 10.1117/12.2634657]
- Wang Y W, Ao N X, Guo R, Mamat H and Ubul K. 2022. Scene Uyghur recognition with embedded coordinate attention//*Proceedings of the 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*. Chengdu, China: IEEE: 253-260 [DOI: 10.1109/PRML56267.2022.9882248]
- Wang Z W, Lu S Y, Wang M Q, Wei X and Qi Y J. 2023. AMRE: an attention-based CRNN for Manchu word recognition on a woodblock-printed dataset//*Proceedings of the 29th International Conference on Neural Information Processing. Virtual Event: Springer*: 267-278 [DOI: 10.1007/978-3-031-30108-7_23]
- Wang Z X, Xie H T, Wang Y X and Zhang Y D. 2023. Hierarchical semantics-fused scene text detection. *Journal of Image and Graphics*, 28(8): 2343-2355 (王紫霄, 谢洪涛, 王裕鑫, 张勇东. 2023. 层级语义融合的场景文本检测. *中国图象图形学报*, 28(8): 2343-2355) [DOI: 10.11834/jig.220902]
- Wei H X and Gao G L. 2006. Feature selection of Mongolian characters in the recognition of printed Mongolian characters. *Acta Scientiarum Naturalium Universitatis Neimongol*, 37(6): 694-697 (魏宏喜, 高光来. 2006. 印刷体蒙古文字识别中蒙古文字特征的选择. *内蒙古大学学报(自然科学版)*, 37(6): 694-697) [DOI: 10.3969/j.issn.1000-1638.2006.06.021]
- Wei H X and Gao G L. 2007a. A method of layout analysis for Mongolian document images based on connected components. *Journal of Inner Mongolia University*, 38(5): 586-590 (魏宏喜, 高光来. 2007a. 一种基于连通域的蒙古文文档图像版面分析方法. *内蒙古大学学报(自然科学版)*, 38(5): 586-590) [DOI: 10.3969/j.issn.1000-1638.2007.05.022]
- Wei H X and Gao G L. 2007b. A skew detection method of Mongolian document images. *Journal of Inner Mongolia University*, 38(4): 458-462 (魏宏喜, 高光来. 2007b. 蒙文文档图像的倾斜检测方法. *内蒙古大学学报(自然科学版)*, 38(4): 458-462) [DOI: 10.3969/j.issn.1000-1638.2007.04.018]
- Wei H X and Gao G L. 2019. A holistic recognition approach for woodblock-print Mongolian words based on convolutional neural network//*Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE*: 2726-2730 [DOI: 10.1109/ICIP.2019.8803226]
- Wei H X, Liu C, Zhang H, Bao F L and Gao G L. 2019. End-to-end model for offline handwritten Mongolian word recognition//*Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing. Dunhuang, China: Springer*: 220-230 [DOI: 10.1007/978-3-030-32236-6_19]
- Wei H X, Liu K X, Zhang J and Fan D E J. 2021a. Data augmentation based on CycleGAN for improving woodblock-printing Mongolian words recognition//*Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer*: 526-537 [DOI: 10.1007/978-3-030-86337-1_35]
- Wei H X, Zhang H, Zhang J and Liu K X. 2021b. Multi-task learning based traditional Mongolian words recognition//*Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE*: 1275-1281 [DOI: 10.1109/ICPR48806.2021.9412326]
- Wu G X, Liu X L, Jiang Z L and Hua R X. 2019. Dongba classical ancient books image classification method based on ReN-Soft plus convolution residual neural network//*Proceedings of the 14th IEEE International Conference on Electronic Measurement and Instruments (ICEMI)*. Changsha, China: IEEE: 398-404 [DOI: 10.1109/ICEMI46757.2019.9101450]
- Wu J J, Zhao K, Yang Z Y, Yin B, Liu C and Dai L R. 2023. End-to-end multilingual text recognition based on byte modeling//*Proceedings of the 12th International Conference on Image and Graphics. Nanjing, China: Springer*: 128-137 [DOI: 10.1007/978-3-031-46311-2_11]
- Wu Y F. 2015. *Manchu Archives and Historical Exploration*. Shenyang: Liaoning Minzu Press (吴元丰. 2015. *满文档案与历史探究*. 沈阳: 辽宁民族出版社)
- Xamxidin N, Mahpirat, Yao Z X, Aysa A and Ubul K. 2022. Multilin-

- gual offline signature verification based on improved inverse discriminator network. *Information*, 13(6): #293 [DOI: 10.3390/info13060293]
- Xie Y R and Dong J E. 2021. Research on Dongba hieroglyph recognition using ResNet network. *Computer Era*, (1): 6-10 (谢裕睿, 董建娥. 2021. 基于ResNet网络的东巴象形文字识别研究. *计算机时代*, (1): 6-10) [DOI: 10.16644/j.cnki.cn33-1094/tp.2021.01.002]
- Xiong L J, Silamu W and Xu M M. 2021. Design and implementation of printed Uyghur recognition system based on Django. *Journal of Zhengzhou University(Natural Science Edition)*, 53(3): 9-14 (熊黎剑, 吾守尔·斯拉木, 许苗苗. 2021. 基于Django印刷体维吾尔文识别系统的设计与实现. *郑州大学学报(理学版)*, 53(3): 9-14) [DOI: 10.13705/j.issn.1671-6841.2020277]
- Xu S, Li M, Zheng R R and Michael S. 2017. Manchu character segmentation and recognition method. *Journal of Discrete Mathematical Sciences and Cryptography*, 20(1): 43-53 [DOI: 10.1080/09720529.2016.1177965]
- Xu X L, Jiang Z L, Wu G X, Wang H J and Wang N. 2017. Identification method of Dongba pictograph based on topological characteristic and projection method. *Journal of Electronic Measurement and Instrumentation*, 31(1): 150-154 (徐小力, 蒋章雷, 吴国新, 王红军, 王宁. 2017. 基于拓扑特征和投影法的东巴象形文识别方法研究. *电子测量与仪器学报*, 31(1): 150-154) [DOI: 10.13382/j.jemi.2017.01.022]
- Xu Y M and Du P P. 2017. Offline handwritten Uighur character recognition based on grapheme analysis//*Proceedings of the 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. Beijing, China: IEEE: 832-835 [DOI: 10.1109/ICSESS.2017.8343040]
- Xu Y M, Lu Z Y and Li J. 2013. Handwritten Uyghur character recognition based on radical dictionary and time division direction feature. *Journal of Jilin University (Engineering and Technology Edition)*, 43(3): 740-746 (许亚美, 卢朝阳, 李静. 2013. 部件字典结合时分方向特征的手写维吾尔字符识别. *吉林大学学报(工学版)*, 43(3): 740-746) [DOI: 10.7964/jdxbgxb201303030]
- Xu Y M and Xue J L. 2019. Offline handwritten Uighur word recognition based on segmentation-driven and two-level DTW//*Proceedings of the 2nd IEEE International Conference on Computer and Communication Engineering Technology (CCET)*. Beijing, China: IEEE: 182-186 [DOI: 10.1109/CCET48361.2019.8989253]
- Yang C, Du J, Xue M B and Zhang J S. 2023. An encoder-decoder based generation model for online handwritten mathematical expressions. *Journal of Image and Graphics*, 28(8): 2356-2369 (杨晨, 杜俊, 薛莫白, 张建树. 2023. 用于在线手写公式合成的编解码网络. *中国图象图形学报*, 28(8): 2356-2369) [DOI: 10.11834/jig.220894]
- Yang C, Liu C, Fang Z Y, Han Z, Liu C L and Yin X C. 2023. Open set text recognition technology. *Journal of Image and Graphics*, 28(6): 1767-1791 (杨春, 刘畅, 方治屿, 韩铮, 刘成林, 殷绪成. 2023. 开放集文字识别技术. *中国图象图形学报*, 28(6): 1767-1791) [DOI: 10.11834/jig.230018]
- Yang Y T and Kang H L. 2018. A novel algorithm of contour tracking and partition for Dongba hieroglyph//*Proceedings of the 13th Conference on Image and Graphics Technologies and Applications*. Beijing, China: Springer: 157-167 [DOI: 10.1007/978-981-13-1702-6_16]
- Yang Y T and Kang H L. 2019a. Research on the extracting algorithm of Dongba hieroglyphic feature curves. *Journal of Graphics*, 40(3): 591-599 (杨玉婷, 康厚良. 2019a. 东巴象形文字特征曲线提取算法研究. *图学学报*, 40(3): 591-599) [DOI: 10.11996/JGJ.2095-302X.2019030591]
- Yang Y T and Kang H L. 2019b. Dongba hieroglyphic classification algorithm based on grid resolution. *Software Guide*, 18(9): 196-198 (杨玉婷, 康厚良. 2019b. 基于网格分解的东巴象形文字分类算法研究. *软件导刊*, 18(9): 196-198) [DOI: 10.11907/rjdk.181810]
- Yimiti A, Liu J C and Sulaiman D. 2014. Study on character segmentation and recognition technology of Uyghur in image with complex background. *Journal of Xinjiang Normal University (Natural Sciences Edition)*, 33(1): 65-68 (阿地力·依米提, 刘吉超, 杜力坤·苏来曼. 2014. 复杂背景图像中维吾尔文字切分与识别技术的研究. *新疆师范大学学报(自然科学版)*, 33(1): 65-68) [DOI: 10.14100/j.cnki.1008-9659.2014.01.015]
- Zhang C and Wang W L. 2021. Character segmentation for historical Uchen Tibetan document based on structure attributes. *Laser Optoelectronics Progress*, 58(20): #2010020 (张策, 王维兰. 2021. 基于结构属性的乌金体藏文古籍字符切分. *激光与光电子学进展*, 58(20): #2010020) [DOI: 10.3788/LOP202158.2010020]
- Zhang C, Wang W L and Zhang G W. 2022a. Construction of a character dataset for historical Uchen Tibetan documents under low-resource conditions. *Electronics*, 11(23): #3919 [DOI: 10.3390/electronics11233919]
- Zhang C, Wang W L, Liu H M, Zhang G W and Lin Q. 2022b. Character detection and segmentation of historical Uchen Tibetan documents in complex situations. *IEEE Access*, 10: 25376-25391 [DOI: 10.1109/ACCESS.2022.3151886]
- Zhang G W, Wang W L, Zhang C, Zhao P H and Zhang M K. 2023. HUTNet: an efficient convolutional neural network for handwritten Uchen Tibetan character recognition. *Big Data*, 11(5): 387-398 [DOI: 10.1089/big.2021.033]
- Zhang G Y, Li J J, He R W and Wang A X. 2004. An offline recognition method of handwritten primitive Manchu characters based on strokes//*Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition*. Kokubunji, Japan: IEEE: 432-437 [DOI: 10.1109/IWFHR.2004.16]
- Zhang G Y, Li J J and Wang A X. 2006. A new recognition method for the handwritten Manchu character unit//*Proceedings of 2006 International Conference on Machine Learning and Cybernetics*. Dalian,

- China: IEEE: 3339-3344 [DOI: 10.1109/ICMLC.2006.258471]
- Zhang H, Wei H X, Bao F L and Gao G L. 2017. Segmentation-free printed traditional Mongolian OCR using sequence to sequence with attention model//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 585-590 [DOI: 10.1109/ICDAR.2017. 101]
- Zhang H R, Chen W, Su X D, Guo H and Xu H L. 2021. An efficient local word augment approach for Mongolian handwritten script recognition//Proceedings of the 16th International Conference on Document Analysis and Recognition. Lausanne, Switzerland: Springer: 429-443 [DOI: 10.1007/978-3-030-86337-1_29]
- Zhang J, Xu S, He J J, Li M and Zheng R R. 2019. Segmentation and extraction method for Manchu words based on seam craving. Journal of Chinese Information Processing, 33(2): 81-88 (张晶, 许爽, 贺建军, 李敏, 郑蕊蕊. 2019. 基于缝隙裁剪的满文单词分割和提取方法研究. 中文信息学报, 33(2): 81-88) [DOI: 10.3969/j.issn. 1003-0077.2019.02.011]
- Zhang J X, Luo C J, Jin L W, Guo F J and Ding K. 2022. Marior: margin removal and iterative content rectification for document dewarping in the wild [EB/OL]. [2024-01-15].
<https://arxiv.org/pdf/2207.11515.pdf>
- Zhang S J, Mahpira, Mamat H, Zhu Y L and Ubul K. 2022. Uyghur offline handwritten signature verification based on texture features. Computer Engineering and Design, 41(3): 770-776 (张淑婧, 麦合甫热提, 吾尔尼沙·买买提, 朱亚俐, 库尔班·吾布力. 2020. 基于纹理特征的维吾尔文离线手写签名鉴别. 计算机工程与设计, 41(3): 770-776) [DOI: 10.16208/j. issn1000-7024.2020. 03.027]
- Zhang X Q, Ma L L, Duan L J, Liu Z Y and Wu J. 2018. Layout analysis for historical Tibetan documents based on convolutional denoising autoencoder. Journal of Chinese Information Processing, 32(7): 67-73, 81 (张西群, 马龙, 段立娟, 刘泽宇, 吴健. 2018. 基于卷积降噪自编码器的藏文历史文献版面分析方法. 中文信息学报, 32(7): 67-73), 81) [DOI: 10.3969/j.issn.1003-0077.2018. 07. 009]
- Zhang Y, Li Q, Shen H W, Zeng G Y, Zhou Y, Ma C, Zhang Y and Wang W P. 2023. Text-centric image analysis techniques: a critical review. Journal of Image and Graphics, 28(8): 2253-2275 (张言, 李强, 申化文, 曾港艳, 周宇, 马灿, 张远, 王伟平. 2023. 以文字为中心的图像理解技术综述. 中国图象图形学报, 28(8): 2253-2275) [DOI: 10.11834/jig.220968]
- Zhao P H, Wang W L, Zhang G W and Lu Y Q. 2021a. Alleviating pseudo-touching in attention U-Net-based binarization approach for the historical Tibetan document images. Neural Computing and Applications, 2021, 35 (19): 13791-13802 [DOI: 10.1007/S00521-021-06512-7]
- Zhao P H, Wang W L, Cai Z Q, Zhang G W and Lu Y Q. 2021b. Accurate fine-grained layout analysis for the historical Tibetan document based on the instance segmentation. IEEE Access, 9: 154435-154447 [DOI: 10.1109/ACCESS. 2021.3128536]
- Zhao J, Li J J, Zhang G Y and Wang J. 2006a. Design and implementation of off-line handwritten document recognition system of Manchu manuscript. Pattern Recognition and Artificial Intelligence, 19(6): 801-805 (赵骥, 李晶皎, 张广渊, 王杰. 2006a. 脱机手写体满文本识别系统的设计与实现. 模式识别与人工智能, 19(6): 801-805) [DOI: 10.3969/j.issn.1003-60592006.06.020]
- Zhao J, Li J J, Wang L J and Zhang J S. 2006b. Research on the post-processing of Manchu character recognition based on hidden Markov model. Journal of Chinese Information Processing, 20 (4): 63-67 (赵骥, 李晶皎, 王丽君, 张继生. 2006b. 基于HMM的满文文本识别后处理的研究. 中文信息学报, 20(4): 63-67) [DOI: 10.3969/j.issn.1003- 0077. 2006.04.009]
- Zhao Q H and Wang W L. 2023. Zero-RADCE: zero-reference residual attention deep curve estimation for low-light historical Tibetan document image enhancement. Visual Communications and Image Processing, 2(1): 1-8 [DOI: 10.23977/vcip.2023.020101]
- Zhao Q H, Wang W L and Yu Y Y. 2022. Retinex-LTNet: low-light historical Tibetan document image enhancement based on improved Retinex-Net//Proceedings of the 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI' 22). Dongguan, China: Association for Computing Machinery: 785-791 [DOI: 10.1145/3584376.3584516]
- Zhao Y C. 2012. Research on Manchu Archives. Shanghai: World Publishing Corporation (赵彦昌. 2012. 满文档案研究. 上海: 世界图书出版公司)
- Zhao Y C and Su Y Y. 2017. A review of the arrangement and research of Manchu archives in the 21st century. Manzu Minority Research, 3: 55-61, 73 (赵彦昌, 苏亚云. 2017. 21世纪以来满文档案整理与研究述评. 满族研究, (3): 55-61, 73) [DOI: 10.3969/j. issn.1006-365X.2010.04.011]
- Zhao Y C and Wang H J. 2010. Research on the development and utilization of Manchu archives. Manzu Minority Research, (4): 47-52 (赵彦昌, 王红娟. 2010. 满文档案开发利用研究. 满族研究, (4): 47-52)
- Zheng R R, Li M, He J J, Bi J J and Wu B C. 2018. Segmentation-Free multi-font printed Manchu word recognition using deep convolutional features and data augmentation//Proceedings of the 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Beijing, China: IEEE: 1-6 [DOI: 10.1109/CISP-BMEI.2018.8633208]
- Zheng R R, Xin S Y, Zhou Y, Liu W P, Dang J W and He J J. 2021. A K-shot Manchu recognition method for large category based on N-ary ECOC. Journal of Zhengzhou University (Natural Science Edition), 53(4): 53-60 (郑蕊蕊, 辛守宇, 周瑜, 刘文鹏, 党佳伟, 贺建军. 2021. 一种基于N元ECOC的大类别K-shot满文识别方法. 郑州大学学报(理学版), 53(4): 53-60) [DOI: 10.13705/j. issn. 1671-6841.2021178]
- Zheng Y T, Li X L, Yin Z X, Gao G and Weng Y. 2023. Multi-feature

fusion based automatic reconstruction in related to Chinese ancient manuscript fragments of Dunhuang. *Journal of Image and Graphics*, 28(8): 2330-2342 (郑玉彤, 李雪龙, 殷梓轩, 高歌, 翁彧. 2023. 多特征融合的敦煌古籍残片自动缀合. *中国图象图形学报*, 28(8): 2330-2342) [DOI: 10.11834/jig.220896]

Zhi X X, Gao D G, Zhao Q J, Li S W and Qu C. 2021. Text detection in Tibetan ancient books: a benchmark//*Proceedings of the 2nd IEEE International Conference on Pattern Recognition and Machine Learning (PRML)*. Chengdu, China: IEEE: 254-259 [DOI: 10.1109/PRML52754. 2021. 9520727]

Zhou F M, Wang W L and Lin Q. 2018. A novel text line segmentation method based on contour curve tracking for Tibetan historical documents. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(10): #1854025 [DOI: 10.1142/S0218001418540253]

Zhu L H and Wang J M. 2010. Off-line handwritten Yi character recognition based on the multi-classifier ensemble with combination features. *Journal of Yunnan University of Nationalities (Natural Sciences Edition)*, 19(5): 329-333 (朱龙华, 王嘉梅. 2010. 基于组合特征的多分类器集成的脱机手写体彝文字识别. *云南民族大学学报(自然科学版)*, 19(5): 329-333) [DOI: 10.3969/j.issn.1672-8513. 2010.05.005]

Zhu Z X and Wu X L. 2012. Principles and implementation of an off-line printed Yi character recognition system. *Computer Technology and Development*, 22(2): 85-88, 92 (朱宗晓, 吴显礼. 2012. 脱机印刷体彝族文字识别系统的原理与实现. *计算机技术与发展*, 22(2): 85-88), 92) [DOI: 10.3969/j.issn.1673-629X.2012.02.022]

作者简介

王维兰, 女, 教授, 主要研究方向为文档图像分析与识别、图像处理与模式识别。E-mail: wangweilan@xbmu.edu.cn

胡金水, 男, 工程师, 主要研究方向为文字识别和计算机视觉。E-mail: jshu@iflytek.com

魏宏喜, 男, 教授, 主要研究方向为文档图像分析与识别、计算机视觉。E-mail: cswhx@imu.edu.cn

库尔班·吾布力, 男, 教授, 主要研究方向为模式识别、计算机视觉和自然语言处理。E-mail: kurbanu@xju.edu.cn

邵文苑, 女, 博士, 讲师, 主要研究方向为彝文编码与古籍数字化。E-mail: wyshao@shu.edu.cn

毕晓君, 女, 教授, 主要研究方向为自然语言处理。

E-mail: bixiaojun@hrbeu.edu.cn

贺建军, 男, 副教授, 主要研究方向为机器学习与文档图像分析。E-mail: jianjunhe@live.com

李振江, 男, 副教授, 主要研究方向为信息内容安全和藏文信息处理。E-mail: lzj6843@gsupl.edu.cn

丁凯, 男, 博士, 高级工程师, 主要研究方向为人工智能、模式识别和自然语言算法。E-mail: danny_ding@intsig.net

金连文, 男, 教授, 主要研究方向为深度学习、机器学习、文字识别、计算机视觉、图像处理和人工智能系统。

E-mail: eelwjin@scut.edu.cn

高良才, 男, 副教授, 主要研究方向为模式识别。

E-mail: glc@pku.edu.cn