



文献 CSTR:

32001.14.11-6035.csd.2025.0069.zh



文献 DOI:

10.11922/11-6035.csd.2025.0069.zh

数据 DOI:

10.57760/sciencedb.25676

文献分类: 计算机科学技术

收稿日期: 2025-04-23

录用日期: 2025-06-10

发表日期: 2026-02-24

* 论文通信作者

毕晓君: bixiaojun@hrbeu.edu.cn

面向 OCR 的满文古籍单词数据集

孙海鹏^{1,2}, 陶文豪², 毕晓君^{1,2*}

1. 中央民族大学中国少数民族语言文学学院, 北京 100081

2. 民族语言智能分析与安全治理教育部重点实验室, 北京 100081

摘要: 满文古籍承载着清代历史的脉络与文化精髓, 是中华文明发展历程中不可或缺的重要遗产。由于满文古籍的释读存在识别困难的问题, 因此基于深度学习的满文单词识别研究显得尤为重要。然而, 在训练过程中, 深度学习模型高度依赖高质量且规模化的标注数据集。为此, 本研究构建了一个高质量且多样化的满文古籍单词数据集。鉴于传统分割方法对数据集的收集存在一定的弊端, 例如人力标注成本高、耗费时间长等。本数据集采用半自动化手段, 通过结合计算机技术辅助和人工核验的方法, 进而实现有效的单词提取。在此基础上, 进一步引入图像质量筛查与标注复核机制, 系统剔除模糊、倾斜和误切样本, 确保图像清晰可辨、标注准确一致。本研究以《国家图书馆藏满文汉合璧古籍珍本丛书》为研究基础, 提取出 24280 张满文古籍单词图像, 涵盖 2428 个满文单词, 其中每个单词图像数量为 10 张。在数据收集和整理过程中, 发现所收集的满文单词图像具有显著的规模优势, 可为相关研究提供坚实的数据基础。

关键词: 满文文献; 满文古籍单词; 光学字符识别; 文化遗产保护

数据库(集)基本信息简介

数据库(集)名称	面向 OCR 的满文古籍单词数据集
数据通信作者	毕晓君 (bixiaojun@hrbeu.edu.cn)
数据作者	孙海鹏、陶文豪、毕晓君
数据时间范围	1733–1867年
地理区域	中国境内
数据量	304.85 MB
数据格式	*.jpg, *.png
数据服务系统网址	https://doi.org/10.57760/sciencedb.25676
基金项目	国家自然科学基金 (62236011); 国家社科基金重大项目 (20&ZD279)。
数据库(集)组成	数据集共包含2428个满文单词, 每个单词对应一个以拉丁转写命名的子文件夹(如toktoho), 每个文件夹内包含该单词的10张图像样本, 总计24280张图像。图像文件采用统一命名规则, 例如“0_010_1_1_toktoho”表示该图像来源于第10页、第1列、第1个位置, 且对应的单词为“toktoho”(为其拉丁转写形式)。

引言

近年来,基于深度学习的满文古籍单词识别方法已成为计算机视觉领域的研究热点^[1-2]。这一方法依托于深度学习模型对满文古籍图像的特征进行自动提取,避免了手工设计特征的烦琐过程,在识别性能上显著优于传统方法。然而,深度学习模型的有效训练高度依赖于高质量且规模化的标注数据集,这在处理低资源语言的识别任务时尤为关键^[3]。满文单词具有字形复杂和高相似度的特性,对数据集的多样性和精确性提出了更为苛刻的要求。因此,构建高质量且多样化的满文古籍图像数据集不仅是提升模型表现的必要条件,也为推动满文 OCR 领域的发展提供了重要保障^[4-6]。

满文作为清朝的官方文字之一,记载着大量珍贵的文献资料,涵盖政治、文化、经济、军事、外交以及天文等多个领域。其形式多样,包括图书、档案以及碑刻等,具有不可替代的历史与学术价值^[7-9]。然而,满文古籍历史悠久、保存条件不佳以及领域专家匮乏等多方面的因素,导致众多的满文史料未能得到有效地挖掘^[10-11]。为此,本研究通过计算机技术辅助和人工核验的方法对满文史料进行收集与整理,构建了一种面向 OCR 的满文古籍单词数据集。鉴于传统分割方法对数据集的收集存在人力标注成本高、耗时间长等因素,本研究采用半自动化的手段对满文古籍图像进行切分,有效克服了传统方法的局限性,提升了单词分割的准确性和效率^[12-14]。本数据集来源于《国家图书馆藏满汉文合璧古籍珍本丛书》^[15],从中提取出 24280 张满文古籍单词图像,涵盖 2428 个满文单词。其中,每个单词图像数量为 10 张。通过调研,本研究所构建的满文古籍单词数据集是目前为止规模最大的,且每张图像均符合严格的规范要求,为后续满文古籍单词识别研究提供了坚实的数据支撑。

根据已发表的文献调研可知,文本识别的相关研究主要集中在中、英文等主流文字上^[16-18],已有多大规模数据集发布,例如:ICDAR2019-LSVT、ICDAR2017-RCTW-17、中文街景文字识别以及中文文档文字识别等。然而,古籍相关的文本识别数据集仍较为稀缺,亟须进一步的研究和开发。近年来,已有一些与古籍相关的公开数据集发布,例如水书字符数据集,该数据集包含 226000 张水书字符图像,主要用于水书字符识别研究^[19];以及甲骨文多模态数据集,该数据集涵盖 10077 件甲骨文图像,提供字符检测、识别以及阅读顺序预测等任务的标注数据^[20]。随着上述大规模数据集的发布,众多先进模型相继被提出,并在文本识别任务中取得了显著的成果。在满文单词识别方面,目前已有学者构建了特定领域的满文单词数据集。例如,2018 年大连民族大学 Zheng 等人^[21]以《满语 365 句》作为研究基础,提取了 666 个常用的印刷体满文单词,并收集了 2135 张图片。此外,2022 年北京交通大学 Wang 等人^[22]选择满汉合璧标音辞典《御制增订清文鉴》,从中提取出 4506 个满文单词和 21245 张图片。尽管上述研究已完成相关数据集的构建,但至今这些数据集尚未公开。因此,为了促进满文单词识别任务的研究,公开发布本满文古籍单词数据集。

1 数据采集和处理方法

1.1 数据采集方法

本研究中的原始数据来源于国家图书馆(<https://www.nlc.cn/web/index.shtml>),其名称为《国家图书馆藏满汉文合璧古籍珍本丛书》^[15],共二十五册,系清代及更早时期刊印的满汉合璧典籍,具有较高的文献价值与研究潜力。该丛书收录《周易》《书经》《诗经》《礼记》《春秋》《孝经》《大学》《中庸》《论语》《孟子》《满汉全字十二头》《清语易言》《醒世要言》等 13 部典籍,

涵盖儒家经典与语言类文献。该文献多为宫廷内府本、刻本或抄本，采用满汉对照排版，兼具历史文献与语言学研究价值。

本研究在此基础上，结合表 1 所列文献内容，开展满文 OCR 相关数字化处理与分析，旨在为清代语言制度研究与儒学典籍的数字人文探索提供基础数据与技术支持。数据集经高质量图像整理与版面结构提取，适用于训练满文 OCR 模型。

表 1 《国家图书馆藏满汉文合璧古籍珍本丛书》

Table 1 Bibliography of the Series of Rare Ancient Books in Manchu and Chinese Housed in the National Library

序号	书名	版本	卷数
1	《周易》	乾隆三十年（1765）武英殿刻本	四卷
2	《书经》	清刻本	六卷
3	《诗经》	乾隆三十三年（1768）武英殿刻本	八卷
4	《礼记》	乾隆四十八年（1783）武英殿刻本	三十卷
5	《春秋》	乾隆四十九年（1784）武英殿刻本	六十四卷附纲领一卷
6	《孝经》	咸丰六年（1856）武英殿刻本	不分卷
7	《大学》	乾隆二十年（1755）武英殿刻本	不分卷
8	《中庸》	乾隆二十年（1755）武英殿刻本	不分卷
9	《论语》	乾隆二十年（1755）武英殿刻本	不分卷
10	《孟子》	乾隆二十年（1755）武英殿刻本	不分卷
11	《满汉全字十二头》	雍正十一年（1733）京都宏文阁刻本	二卷
12	《清语易言》	乾隆三十一年（1766）刻本	不分卷
13	《醒世要言》	同治六年（1867）刻本	四卷

1.2 数据处理方法

在数据处理过程中，对上述满文古籍文献进行扫描，并将其转化为电子版。其扫描示例如图 1 所示。从图中可以看出，满文古籍文献排版较为规整，遵循自上而下、由左至右的阅读顺序，且满文与转汉文并行排列。为确保文献管理的高效性，对于每一册扫描的文档均独立存储，按页数顺序命名。该命名方式不仅便于后续的查找与处理，还为文献的进一步分析与数字化提供了有力的保障。此外，在扫描满文古籍文献的过程中，严格遵循高精度的标准，确保图像有较高的清晰度，从而为后续的满文单词识别任务提供了可靠的数据支撑。

进一步，采用现有的满文识别方法对扫描到的满文古籍文档进行初步识别。在初步识别的过程中，首先对文档中的每一个满文单词进行切割。由于传统切割方法存在一定的局限性，本研究通过计算机辅助技术和人工核验的方法，第一步采用人工标注的方式，通过多边形框对每个单字进行精准框选，确保字符边界的准确性；第二步依据标注的坐标信息进行图像切割，从而实现了更为精确的分割效果。其次，对分割出的每一个满文单词进行转写，生成相应的字母形式。最终，对所有识别出的结果按页进行排序，每一页上的满文单词均依照从上至下、从左至右的顺序进行编号，其初步识别结果如图 2 所示。

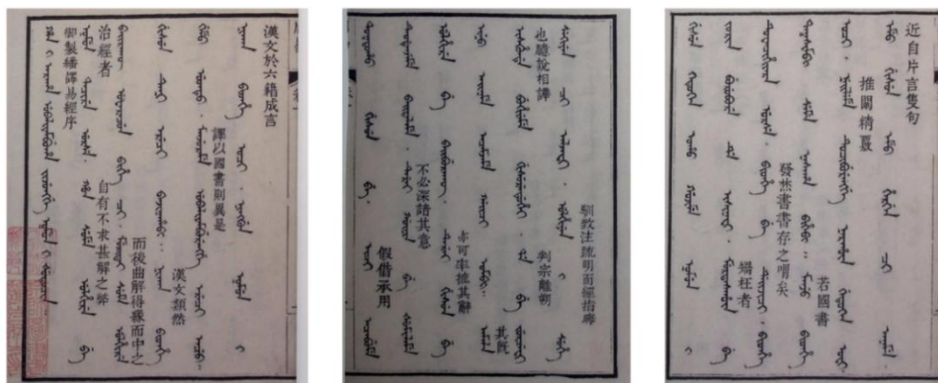
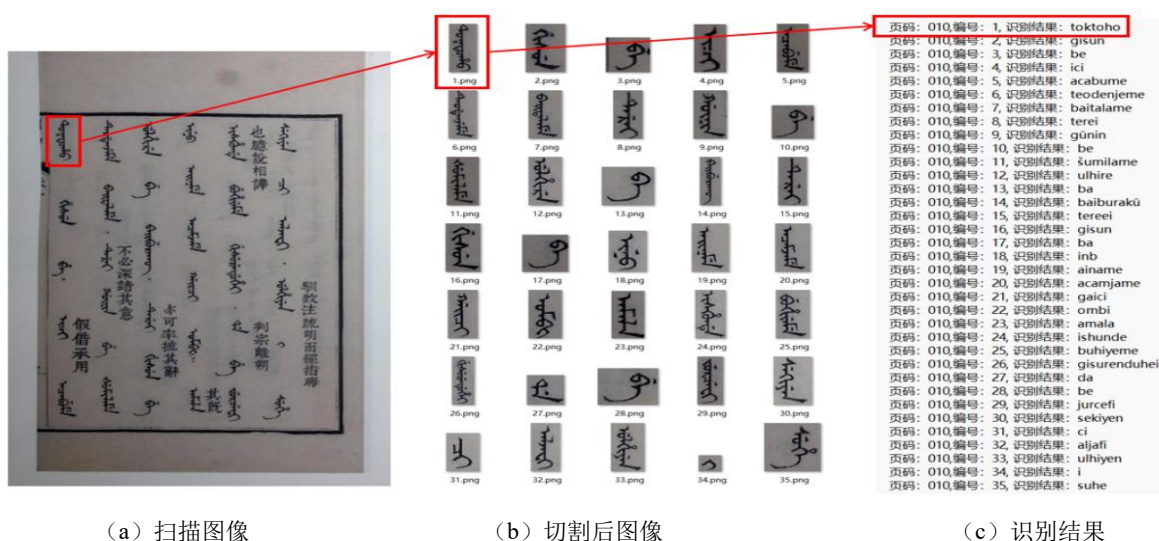


图 1 满文古籍扫描示例图

Figure 1 Example of scanned ancient Manchu books



(a) 扫描图像

(b) 切割后图像

(c) 识别结果

图 2 初步识别结果示例图

Figure 2 Example of preliminary recognition results

上述步骤得到识别结果后, 需要对其进行核验, 并对分割后的满文单词图像进行重命名操作。在核验过程中, 对于识别出的转写字母, 通过查询在线满文词典 (<http://hkuri.cneas.tohoku.ac.jp/p06/manchu/list?groupId=11>) 进行比对, 从而判断满文单词与转写字母是否匹配。根据两者比对结果, 满文图像按照以下规则进行命名, 以“_”作为分隔符。其中, 第一部分表示识别的准确性, 若识别正确, 则命名为“0”, 否则为“1”; 第二部分为页号, 表示该满文单词所在的页数; 第三部分为列号, 指该单词在该页中的列位置; 第四部分为满文单词在该列中的位置编号; 第五部分为该满文单词所对应的转写字母。

例如, 命名为“0_010_1_1_toktoho”的文件表示该满文单词识别正确, 位于第 10 页, 第 1 列, 第 1 个位置, 且其转写字母为“toktoho”。通过这种命名方式, 可以有效地对每个满文单词进行追踪, 并方便进行管理。其中, 该满文单词校验结果示例如图 3 所示。

最后将所有识别正确的满文古籍单词图像进行整理, 并根据转写字母进行分类存放, 最终形成了按转写字母划分的满文古籍单词数据集, 其具体样式如图 4 所示。



图 3 校验结果示例图

Figure 3 Example of Verification results

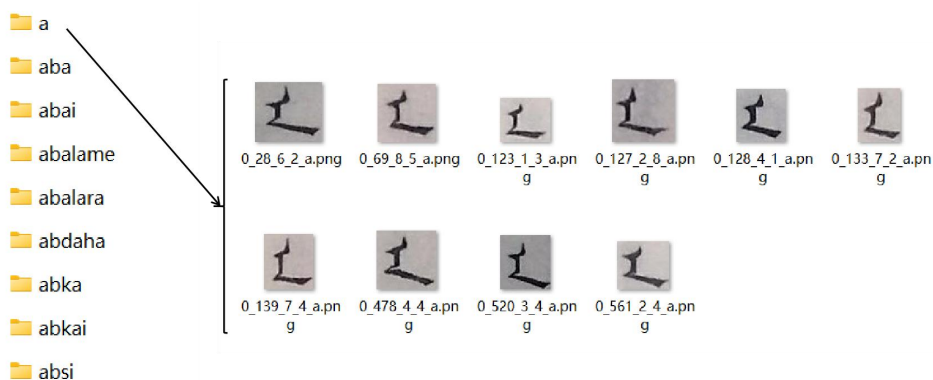


图 4 转写字母为“a”的满文单词文件夹示意图

Figure 4 Diagram of the Manchu word folder with the transliteration starting with the letter "a"

2 数据样本描述

本研究构建的面向 OCR 的满文古籍单词数据集共包含 24280 张高质量图像，覆盖 2428 个不同的满文单词类别。每个单词类别包含 10 张具有代表性的图像样本，数据来源于《国家图书馆藏满汉文合璧古籍珍本丛书》，保证了数据的历史真实性与多样性。数据集旨在支持 OCR 技术的深入研究与应用，提升模型在满文识别中的准确性和泛化能力。

2.1 数据结构与样本说明

(1) 单词类别：数据集中共包含 2428 个独立的满文单词类别，每个类别对应一种独特的字符组合，涵盖了满文的多种书写形式和结构复杂度。

(2) 样本数量：为了增强模型的泛化能力并避免过拟合，每个单词都有 10 张独立的图像样本。这些图像样本的选择考虑了不同的书写风格、字体变化及可能的噪声影响，从而确保了数据集的多样性和代表性。

(3) 图像质量：每张图像均经过高质量扫描或拍摄，保证了清晰度和细节的保留，以便在 OCR

任务中准确识别字符和单词。其中，图像尺寸和分辨率均经过优化，确保可以在各种 OCR 模型中使用。

2.2 典型样例展示

图 5 展示了数据集中“tucirakv”单词类别下的 10 张典型图像文件命名结构及样本（文件格式示例：0_101_1_4_tucirakv.png），反映了样本的多样性和编号规则。

```
Manchu OCR/tucirakv/  
├── 0_101_1_4_tucirakv.png  
├── 0_128_3_7_tucirakv.png  
├── 0_184l_7_8_tucirakv.png  
├── 0_217_3_3_tucirakv.png  
├── 0_217_5_5_tucirakv.png  
├── 0_217_6_5_tucirakv.png  
├── 0_328_7_9_tucirakv.png  
├── 0_330_4_6_tucirakv.png  
├── 0_440_7_9_tucirakv.png  
└── 0_530_4_8_tucirakv.png
```

图 5 转写字母为“tucirakv”的典型样例

Figure 5 Typical example of the transliteration "tucirakv"

图像文件命名遵循固定格式：[文档编号]_[页码]_[列号]_[单词位置]_[单词拼音].png。其中，文档编号表示原始文献的序号，用于区分不同文档；页码表示单词所在页码，支持带字母的页码编号；列号表示所在页的列数，便于横排/竖排版式定位；单词位置表示在当前列中的顺序位置；单词拼音：满文单词的拉丁转写，作为类别标签。

2.3 数据来源与代表性

本数据集主要来自《国家图书馆藏满汉文合璧古籍珍本丛书》，涵盖多个历史时期和不同书写风格，确保数据在历史和书写形式上的多样性。此外，数据集内的图像经过人工筛选和质量控制，确保每张图像均具备清晰度和信息完整性，适合用于高精度 OCR 模型的训练和评估。

3 数据质量控制和评估

在构建本数据集的过程中，通过采用高精度的扫描技术对原始古籍文献进行数字化处理，以确保满文单词图像的清晰度。此过程分为两个阶段，第一阶段为数据预处理，严格规范单词分割和转写的操作流程，以最大程度保证样本的完整性与准确性。第二阶段为人工核验，对初步识别出的结果进行质量筛选，剔除掉较为模糊、破损以及笔画断裂的样本，从而为后续的 OCR 任务提供了可靠的数据支撑。

在质量评估方面，由于本研究所选用的数据来源于特定的古籍书目，特别注重确保数据集样本的均衡性。每个满文单词均有 10 张图像样本，这一均衡的样本分布有效避免了样本数量的不均衡问题，为模型训练提供了稳定的支持。该分布有助于提升 OCR 模型的识别精度和泛化能力，确保模型能够在不同类型的满文单词上取得一致且优异的表现。

4 数据价值

我国作为一个多民族和谐共存的国家，拥有 56 个民族，形成了多元一体的语言体系，展现出丰富而独特的语言资源。其中，满文作为我国重要的民族文字之一，在辽宁、吉林、黑龙江及内蒙古东部的满族聚居地区仍被广泛使用。然而，由于缺乏大规模公开的满文单词标注语料库，一些关键的满文信息处理任务仍处于起步阶段。例如，满文古籍修复、机器翻译以及知识图谱的构建等。鉴于此，本数据集的构建与发布不仅为满文 OCR 任务提供了重要基础，同时也为训练满文单词识别模型提供了数据支撑。本数据的公开发布，将为古籍智能识别、语言处理与清代文献研究提供有力支撑，具有重要的技术与学术双重价值。

5 数据使用方法和建议

本研究所构建的数据集以单词为单位进行划分，图像质量经过严格控制，具有较高的清晰度和规范性，可应用于满文 OCR 任务中。其数据格式采用 JPG 和 PNG 两种，支持主流的图像处理工具和编程语言进行操作与分析。

对于用户而言，若直接下载并使用本研究构建的数据集，将其应用于图像识别任务中，可能会面临一定的挑战。由于满文古籍图像留存年代较为久远，加之文献保存条件差，普遍存在文字残缺、图像模糊以及阴影干扰等问题，导致模型在处理此类低质量样本时，识别效果不佳。为了有效提高识别的准确率，建议使用者结合满文书写特点及上下文语义信息，对视觉和语义特征进行协同建模，以增强模型的鲁棒性。此外，为了确保模型训练的有效性和评估的准确性，建议将数据集按 8:2 的比例划分为训练集和测试集，其中 80% 的数据用于训练，20% 的数据用于测试。这种划分方式有助于保持训练和测试数据的平衡，并有效评估模型在实际应用中的性能。

作者分工职责

孙海鹏（1997—），男，山东人，博士研究生，研究方向为满文单词识别、机器翻译。主要承担工作：数据收集、人工数据处理、论文撰写。

陶文豪（2001—），男，安徽人，硕士研究生，研究方向为满文单词识别。主要承担工作：数据收集、人工数据处理、论文撰写。

毕晓君（1964—），女，黑龙江人，博士，教授，研究方向为自然语言处理、数字图像处理、机器学习及智能优化算法。主要承担工作：研究设计及论文撰写指导。

参考文献

- [1] JIA X. Image recognition method based on deep learning[C]. 2017 29th Chinese Control and Decision Conference (CCDC). Chongqing, China. IEEE, 2017: 4730–4735. DOI:10.1109/CCDC.2017.7979332.
- [2] WANG J Y, LI X C, JIN Y X, et al. Research on image recognition technology based on multimodal deep learning[EB/OL]. (2024-10-15) [2026-02-12]. <https://arxiv.org/abs/2405.03091>.
- [3] 王维兰, 胡金水, 魏宏喜, 等. 少数民族文字文本分析与识别的研究进展[J]. 中国图象图形学报, 2024, 29(06): 1685–1713. [WANG W L, HU J S, WEI H X, et al. Survey on text analysis and recognition

- for multiethnic scripts[J]. Journal of Image and Graphics, 2024, 29(6): 1685–1713. DOI:10.11834/jig.240015.]
- [4] WANG Z W, LU S Y, WEI X, et al. Learn more Manchu words with a new visual-language framework[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024: 3652992. DOI:10.1145/3652992.
- [5] LI H R, LU S Y, WEI X, et al. Probing handwritten Manchu word recognition foundation model[C]. 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). Ocean Flower Island, China. DOI:10.1109/ICPADS60453.2023.00182.
- [6] ZHANG D D, LIU Y, WANG Z W, et al. OCR with the deep CNN model for ligature script-based languages like Manchu[J]. Scientific Programming, 2021, 2021(1): 5520338. DOI:10.1155/2021/5520338.
- [7] 张文亮, 于姗姗. 满文古籍信息组织研究的系统综述[J]. 图书馆学研究, 2024(10): 2–13. DOI:10.15941/j.cnki.issn1001-0424.2024.10.006. [ZHANG W L, YU S S. A systematic review of research on the information organization of Manchu ancient books[J]. Research on Library Science, 2024(10): 2–13. DOI:10.15941/j.cnki.issn1001-0424.2024.10.006.]
- [8] 张玲玲. 新时代满文古籍出版现状及未来发展路径[J]. 临沂大学学报, 2024, 46(3): 143–151. DOI:10.13950/j.cnki.jlu.2024.03.015. [ZHANG L L. The current situation and future development path of publishing Manchu ancient books in the new era[J]. Journal of Linyi University, 2024, 46(3): 143–151. DOI:10.13950/j.cnki.jlu.2024.03.015.]
- [9] 吴智嘉. 民族古籍文献整理与各民族交往交流交融史料开发: 以清代满文档案整理开发为中心[J]. 满族研究, 2022(2): 62–68. DOI:10.3969/j.issn.1006-365X.2022.02.012. [WU Z J. Ethnic ancient books and documents should communicate and blend with the Chinese nation[J]. Manzu Minority Research, 2022(2): 62–68. DOI:10.3969/j.issn.1006-365X.2022.02.012.]
- [10] 李刚. 从传统到现代: 中国第一历史档案馆满文档案编纂工作的回顾与展望[J]. 中国档案, 2022(8): 40–41. [LI G. From tradition to modernity: review and prospect of Manchu archives compilation in China No.1 historical archives[J]. China Archives, 2022(8): 40–41.]
- [11] 吴元丰, 李刚. 中国满文档案工作 70 年: 以中国第一历史档案馆为中心[J]. 满语研究, 2019(2): 81–87. DOI:10.3969/j.issn.1000-7873.2019.02.013. [WU Y F, LI G. 70 years' research on Manchu archives from China[J]. Manchu Studies, 2019(2): 81–87. DOI:10.3969/j.issn.1000-7873.2019.02.013.]
- [12] 张晶, 许爽, 贺建军, 等. 基于缝隙裁剪的满文单词分割和提取方法研究[J]. 中文信息学报, 2019, 33(2): 81–88. DOI:10.3969/j.issn.1003-0077.2019.02.011. [ZHANG J, XU S, HE J J, et al. Segmentation and extraction method for Manchu words based on seam craving[J]. Journal of Chinese Information Processing, 2019, 33(2): 81–88. DOI:10.3969/j.issn.1003-0077.2019.02.011.]
- [13] 许爽, 李敏, 朱满琼, 等. 满文识别技术研究与分析[J]. 大连民族学院学报, 2014, 16(5): 546–551. DOI:10.13744/j.cnki.cn21-1431/g4.2014.05.017. [XU S, LI M, ZHU M Q, et al. Research and analysis of Manchu character recognition technology[J]. Journal of Dalian Nationalities University, 2014, 16(5): 546–551. DOI:10.13744/j.cnki.cn21-1431/g4.2014.05.017.]
- [14] 周兴华, 李敏, 郑蕊蕊, 等. 印刷体满文文字数据库的构建与实现[J]. 大连民族学院学报, 2015, 17(3): 270–273. DOI:10.13744/j.cnki.cn21-1431/g4.2015.03.019. [ZHOU X H, LI M, ZHENG R R, et al. Construction and implementation of printed Manchu character database[J]. Journal of Dalian Nationalities

University, 2015, 17(3): 270–273. DOI:10.13744/j.cnki.cn21-1431/g4.2015.03.019.]

[15] 国家图书馆古籍馆, 全桂花, 朱志美, 等. 国家图书馆藏满汉文合璧古籍珍本丛书 (25 册) [M]. 北京: 学苑出版社, 2018. [National Library of China Ancient Books Library, QUAN G H, ZHU Z M, et al. A series of rare ancient books in Manchu and Chinese in the national library (25 volumes) [M]. Beijing: Xueyuan Press, 2018.]

[16] 刘成林, 金连文, 白翔, 等. 文档智能分析与识别前沿: 回顾与展望[J]. 中国图象图形学报, 2023, 28(8): 2223–2252. DOI:10.11834/jig.221112. [LIU C L, JIN L W, BAI X, et al. Frontiers of intelligent document analysis and recognition: review and prospects[J]. Journal of Image and Graphics, 2023, 28(8): 2223–2252. DOI:10.11834/jig.221112.]

[17] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别研究[J]. 计算机学报, 2020, 43(10): 1943–1957. DOI:10.11897/SP.J.1016.2020.01943. [LUO L, YANG Z H, SONG Y W, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning[J]. Chinese Journal of Computers, 2020, 43(10): 1943–1957. DOI:10.11897/SP.J.1016.2020.01943.]

[18] 金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述[J]. 自动化学报, 2016, 42(8): 1125–1141. DOI:10.16383/j.aas.2016.c150725. [JIN L W, ZHONG Z Y, YANG Z, et al. Applications of deep learning for handwritten Chinese character recognition: a review[J]. Acta Automatica Sinica, 2016, 42(8): 1125–1141. DOI:10.16383/j.aas.2016.c150725.]

[19] TANG M L, XIE S M, LIU X R. Ancient character recognition: a novel image dataset of Shui manuscript characters and classification model[J]. Chinese Journal of Electronics, 2023, 32(1): 64–75. DOI:10.23919/cje.2022.00.077.

[20] LI B, LUO D H, LIANG Y J, et al. Oracle bone inscriptions multi-modal dataset[EB/OL]. (2024-07-04) [2026-02-12]. <https://arxiv.org/abs/2407.03900>.

[21] ZHENG R R, LI M, HE J J, et al. Segmentation-free multi-font printed Manchu word recognition using deep convolutional features and data augmentation[C]. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Beijing, China. IEEE, 2019: 1–6. DOI:10.1109/CISP-BMEI.2018.8633208.

[22] WANG Z W, LU S Y, WANG M Q, et al. AMRE: an Attention-Based CRNN for Manchu Word Recognition on a Woodblock-Printed Dataset[C]. Neural Information Processing. Cham: Springer, 2023: 267–278. DOI:10.1007/978-3-031-30108-7_23.

论文引用格式

孙海鹏, 陶文豪, 毕晓君. 面向 OCR 的满文古籍单词数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-24). DOI: 10.11922/11-6035.csd.2025.0069.zh.

数据引用格式

孙海鹏, 陶文豪, 毕晓君. 面向 OCR 的满文古籍单词数据集[DS/OL]. V2. Science Data Bank, 2026. (2026-02-24). DOI: 10.57760/sciencedb.25676.

A dataset of Manchu ancient book words for OCR

SUN Haipeng^{1,2}, TAO Wenhao², BI Xiaojun^{1,2*}

1. School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, P.R. China

2. Key Laboratory of Ethnic Language Intelligent Analysis and Security Management of MOE, Beijing 100081, P.R. China

*Email: bixiaojun@hrbeu.edu.cn

Abstract: Manchu ancient books carry the historical context and cultural essence of the Qing Dynasty, serving as an indispensable and important heritage in the development of Chinese civilization. Due to the challenges in interpreting these texts, particularly the difficulty in recognizing Manchu words, the study of Manchu word recognition based on deep learning is particularly important. However, deep learning models highly depend on high-quality and large-scale annotated datasets during training. To address this, this study constructs a high-quality and diverse dataset of Manchu ancient book words. Given the limitations of traditional segmentation methods for data collection, such as the high costs and time-consuming nature of manual annotation, this dataset adopts a semi-automated approach, combining computer-assisted technology and manual verification to achieve effective word extraction. On this basis, we further introduced an image quality screening and annotation review mechanism to systematically eliminate blurred, tilted and miscut samples to ensure the clarity of images and the accuracy and consistency of annotations. Based on the *Series of Rare Ancient Books in Manchu and Chinese Housed in the National Library*, this study extracts a total of 24,280 word images from Manchu ancient books, covering 2,428 unique Manchu words, with 10 images corresponding to each word. In the process of data collection and collation, we found that this dataset is the largest collection of Manchu word images so far, providing a solid data foundation for related research.

Keywords: Manchu Literature; Manchu ancient book words; optical character recognition; cultural heritage protection

Dataset Profile

Title	A dataset of Manchu ancient book words for OCR
Data corresponding author	BI Xiaojun (bixiaojun@hrbeu.edu.cn)
Data authors	SUN Haipeng, TAO Wenhao, BI Xiaojun
Time range	1733–1867
Geographical scope	China
Data volume	304.85 MB
Data format	*.jpg, *.png
Data service system	https://doi.org/10.57760/sciencedb.25676
Sources of funding	National Natural Science Foundation of China (62236011); National Social Science Foundation of China (20&ZD279).

Dataset composition	The dataset contains a total of 2,428 Manchu words. Each word is assigned to a subfolder named after its Latin transliteration (e.g., toktoho). Each subfolder contains 10 image samples of the respective word, totaling 24,280 images. The image files follow a standardized naming convention. For example, "0_010_1_1_toktoho" indicates that the image comes from page 10, column 1, position 1, and the corresponding word is "toktoho" (its Latin transliteration).
----------------------------	---