

文章编号: 1004-4353 (2024) 02-00101-06

基于中朝统一 IDS 编码的朝鲜语 古籍文字识别方法

赵梦玲, 金小峰

(延边大学 融合学院, 吉林 延吉 133002)

摘要: 为解决朝鲜语古籍中的中文和朝鲜文字混排的识别难题, 提出一种中朝文字的表意文字描述序列 (IDS) 统一编码方案, 旨在通过利用偏旁分解字符识别模型 (CCR-CLIP) 识别朝鲜语古籍文字。首先, 根据中朝文字结构的相似性, 对文字中出现的汉字偏旁、朝鲜文字字母和 12 种基本结构进行了统一编码; 其次, 通过加入朝鲜文字的 IDS 序列扩充了 CCR-CLIP 原模型中提供的汉字的 IDS 序列文件; 最后, 通过在训练阶段使用印刷体文字训练的方式解决了朝鲜语古籍样本少的问题。

关键词: 朝鲜语古籍; 零样本; 文字识别; 文字编码; 表意文字描述序列

中图分类号: TP391.1 **文献标志码:** A

Korean ancient books character recognition method based on unified Chinese and Korean characters ideographic description sequences coding

ZHAO Mengling, JIN Xiaofeng

(College of Integration Science, Yanbian University, Yanji 133002, China)

Abstract: In order to solve the problem of recognition of mixed Chinese and Korean characters in ancient Korean books, this paper proposes a unified ideographic description sequence (IDS) encoding scheme for Chinese and Korean characters, which aims to recognize ancient Korean books by using a side decomposition chinese character recognition-contrastive language-image pre-training) (CCR-CLIP). Firstly, according to the similarity of Chinese and Korean characters, the Chinese characters' side edges, Korean characters' letters and 12 kinds of basic structures are uniformly coded. Secondly, the IDS sequence file of Chinese characters provided in the original model of CCR-CLIP is extended by adding IDS sequence of Korean characters. Finally, the problem of few samples of Korean ancient books was solved by using printed characters in the training stage. The results show that compared with the CCR-SLD method, the character recognition accuracy of this method is improved by 13.8% in the experiment of Korean ancient books. In the printed text experiment, the accuracy of character recognition improved by 5.38%. The established method is better than other methods in solving the problem of Korean ancient text recognition, and can provide reference for solving the problem of Korean ancient text recognition.

Key words: Korean ancient books; zero-shot; character recognition; character coding; ideographic description sequences

收稿日期: 2023-12-19

基金项目: 吉林省教育厅人文社科基础研究项目 (JJKH20230608SK)

第一作者: 赵梦玲 (1998—) 女, 硕士研究生, 研究方向为文字识别。

通信作者: 金小峰 (1970—) 男, 教授, 研究方向为语音信息处理、计算机视觉。

0 引言

朝鲜语古籍具有多文种混排的特点,尤其是中文与朝鲜语混排的古籍较多.目前,大多数的文字识别模型只能识别一种文字,且缺乏适用于多语种通用的文字识别方法.近年来,针对汉字识别(Chinese character recognition, CCR)的方法研究主要集中在字符级、偏旁级和笔画级^[1-8]3个层次上.由于汉字的类别和结构较为复杂,在实际应用中面临着零样本的问题;因此,现有的汉字识别方法常依靠预测偏旁或笔画序列来实现字符的识别.

由于朝鲜文字是由有限的字母按特定的组字规则构造的文字,且朝鲜语中的字母类似于汉字中的偏旁部首,因此也可以按照笔画对其进行分解.2022年, Kim等^[9]提出了一种将朝鲜语字符分解为字母后再通过识别字母进而识别朝鲜语字符的方法.2023年, Yeongseo等^[10]进一步证明了可以将朝鲜语字符分解为字母后进行识别.基于上述研究,本文针对朝鲜语古籍中的中朝2种文字混排现状,利用中朝2种文字的表意描述序列(ideographic description sequences, IDS),进行统一编码,采用偏旁分解字符识别模型(CCR-CLIP)^[4],对朝鲜语古籍中的中朝2种文字识别方法进行研究,提出了朝鲜语古籍的文字识别方法.

1 中朝文字相关背景知识

根据 GB18030—2005《信息技术中文编码字符集》,汉字共有 70 244 个类别.其中,一级常用汉字有 3755 个.每个汉字都可以按照特定的顺序将其分解成相应的偏旁序列.一级常用汉字由 514 个偏旁和 12 个基本结构所组成,如图 1 所示.

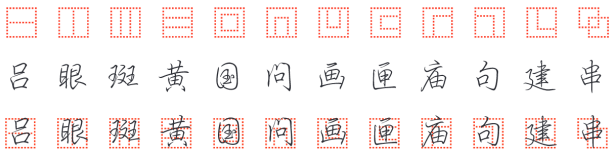


图 1 12 种汉字结构及示例

按图 1 所示的汉字分解方式,每个汉字可以用树表示,如图 2 所示.由遍历树可以得到相应的表意文字描述序列(IDS).

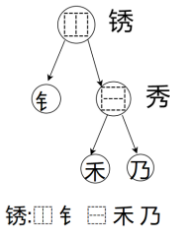


图 2 汉字“锈”字的分解过程及 IDS 序列

朝鲜文字和汉字虽然都属于方块文字,但朝鲜文字由 2 个或 3 个字母构成.其中,元音字母有 21 个,辅音字母有 19 个,收音字母有 27 个,共可以构造出 11 172 个字符,且这些字母大多数与汉字的偏旁不同.虽然 11 172 个字符都已经在 Unicode 中被编码,但其中的大多数字符在现代朝鲜语中并不常用.当前,现代朝鲜语中常用的字符集仅为 2350 个字符.由于朝鲜语文字与汉字结构类似,因此同样可以生成相应的序列.图 3 为朝鲜语文字“넉”的分解过程及其 IDS 序列的生成.

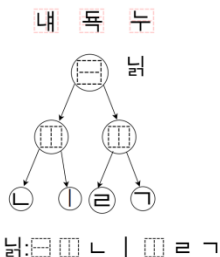


图3 朝鲜语中的3个结构, 字符“ ㄴ ”的分解过程及IDS序列

2 朝鲜语古籍文字识别模型

为识别朝鲜语古籍字符, 本文将朝鲜语中的字母与汉字中的偏旁视作同等层次, 即对汉字偏旁和朝鲜语字母进行统一编码, 使模型可以同时识别 2 种字符。同时, 为让模型在训练和测试阶段可以对 2 种字符进行训练和测试, 本研究对支持集样本进行了扩充, 即加入了朝鲜语字符分解后的偏旁序列, 使模型适合于识别朝鲜语古籍字符。本文在采用 CCR-CLIP 模型对朝鲜语古籍字符进行识别时, 主要是通过识别偏旁表意描述序列来识别字符。该模型在训练阶段学习中朝字符的规范表示, 在测试阶段通过对比得到的中朝字符的图像特征和文本特征来选择相似度最高的字符, 并将其作为最终的预测字符。

2.1 中朝文字 IDS 统一编码方案

汉字的 IDS 中有 12 种基本结构, 其中适合朝鲜语的只有 3 种: 左右结构、上中下结构、上下结构, 如图 3 中最上方的 3 个朝鲜语的基本结构. 在进行朝鲜语字符分解时, 为减少基本字母的个数, 本文对双收音“ㄱ, ㅋ, ㆁ, ㄷ, ㅌ, ㄴ, ㄷ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ”进行了分解, 即将其分解为“ㄱ, ㅋ, ㆁ, ㄷ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ”进行了分解, 即将其分解为“ㄱ, ㅋ, ㆁ, ㄷ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ, ㅌ, ㄴ”. 为使朝鲜语中出现的结构可以和汉字中的结构相同, 本文对“ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅟ, ㅢ”进行了分解, 即将其分解为“ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅟ, ㅢ”. 按上述方法分解后, 最终得到 33 个朝鲜语的基本字母和 3 种基本文字结构. 由于 IDS 序列中不止有中文偏旁和朝鲜语字母, 还有 12 种基本文字结构, 因此在最终的中朝文字 IDS 统一编码中, 除了包含中文偏旁、朝鲜语字母, 还包括了 12 种基本结构. 其中, 每个偏旁、每个字母、每个结构对应一个唯一的编码. 由此所得的最终的中朝文字 IDS 的统一编码方案如表 1 所示, 其中ㄱ、ㅋ是在 Unicode 编码中, 中文和朝鲜语共有的编码相同的 2 个字母. 本文将它们进行了统一编码, 即将ㄱ编码为 526, 将ㅋ编码为 517.

表 1 中朝文字 IDS 统一编码表

编码	1	2	3	4	5	6	7	8	…
偏旁 / 结构	𠂇	讠	𠂇	丩	口	儿	亻	内	…
编码	513	514	515	516	…	541	542	543	544
偏旁 / 结构	亻	亠	丁	一	…	匕	𠂇	从	双

2.2 基于 CCR-CLIP 的朝鲜语古籍文字识别模型

图4为本文提出的朝鲜语古籍文字识别模型.在训练阶段,CCR-CLIP模型由1个图像编码器和1个文本编码器组成.其中:图像编码器负责提取输入的中朝字符图像的视觉特征;文本编码器则负责提取中朝字符的偏旁序列特征,即通过对比损失对模型进行监督,以确保模型在学习过程中能够有效地将中朝字符图像的视觉特征与中朝字符偏旁序列的文本特征相匹配.在测试阶段,模型利用图像编码器来获取输入的中朝字符图像的视觉特征,然后再对比提取的中朝字符图像的视觉特征与中朝字符的文本特征来选取相似度最高的字符,并将其作为模型最终的预测字符.

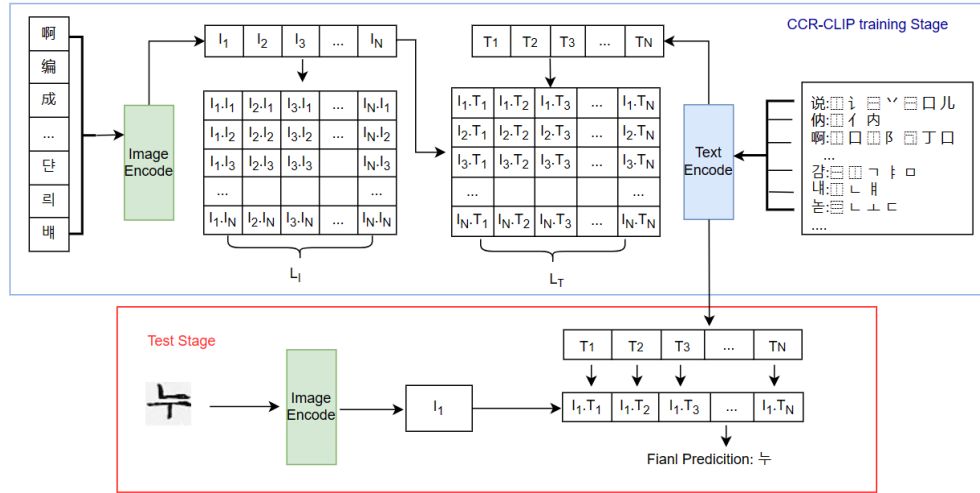


图 4 CCR-CLIP 模型总体结构

2.2.1 Image Encoder

Image Encoder 采用 ResNet-50 模型实现. ResNet 是一种深度神经网络架构, 目前被广泛用于计算机视觉任务中, 特别是图像分类领域. 在 CCR-CLIP 模型中, 本文采用 ResNet-50 从输入的图像中提取特征图 $F^C \in \mathbb{R}^{H/8 \times W/8 \times C}$, 其中 H 为特征图的高度, W 为特征图的宽度, C 为通道数. 为了用一维矢量表示输入图像, 本文采用全局平均池化压缩特征图 F^C , 即 $f^C = \text{GlobalAvgPool}(F^C)$, 其中 f^C 为压缩后的特征向量.

将 f^C 投影到视觉特征空间中, 即可得到输入图像的视觉特征, 其可表示为 $I = f^C W^C$, 其中 I 表示输入字符图像的视觉特征, W^C 表示投影矩阵.

2.2.2 Text Encoder

文本编码器由 K 层 transformer 编码器和一个嵌入层组成. 通过 transformer 编码器, 可将偏旁序列 $D = \{r_1, r_2, \dots, r_L\}$ (其中: r_i 表示偏旁序列中的第 i 个偏旁, L 表示偏旁序列的长度, r_L 表示结束标记) 标记为 $F^r = \{f_1^r, f_2^r, \dots, f_L^r\}$. (其中: $f_i^r \in \mathbb{R}^{1 \times D}$ 为 D 的全部特征), 然后再将 f_i^r 投影到 T 中即可得到文本特征. 该过程可表示为 $T = f_i^r W^r$, 其中 T 为文本特征, W^r 为投影矩阵.

2.2.3 损失函数

本文使用对比损失 L_T 对提取的视觉特征与其相应的偏旁序列特征进行校准. 具有 N 个字符样本的损失函数 L_T 其计算公式为 $L_T = -\sum_{j=1}^N \log \frac{\exp(I_j \cdot T_j)}{\sum_{n=1}^N \exp(I_j \cdot T_n)} - \sum_{j=1}^N \log \frac{\exp(I_j \cdot T_j)}{\sum_{n=1}^N \exp(I_n \cdot T_j)}$, 其中 I_j 和 T_j 分别表示数据中第 j 个样本的视觉特征和偏旁序列特征.

为了减少不同字体样式和相似字符而造成的预测误差, 本文还采用了具有相同标签的输入图像视觉特征之间的对比损失函数 L_I . 给定数据 $B = \{(C_1, D_1), \dots, (C_N, D_N)\}$, 其中: C_i 和 D_i 分别表示第 i 个字符图像及其对应的偏旁序列, 然后再利用图像编码器将第 i 个字符图像 C_i 编码为相应的视觉特征 I_i . 因此, 计算损失

函数 L_I 的公式可表示为 $L_I = -\sum_{j=1}^N \log \frac{\sum_{I' \in u_j} \exp(I_j \cdot I')}{\sum_{n=1}^N \exp(I_j \cdot I_n)}$, 其中 u_j 为具有相同偏旁序列 D_j 的视觉特征的集合, I' 为具有相同偏旁序列 D_j 的其中一个字符图像的视觉特征.

基于上述方法, CCR-CLIP 模型的整体损失函数可表示为 $L_{\text{pre}} = L_T + L_I$.

3 实验结果与分析

3.1 数据集

为验证所提出的方法在识别朝鲜语古籍字符的有效性, 本文针对朝鲜语古籍数据集^[11]进行对比实验. 朝鲜语古籍数据集来源于《同文类解》《阐义昭鉴谚解》和《谚解胎产集》等 3 个文本图像数据集, 合计包含 875 张图片, 如图 5 所示. 其中: 图 (a) 555 张, 图 (b) 160 张, 图 (c) 160 张. 对朝鲜语古籍文本进行切分和进行标注后, 共得到 4100 类已标记的朝鲜语古籍文字图像.

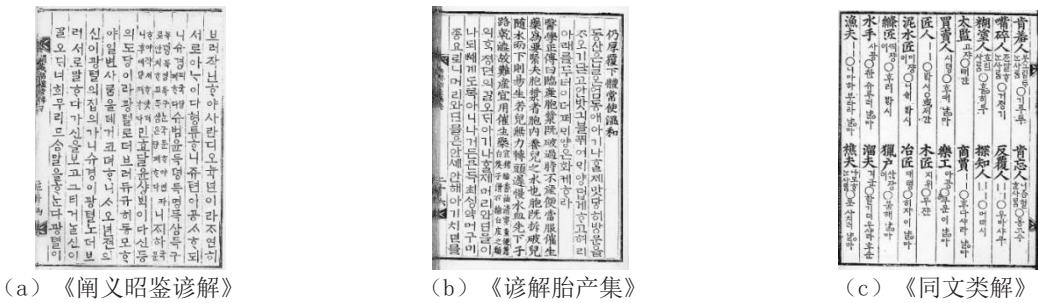


图 5 朝鲜语古籍文本图像示例

为缓解样本少的问题, 本文根据 Unicode 编码生成了常用 3755 类一级汉字文字图像集和 2350 类常用朝鲜语文字图像集. 在生成的 2 种文字图像集中, 字体选用与手写体类似的字体形式. 其中, 汉字文字图像涉及宋体、楷体、隶书等 3 种字体形式, 朝鲜语文字图像涉及 Batang 和 gungsuh2 种字体形式. 合并常用的一级汉字文字图像集和常用朝鲜文字图像集后, 即可得到中朝印刷体文字图像数据集. 该数据集共计包含 6105 类文字图像集, 每类含有 150 张图片.

3.2 实验结果

本文在朝鲜语古籍数据集进行零样本文字图像识别, 由于朝鲜语古籍数据集中存在中朝类别混杂现象 (朝鲜语与汉语的字符类别比例约为 2:3), 因此在实验中抽取的训练集和测试集, 也按照该比例进行采样. 依此, 首先在 4100 类标注数据集中随机选取了 1000 类作为测试样本, 然后再从剩余 3100 类中按照 (500, 1000, 1500, 2000, 3100) 规模任意抽取样本创建训练集. 本文采用字符识别的准确率作为评价模型性能的指标. 采用朝鲜语古籍数据集进行 0 样本字符识别的实验结果如表 2 所示.

表 2 不同模型在朝鲜语古籍数据集上的识别准确率

方法	不同规模的训练集				
	500	1000	1500	2000	3100
CCR-SLD ^[6]	4.83%	6.75%	8.95%	11.07%	13.38%
CCR-SLD+CA ^[11]	5.47%	7.83%	9.45%	13.28%	15.96%
本文方法	6.31% (↑0.84)	12.39% (↑4.56)	17.72% (↑8.27)	21.62% (↑8.34)	29.76% (↑13.8)

由表 2 可以看出, 本文方法在字符零样本的设置上明显优于其他 2 种方法, 并且字符识别的准确率随着训练样本的增多而明显增加, 且均高于其他 2 种方法. 本文方法由于采用了 IDS 和字符图像对比的架构, 因此其可有效规避 CCR-SLD 和 CCR-SLD+CA 模型既有的在笔画层对朝鲜语古籍字符进行分解可能导致的笔画预测错误. 而 CCR-CLIP 模型由于能够识别偏旁部首, 因此其可有效规避 CCR-SLD 和 CCR-SLD+CA 模型在笔画预测方面的误差.

为解决朝鲜语古籍样本少的问题, 本文还利用中朝印刷体文字图像数据集进行了字符识别实验. 测试

集由在朝鲜语古籍数据集中任意选取的 1000 类样本组成,训练集采用中朝印刷体文字图像数据集,规模分别为 (1200, 2400) . 中朝印刷体文字图像数据集为训练集、朝鲜语古籍为测试集时的字符识别准确率,见表 3. 由表 3 可以看出,采用中朝印刷体数据集进行训练时,该方法能够增加训练集的种类和数目,即可为模型提供更多的样本进行学习,因此可提高模型的识别效果. 这表明,使用中朝印刷体数据集不仅可以解决数据集不足的问题,还可以解决类别不均衡的问题.

表 3 中朝印刷体文字图像数据集为训练集时的朝鲜语古籍字符识别准确

方法	不同规模的训练集	
	1200	2400
CCR-SLD ^[6]	5.97%	11.25%
CCR-SLD+CA ^[11]	7.34%	13.76%
本文方法	10.31% (↑2.97)	19.14% (↑5.38)

4 结语

实验表明,本文提出的基于 CCR-CLIP 的朝鲜语古籍文字识别模型,不但实现了中朝字符 IDS 序列的统一编码,而且可有效提高对朝鲜语古籍文字的识别效果. 研究结果可为朝鲜语古籍文献的数字化保存和研究提供参考. 基于研究的样本有限,在后续研究中,我们将进一步优化 CCR-CLIP 模型,以验证研究结果的科学性,并提升本文方法的普适性.

参考文献:

[1] HUANG G, LUO X, WANG S, et al. Hippocampus-heuristic character recognition network for zero-shot learning in Chinese character recognition[J]. Pattern Recognition, 2022, 130: 108818.

[2] CHEN Z, YANG W, LI X. Stroke-Based Autoencoders: Self-Supervised Learners for Efficient Zero-Shot Chinese Character Recognition[J]. arXiv preprint arXiv:2207.08191 2022. <https://arxiv.org/abs/2207.08191>.

[3] GAN J, CHEN Y, HU B, et al. Characters as graphs: Interpretable handwritten Chinese character recognition via Pyramid Graph Transformer[J]. Pattern Recognition, 2023, 137: 109317.

[4] YU H, WANG X, LI B, et al. Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE/CVF 2023 11943-11952.

[5] LUO G F, WANG D H, DU X, et al. Self-information of radicals: A new clue for zero-shot Chinese character recognition[J]. Pattern Recognition, 2023, 140: 109598.

[6] CHEN J, LI B, XUE X. Zero-shot Chinese character recognition with stroke-level decomposition[J]. arXiv preprint arXiv: 2106.11613 2021. <https://arxiv.org/abs/2106.11613>.

[7] ZU X, YU H, LI B, et al. Chinese Character Recognition with Augmented Character Profile Matching[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisbon ACM 2022 6094-6102.

[8] LI M, YU Y, YANG Y, et al. Stroke extraction of chinese character based on deep structure deformable image registration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(1): 1360-1367.

[9] KIM G, SON J, LEE K, et al. Character decomposition to resolve class imbalance problem in Hangul OCR[J]. arXiv preprint arXiv: 2208.06079 2022. <https://arxiv.org/abs/2208.06079>.

[10] 하영서, 황호석, 김민준, 等. 객체 검출에서 클래스 압축 및 분할을사용한 중세 한글 인식 인식률 향상 [J]. Journal of Korea Multimedia Society, 2023; 26(6) 795-803.

[11] 刘晓童. 基于笔画分解的朝鲜语古籍字符识别方法研究与应用 [D]. 延吉: 延边大学, 2023.