国际音标字符识别及其应用研究*

龙从军　郭承禹**

[提要] 国际音标是一种运用范围广泛的语音记录和书写符号系统。国际音标光学字符识别技术可实现国际音标的批量识别，将极大地提高国际音标的处理效率，对民族语言和汉语方言资源数字化及基础研究都极具价值。本文推介一款由作者研发的国际音标识别模型“IPAOCR-IEA”。该模型在 63 万条国际音标图文对照数据的基础上利用卷积神经网络模型训练而成，可以高效识别国际音标。其整词精度、字符查准率和查全率等多项指标的最好结果均高于 98%。此外，该模型轻量化的“龙水国际音标识别软件”现已推出，用户可以用其截取国际音标图片并自动获取截图中的国际音标字符。

[关键词] 国际音标　光学字符识别　模型　国际音标自动识别

国际音标是一种语音记录和书写符号系统，由国际语音学会设计并不断完善，其字符可用于标示人类语言的语音单位，包括音素（元音和辅音）及声调等。国际音标符号大多来自拉丁字母，同时也借用了少量希腊字母，还有一些为了准确表示特定语音而设计的符号。国际音标基本遵循“一音一符，一符一音”的原则，具有高度的准确性和系统性，被广泛应用于词典标音、语音工程、言语治疗、语言调查及其他适用场合。

国际音标在我国学术界使用广泛。中国语言学界使用国际音标记录了大量的民族语言和汉语方言材料，这些材料对学术研究、文化传承都有十分重要的价值。随着时代发展，大数据的应用领域不断扩展。迄今为止，电子化资源处理领域内仍缺乏对国际音标数据的采集、识别和规范整理，这一需求尚未解决。为此，国际音标光学字符识别（Optical Character Recognition，简称 OCR）的项目开发十分必要。国际音标 OCR 技术能满足国际音标数据的大规模资源采集，有利于中国语言音标资源数据库的建设，能够有效促进语言文化理论研究。本文推介一个基于神经网络模型的国际音标 OCR 系统，并展示其研发过程及测试结果。

一　国际音标 OCR 技术的应用现状

目前，大多数通用文字的 OCR 技术比较成熟，如汉王科技公司的“汉王 OCR”系统对汉字的识别率高达 99%，并支持多种字体及简繁混识，特别是中英文混排识别技术和表格识

* 本文得到中国社会科学院实验室孵化专项项目“基于民族语言多模态数据的共性特征计算研究（2024SYFH008）”的资助。感谢江荻研究员和李大勤教授对本文所提出的修改意见。感谢《民族语文》匿名审稿人详尽的意见与建议，对文章修改大有裨益。文中不当之处，概由作者负责。

** 通讯作者，北京师范大学文理学院，E-mail：guochengyu@bnu.edu.cn。

别技术世界领先[①]。我国民族文字的识别工作肇始于 20 世纪中后期。由清华大学电子工程系智能图文信息处理研究室主导开发的 TH-OCR 2007“统一平台民族文字文档识别系统”集民族文字识别之大成，使多种民族文字识别达到实用化水平（钱丽花 2007）。江荻（2012:7）对藏文 OCR 工作的进展作了介绍。目前，我国一些民族文字 OCR 的测试结果也达到了实用水平，如 2017 年内蒙古大学推出的“奥云蒙古文印刷体文字识别（OCR）系统”[②]，以及捷通华声公司 2018 年推出的维吾尔文、藏文、蒙古文、彝文、哈萨克文、朝鲜文识别系统等[③]。国外也有大量 OCR 软件，例如 ABBYY FineReader 和 Google Docs 等商业软件已能基本完成对不同文字的识别工作（Tafti et al. 2016）。但是，市场上始终没有出现一款针对国际音标识别的软件。其背后原因可能是潜在用户较少，使用领域受限等。实际上，在语言教学与研究、词典编纂、声学工程、言语治疗、文化记录等特定领域，国际音标有着不可替代的作用。不仅如此，当代大数据智能工程正不断完善，其中不应该忽视特定领域的资源和知识。由此可以判定，国际音标 OCR 技术是有真实需求的，值得开发。

一套 OCR 系统主要有以下几个性能衡量标准：精度（Accuracy）、查准率（Precision）、查全率（Recall），以及 F1 度量（周志华 2016:30）。具体而言，精度是指正确识别的样本数占样本总数的比例，精度越高，模型识别整体效果越好。查准率是指在识别的正例中，有多少是准确的正例，可表示为 [TP/(TP+FP)][④]；查全率是指准确识别的正例在实际正例中的占比，也即 [TP/(TP+FN)]。这两者通常存在相互制约的关系，查准率越高，查全率则可能越低，反之亦然。F1 度量则用于衡量查准率与查全率的相对重要性。当 F1 值小于 1 时，表示查准率影响更大；F1 值大于 1 时，则是查全率更有影响。总之，应该采取多种指标对模型进行整体评估，这样才能多维评价一套 OCR 系统的优劣。

目前，已有一些研究者开展过国际音标 OCR 工作，并取得了一定的成果。邱立松（2015:50）收集了印刷体和手写体两种国际音标，两者都基于 400 的训练样本与 100 的测试样本进行识别，印刷体识别精度达到了 95%，手写体识别精度也达到了 73%。该项研究取得了良好的开端，特别是对于手写体音标识别的尝试颇有意义，可能直接有助于语言调查等应用场景。不过，此项研究并未公布查准率与查全率，且实验样本量也比较小。此外，Li & Hill（2023）以《土家语简志》（田德生等 1986）为材料，利用 Transkribus 这一基于人工智能的文本识别平台，对民族语言印刷体文献进行识别，并使用字符错误率（Character error rate，简称 CER）对结果进行评估。从 OCR 结果来看，4000 多个字符（包括汉字与音标）的识别耗时 40 秒，CER 为 8.65%。换言之，文本识别的精度为 91.35%。尽管如此，此研究无法证明这一文本识别平台能否适用于土家语之外的其他语言，因而语料范围仍有局限性。

尽管现有研究取得了一定的成果，但仍存在以下不足：其一，训练和测试的数据数量偏少；其二，对 OCR 结果的评估还不够完善，缺乏对查准率、查全率及 F1 度量的评估；其三，未达到大规模使用且易用的水平。目前，学界未见广泛使用以及定期更新的音标识别软件。

① https://m.zol.com.cn/article/276660.html。

② http://ocr.mglip.com。

③ https://www.sinovoice.com/news/610.html。

④ TP 为 true positives（正样本预测为正样本的数量）的缩写，FP 为 false positives（负样本预测为正样本的数量）的缩写，FN 为 false negatives（正样本预测为负样本的数量）的缩写。

二　国际音标字符特点及识别难点

早在国际语音学会成立之初（1888 年），其理事会提出并修订了《国际音标使用原则》（简称《原则》），其中第一条规定“每一个不同语音都应该有独立的符号”（国际语音学会 2020:278）。这一规定有着深远的影响。在 1949 年发表的《原则》（详见 International Phonetic Association 1949）中，该规定被重新表述为“同一语言中的两个语音被用来区分两个词的时候，应尽可能用不带附加符号的两个不同符号表示”（国际语音学会 2020:232）。此后，1989 年的《原则》仍然沿用了该项规定（详见 Roach 1989）。

由于常规拉丁字母无法覆盖世界语言复杂的语音单位，因而在其基础上创制了新的符号，包括利用小型大写字母、字符对称翻转、字符延伸或加符、字符组合及附加符号等[①]。小型大写字母如小舌颤音[ʀ]。字符对称翻转包括左右翻转（如声门音[ʔ]与咽部浊擦音[ʕ]）、上下翻转（如央半高元音[ɘ]与央元音[ə]）、中心对称（如后低不圆唇元音[ɑ]与后低圆唇元音[ɒ]）。

字符延伸是指在原有拉丁字母的基础上向上或向下延伸一段带有“弯钩”的符号。顶部向上延伸的国际音标符号如 1a 所示，包含唇齿闪音、声门浊擦音、内爆音等。底部向下延伸的国际音标符号如 1b 所示，包括唇齿鼻音、卷舌音、硬腭音等。1c 则为拉丁字母同时向顶部和底部两端延伸的音标符号，如龈后清擦音、小舌清擦音等。舌尖前元音 [ɻ/ʯ]（非圆唇/圆唇）与舌尖后元音 [ɻ/ʯ]（非圆唇/圆唇）[②]也属于字符延伸符号。此外，字符加符一般是在字符中间添加横线、曲线或斜线，具体如 1d 所示。

1a. 顶部延伸：[ⱱ̟、ɦ̈、ɓ̈、ɗ̈、ɠ̈、ɠ̊]、“ɻ̟、ʯ̟”

1b. 底部延伸：[ɱ̠、ʈ̠、ɖ̠、ɳ̠、ɽ̠、ʂ̠、ʐ̠、ɻ̠、ɭ̠、ɲ̠、ç̠、ʝ̠、ɕ̠、ʑ̠、ŋ̠、ɯ̠]

1c. 两端延伸：[ʃ̴、χ̴、ɧ̴]、“ɻ̴、ʯ̴”

1d. 字符加符：[ɬ̈、ʝ̈、ħ̈、ɬ̈、ʃ̈、ʔ̈、ʕ̈、ʉ̈、ø̈、ɵ̈]

字符组合指音标由两个辅音符号组合而成[③]，在必要时可以用连音符[◌͡◌]标示，如汉藏语言中的塞擦音 [pf、bv、tθ、dð、ts、dz、tʂ、dʐ、tʃ、dʒ、tɕ、dʑ] 等。这种塞擦音在近年的国际音标表中都是以组合的形式出现（燕海雄、江荻 2007），一般不会对识别造成太大影响。

附加符号（diacritic）专指较小的字母或标记，这些符号使得原音标的意义发生变化或更加精确。添加附加符号后的国际音标仍可以视为单一符号。以出现位置分类，附加符号有上、中、下、左、右共五类。上加符号如 2a 所示，诸如央化、中央化、鼻化符号等；中加符号包括卷舌符号 [◌˞] 和软腭化或咽音化符号 [◌ˠ]。下加符号诸如清化和气声符号等，如 2b 所示。

相比上下纵向的附加符号，左右横向的附加符号与原音标的整体结合度并不紧密。左加符号如 2c 所示，大多是超音段符号；汉藏语言常见的声调调类标记符号也在其中，包括轻声“˙◌”阴平“˩◌”等等。右加符号除包括诸如外挤气、送气、唇化等符号，也包括声调符号。这些声调符号涵盖国际音标表中的声调符号（如 [◌˥]）、传统的调类标记符号（如阴入“◌˳”），

① 除国际音标外（本文以“[ ]”表示），国内语言学界创制了其他音标进行记音。由于这些音标尚未在国际音标表上出现，本文暂且用双引号（“”）表示。

② 高本汉（[1940]2003:33）借用瑞典方言字母标写汉藏语言中的舌尖元音，详见“音标对照说明”。

③ 此处的“组合”是指两个音标符号一左一右放置在一起。

以及“五度标记法”的标调数字，具体参见 2d。

2a. 上加符号：[◌̈、◌̽、◌̃、◌̆、◌̋、◌́、◌̄、◌̏、◌̀、◌̌、◌̂]

2b. 下加符号：[◌̥、◌̬、◌̹、◌̜、◌̟、◌̠、◌̩、◌̯、◌̤、◌̰、◌̼、◌̝、◌̞、◌̘、◌̙、◌̺、◌̪、◌̻]

2c. 左加符号：[ꜛ◌、ꜜ◌、↘◌、↗◌、ˌ◌、ˈ◌]、“·◌、꜀◌、꜁◌、꜂◌、꜃◌”

2d. 右加符号：[◌’、◌ʰ、◌ʷ、◌ʲ、◌ˠ、◌ˤ、◌ⁿ、◌ˡ、◌˞、◌ː、◌ˑ、◌|(◌)、◌‖(◌)、◌.(◌)、◌˥、◌˦、◌˧、◌˨、◌˩、◌˩˥、◌˥˩、◌˦˥˦、◌˨˩˨、◌˧˥˧]、“◌꜄、◌꜅、◌꜆、◌꜇、◌^{33}”

在上述符号中，对 OCR 技术最具挑战性的是上加与下加符号，原因有二：其一是附加符号本身较小；其二是不同附加符号之间的差异并不大。此外，字符延伸形式也容易识别为其本体字符。如何提高音标符号 OCR 的准确性，也是本研究面临的难点之一。

三　国际音标 OCR 模型的研发过程及测试结果

（一）图文对齐数据集

数据集构建是国际音标识别的首要工作。本文所用数据主要是多个民族语言和方言词汇数据，累计制作了词汇级图文对照数据达 633406 个，其来源为《汉藏语语音和词汇》（简称 HZ，217023 个）（孙宏开等 2016）、《藏缅语语音和词汇》（简称 ZM，48765 个）（《藏缅语语音和词汇》编写组 1991）、《广西民族语言方音词汇》（简称 GX，100387 个）（广西壮族自治区少数民族语言文字工作委员会 2008），以及其他来源数据（简称 QT，267231 个）[①]。我们通过对原书扫描获得图片，利用图片处理技术把每个词语提取出来，存储为一个单独的图片文件，然后把词语文本与图片一一对齐，转换成模型所需要的数据格式，如图 1 所示。

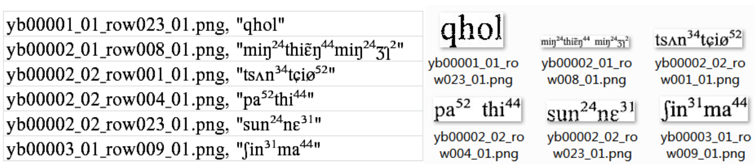


图 1　图文对齐数据集构建示例

（二）图片数据预处理

图片预处理的主要目的包括消除图像中的无关信息，恢复有用的真实信息，增强有关信息的可检测性，以及最大限度地简化数据等等。预处理过程一般包含数字化、几何变换、归一化、平滑、复原和增强等步骤（常晨霞、凌永发 2013）。本研究的预处理工作包括图片增强、图片二值化，以及图片边缘检测和剪切三个步骤。

图片增强是提高图像质量和改善图像视觉效果的一种方法。它通过调整图像的亮度、对比度、颜色等属性，使图像更清晰、更具吸引力。本研究所用图片主要由 PDF 转换而来。由于图片质量不高，我们调用了 enhance_pixels 函数，对其进行像素点加强处理。

图片二值化是将图像像素值限定为两个特定值的处理方法。在二值化后的图像中，灰度值仅为 0 或 255，分别表示黑色和白色。这种处理方法可以将图像颜色简化为黑白两色，有

[①] 数据来自于对“中国少数民族语言简志丛书”“新发现语言系列丛书”和“中国少数民族语言参考语法研究系列丛书”原书扫描后获取的图片文件。

利于后续的图像分析和处理。在处理二值化时，我们利用自定义的 binarize_image 函数和 enhance_pixels 函数，以及 Pillow 库的功能来实现图像二值化和图像增强。关于二值化阈值，我们根据不同来源的文本原图进行调整，同时移除图片颜色配置文件信息。

图片边缘检测和剪切是指把图中（与有效信息无关的）空白部分裁剪掉的方法。首先，要进行有效信息的检测，计算图像中每行和每列的像素值之和，以确定图像的有效区域。然后，再把多余的空白切除掉。为了保证在剪切时不损坏有效信息，我们根据图片的来源和质量、大小等情况，手动调整剪切边缘，预留空白阈值，一般是 2 至 5 个像素值。

（三）算法选择

常用的字符识别算法经历了从基于模板匹配、基于统计特征的分类算法（如支持向量机和神经网络），到基于深度学习的卷积神经网络（Convolutional Neural Networks，简称 CNN）算法（何力等 2021）。目前，OCR 模型通常采用基于深度学习的 CNN 算法。这种算法是一种在图像处理和计算机视觉领域内已得到广泛应用的深度学习模型。CNN 通过一系列的处理步骤，包括卷积操作和下采样操作等，能够自动提取图像中的重要特征，从而进行分类和识别。与传统 OCR 识别技术相比，其优点在于采用端到端技术，即从原始图像输入到最终识别结果输出，整个过程都由同一个模型完成，提高了处理大规模图像数据的效率。

卷积神经网络算法内部的连接时序分类法（Connectionist Temporal Classification，简称 CTC）在文字 OCR 领域被广泛使用（Graves et al. 2006）。CTC 可对不定长的序列进行对齐。以国际音标识别来说，传统方法需要把一个音节切分成一个个单独的音素，构造成一个小图对应一个音素。CTC 则不需要，只需把一个音节或者一个词的图片与相应的音节和词的文本对照即可，音素内部对应关系交给模型自动处理。据报道，CTC 方法在端到端文字识别任务中取得了较好的成果（Ayad et al. 2024）。这是我们选择这一算法来训练国际音标识别模型的主要原因。此外，我们还联合使用了 CNN 和循环神经网络（Recurrent Neural Network，简称 RNN）对图像特征进行建模（Dhruv & Naskar 2020）。其中，CNN 能捕捉图片全局信息，提取图像特征，而 RNN 能提取图像上下文特征。

（四）模型训练及性能评价

在训练模型时，我们按照 7:2:1 的比例随机将训练数据划分成训练集、验证集和测试集。模型训练的硬件环境如下：在四个 GPU 卡的服务器上进行，使用 GPU-0 卡，迭代次数都为 400 次。首先，根据数据来源不同，把不同质量的图片数据分开训练，使用来源于 HZ、ZM 和 GX 的三类数据训练了三个模型，分别命名为 IPAOCR-HZ、IPAOCR-ZM 和 IPAOCR-GX。然后，把所有数据合并后，又训练出一个模型，命名为 IPAOCR-IEA。训练结果如表 1 所示。

表 1　不同数据集训练模型的基本情况

数据来源	数据量	训练用时	整词精度	字符查准率	字符查全率	模型大小
IPAOCR-HZ	217023	31:38:50	0.9908	0.9977	0.9977	172M
IPAOCR-ZM	48765	3:28:12	0.9500	0.9908	0.9900	120M
IPAOCR-GX	100387	7:09:02	0.8484	0.9615	0.9756	142M
IPAOCR-IEA	633406	50:59:04	0.9477	0.9808	0.9826	218M

从表 1 数据可知，整词精度在不同数据集上的表现具有明显差异。其中，最低的是由 GX

数据集训练的 IPAOCR-GX 模型，只有 84.84%；最高的是在 HZ 数据集上训练的 IPAOCR-HZ 模型，为 99.08%；在 ZM 和总数据集上训练的 IPAOCR-ZM 和 IPAOCR-IEA，整词精度比较接近，约为 95%。模型性能差异较大的主要原因是 GX 数据集的数据质量相对较差，原文本数据中存在多种音标编码。虽然在制作训练数据时，对编码进行了统一处理，但没有进行全面人工核对，可能存在数据转码错误。不仅如此，这种错误也被带入 IPAOCR-IEA 模型，对其性能造成影响。IPAOCR-HZ 模型的效果最好，其原因是 HZ 的数据质量高，而且图片是原文文档转成 PDF 后获得，类似于合成图片。在利用总数据集训练 IPAOCR-IEA 模型时，我们并没有直接使用数据，而是先将其打印出来进行扫描，再利用扫描图片制作训练数据。

四个模型的字符查全率和字符查准率都超过了 96%。可以说，数据集的质量直接决定了模型的效果，模型 IPAOCR-HZ、IPAOCR-ZM 的查全率和查准率都超过了 99%，而模型 IPAOCR-GX 的查全率和查准率分别是 97.56% 和 96.15%。此外，迭代次数是指算法在训练过程中对数据集进行循环处理和参数调整的次数。我们在“中国少数民族语言参考语法研究系列丛书”（简称“参考语法丛书”）中随机选择一部专著的 PDF 版文件，然后从其国际音标行的图片中，随机抽取 20 行音标数据，在 IPAOCR-IEA 模型中分别迭代 30 次、200 次和 400 次进行测试，其置信度分别是 95.99%、96.83% 和 97.42%。这说明随着迭代次数的增加，置信度也随之增高。但是当迭代次数达到 600 次以后，置信度几乎不再增加。

我们还从“参考语法丛书”原书扫描后的图片中抽取了 20 行音标数据，对 IPAOCR-GX、IPAOCR-ZM 和 IPAOCR-HZ 模型进行性能测试，结果置信度分别是 93.18%、95.48% 和 88.28%。虽然 IPAOCR-HZ 模型在表 1 中的各项指标都是最好的，但是其泛化能力不如另两个模型，主要原因是这个数据集的图片来自文本转换，而不是实际扫描，真实性不足。

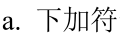
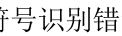
（五）真实数据测试结果

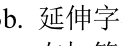
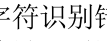
模型的最终价值在于能否处理真实数据。为了测试本研究训练的 IPAOCR-IEA 模型，我们准备了两套词汇数据以检测识别效果。测试数据集 1 是将《那斗语研究》（符昌忠 2021）词汇部分的 100 页转换成图片，然后截取图片 1954 张，并利用模型识别，最后再经过人工核对后形成的标准集。测试数据集 2 是从“新发现语言系列丛书”PDF 文本中抓取图片 3017 张后，再由人工录入的数据。测试结果如表 2 所示。

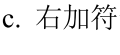
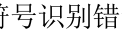
表 2　不同数据集训练模型的基本情况

数据集	数据量	整词精度	字符查准率	字符查全率	字符 F1 度量
测试数据集 1	1954	98.06%	98.99%	100%	99.49%
测试数据集 2	3017	69.80%	93.59%	100%	96.67%

由于测试数据集 1 扫描图片质量高，识别效果非常好。其识别中的主要错误是由于附加符号的相似性导致的。例如，《那斗语研究》中存在表示嘎裂声发声态（creaky voice）的下加符号[◌̰]，IPAOCR-IEA 模型把该符号识别成一般的下划线“_”。总体上看，测试数据集 1 数据测试的主要错误包括三类，如 3a、3b、3c 所示。其中，3a 与 3b 两类错误占比较大，其他错误非常少。识别错误案例如下所示（左边为原书截取图片，右边是识别结果，下同）。

3a. 下加符号识别错误：  01_row011.png:vɛn^{42}ɗaŋ^{25}　  01_row011.png:vɛn^{42}ɗaŋ^{25}

3b. 延伸字符识别错误：  14_row014.png:zɛŋ^{42}　  14_row014.png:zɛŋ^{42}

3c. 右加符号识别错误：  23_row018.png:thun^{25}ha$^{42-52}$  23_row018.png:thun^{25}ha$^{42,\ 52}$

测试数据集 2 是从 PDF 文档转换成图片后再截取的数据。由于原书扫描质量不高，背景噪声较大，导致尽管字符级的测试表现还不错，但是整词精度较低，只有约 70%。这主要是由于某些词的单个字符识别错误导致的，进而造成了整个词条错误，如下所示。

3d. 右加符号识别错误：4196.png:n̥yɛ^{35}gɔ^{35}tʃa^{13}tɕhɛ^{13} 4196.png:n̥yɛ^{35}gɔ^{35}tʃɑ^{13}tɕhɛ^{48}

3e. 延伸字符识别错误：4200.png:d͡za^{13}gɔ^{35}tʃa^{13}tɕhɛ^{13} 4200.png:d͡zɑ^{13}gɔ^{35}tʃɑ^{53}tɕhɛ^{13}

3f. 其他字符识别错误： 7163.png:no^{13}ko^{53}nɒ̃o^{53}ko^{53} 7163.png:no^{13}ko^{53},nɒ̃o^{53}ko^{53}

我们也利用带有文本行的数据对模型进行检测。在许多民族语言和汉语方言的材料中，国际音标的应用场景如图 2 所示。这个格式被称为隔行对照标注，国际音标与解释文本各占一行。对于使用者来说，如果 OCR 模型能同时识别混合排列的中英文和国际音标，那是最理想的。但是，IPAOCR-IEA 模型现阶段只能识别音标，即把音标行单独截取出来进行识别。为了观察效果，我们从“参考语法丛书”中随机截取了 20 行音标数据进行测试。示例如图 2。

图 2 民族语言隔行对照数据例示

具体而言，国际音标数据原图及识别结果如 4a、4b 所示。每一组的第一行是被识别图片，第二行是识别结果，下划线表示识别错误处。模型对 4a 和 4b 识别的置信度分别是 99.16% 和 99.94%。由此来看，识别效果不错。需要说明的是，用于训练模型的数据主要来自于声调语言的语料，右加的声调数字在某种意义上可以充当音节之间的分隔符号。

4a. 01.png： ɑ^{55} də^{31}dzə̯^{55} min^{55} mɑŋ^{55} din^{24}，ɑ^{55} zu̯ɛ$^{24\ 31}$ ɖa$^{24\ 31}$ dzə̯^{55} xɑ^{55} jin^{31} san^{3}

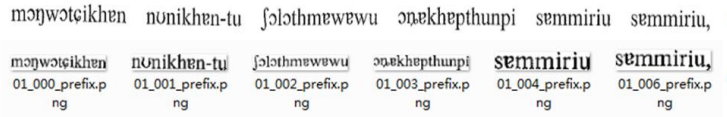
识别结果：[ɑ^{55}də^{31}dzə̯^{55}min^{55}mɑŋ^{55}din^{24}…ɑ^{55}zu̯ɛ^{2431}ɖa^{2431}dzə̯^{55}xɑ^{55}jin^{31}san^{3}]

4b. 02.png： o^{55}ti^{55} nie^{31}stie^{55} gɯ^{55} ɦan^{24} khu^{55} nə^{31}tsiu$^{55\ 31}$ si^{31} mɑŋ^{55} də^{24}.

识别结果：[o^{55}ti^{55}nie^{31}stie^{55}gɯ^{55}ɦan^{24}khu^{55}nə^{31}tsiu^{5531}si^{31}mɑŋ^{55}də^{24}]

但是，当输入的是无声调语言的文本时，由于没有声调符号作为间隔，会导致输出结果中所有字符都连接在一起。为了解决这个问题，我们把无声调语言的文本图片先进行预处理，沿空白间隙切分成短的片段，识别后再进行拼接，如 4c 所示。4c 中，第一行是原图片，第二行是切分后的图片。切分后的小图片的名称表明了其位置，如“01_000_prefix.png”中 01 表示原图像在原文中的行数，000 表示原图切分后的第一个小图，prefix 表示切分后的非空白部分，程序会自动将空白部分删除。

4c.

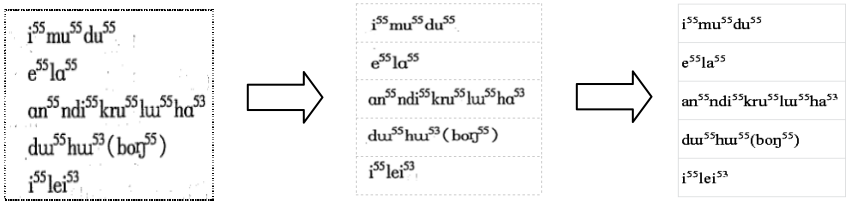


切分处理不仅可解决字符粘连问题，还可以提高识别效率。通过拼接识别结果，模型会留出空格，这使得识别结果格式与输入样式保持一致。例如，4c 未经切分处理的识别结果为：['mɔŋwɔtɕikhɐnnʊnikhɐn-tuʃɔlɔthmɐwɐwuɔŋɐkhɐpthunpisɐmmiriusɐmmiriu']，置信度为 99.96%。虽然置信度较高，但由于丢失了空格，输出的文本格式与输入图片不太一致。而经过切分预处理后，图片的识别结果为：['mɔŋwɔtɕikhɐn nʊnikhɐn-tu ʃɔlɔthmɐwɐwu ɔŋɐkhɐpthunpi sɐmmiriu sɐmmiriu']，置信度为 98.99%；经过切分预处理后，最后输出的文本格式与输入图片类似。由此可见，该模型用于无声调语言的音标识别需要增加图片切分的预处理环节。

除切分预处理外，字符粘连问题还可通过模型来解决。当输入文本时在词语之间加上空格分隔符号，模型就会学习自动识别图片中的空白，并在识别输出的文本中保留空格。实际上，本研究也训练了这样的模型。特别是对有较长空白的图片，通过切断后再识别，就能减少识别的错误。不过，我们最终还是选择了切分预处理的方式，因为预处理后的输出结果与输入图片更加一致，并与能识别空白的模型配合使用，识别的效果也更好。

四　国际音标 OCR 模型的应用

国际音标 OCR 模型主要面向两类应用场景：一是依赖音标数据的数字人文研究、语料库和知识库建设；二是供专家学者学习和写作。第一类处理的文献量大，需要调用国际音标 OCR 模型，例如民族语言或汉语方言的隔行标注语料库构建及词典编纂等。针对这种带有混排版面的应用场景，我们先将中文和国际音标自动分离，然后分别调用中文 OCR 模型和国际音标 OCR 模型，再把识别结果根据分离图片名称的 ID 号拼接起来，即可实现混排版面的国际音标识别。第二类是研究者在学习或研究中需要获取国际音标书写的例词或例句，通常只需截取一行或几行文本进行数据处理即可。针对这种需求，我们设计了一款轻量化的国际音标识别软件——“龙水国际音标识别软件”。该软件可通过截取图片的方式识别国际音标字符，并输出识别结果。其工作流程如图 3 所示，主要包括图片截取（图 3a）、行切分及图片预处理（图 3b）、调用国际音标识别模型并输出文本（图 3c）三个步骤。



（a）　　（b）　　（c）

图 3　“龙水国际音标识别软件”的识别流程

“龙水国际音标识别软件”遵循“有需即取”的设计理念。该软件安装简单，使用方便。用户在阅读和写作中，如需从 PDF 或者图片中获取由国际音标记录的词、短语、句子，只需选中目标区域，就可以提取国际音标文本。该软件能实现对单个词语、单行或多行文本的音标识别。对带有声调的国际音标进行图片识别，既可以选择“单行文本”，也可以使用“单个词语”选项（如图 4）。如需对不带声调的音标进行图片识别，推荐使用“单行文本”选项。

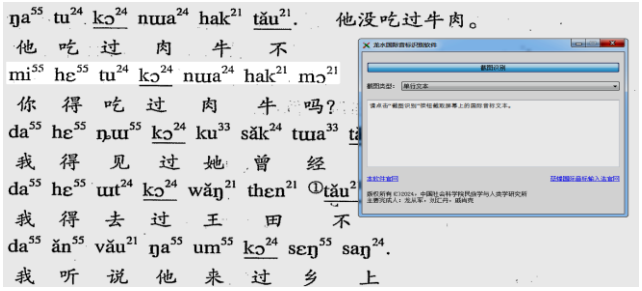


图 4　“龙水国际音标识别软件”识别示例

五 结 语

国际音标的自动转写对诸多人文社会科学研究具有重要的意义。国际音标自动识别模型 IPAOCR-IEA 正是以此为目标而研发的。该模型的成功开发不仅得益于现阶段成熟可用的深度学习算法，也得益于中国语言学界所提供的越来越多的语料支持，两者缺一不可。

不过，该模型还存在以下两点不足：其一，对于少数符号及特殊附加符号的识别仍然可能存在错误；其二，对于铅印版文本（较早期的出版物）的识别效果不如激光版印刷文本，印刷技术的差异会影响识别效果。在后续研究中，我们将继续改善模型识别的精度，提升识别效果和模型的泛化能力，以便更好地满足用户的需求。

参考文献

常晨霞、凌永发. 2013.《图像目标识别与跟踪方法研究》,《云南民族大学学报（自然科学版）》第 S1 期.

符昌忠. 2021.《那斗语研究》，北京：民族出版社.

高本汉. [1940]2003.《中国音韵学研究》，赵元任、罗常培、李方桂译，北京：商务印书馆.

广西壮族自治区少数民族语言文字工作委员会编. 2008.《广西民族语言方音词汇》，北京：民族出版社.

国际语音学会编. 2020.《国际语音学会手册：国际音标使用指南》（中文修订版），江荻、孟雯译，上海：上海教育出版社.

何　力、郑灶贤、项凤涛、吴建宅、谭　林. 2021.《基于深度学习的文本分类技术研究进展》,《计算机工程》第 2 期.

江　荻编著. 2012.《藏文识别原理与应用》，北京：商务印书馆.

钱丽花. 2007.《统一平台的多民族文字文档识别系统研制成功》,《中国民族报》1 月 30 日第 1 版.

邱立松. 2015.《国际音标字符识别算法的研究》，上海师范大学博士学位论文.

孙宏开、丁邦新、江　荻、燕海雄主编. 2016.《汉藏语语音和词汇》，北京：民族出版社.

田德生、何天贞、陈　康、李敬忠. 1986.《土家语简志》，北京：民族出版社.

燕海雄、江　荻. 2007.《国际音标符号的分类、名称、功能与 Unicode 编码》,《语言科学》第 6 期.

《藏缅语语音和词汇》编写组编. 1991.《藏缅语语音和词汇》，北京：中国社会科学出版社.

周志华. 2016.《机器学习》，北京：清华大学出版社.

Ayad, Abdessamad, Habib Ayad and Abdellah Adib. 2024. A survey on scene text recognition in natural images. In Brahim El Bhiri, Rajaa Saidi, Mohammed Essaaidi and Naima Kaabouch (eds.), *Smart Mobility and Industrial Technologies: The Quality of Life in Sustainable Cities*, pp. 81-87. Cham: Springer.

Dhruv, Patel and Subham Naskar. 2020. Image classification using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A review. In Debabala Swain, Prasant Kumar Pattnaik and Pradeep K. Gupta (eds.), *Advances in Intelligent Systems and Computing: Machine Learning and Information Processing, Proceedings of ICMLIP 2019*. Vol 1101, pp. 367-381. Singapore: Springer.

Graves, Alex, Santiago Fernández, Faustino John Gomez and Jürgen A Schmidhuber. 2006. Connectionist temporal classification: Labeling unsegmented sequence data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369-376. Pittsburgh, Pennsylvania, June 25-29.

International Phonetic Association. 1949. The principles of the International Phonetic Association. *Le Maître*

Phonétique 27(64): 1-53.

Li, Shihua and Nathan Hill. 2023. Printed text recognition for lexical lists in Chinese-International Phonetic Alphabet (IPA) glossing. *Journal of Open Humanities Data* 9(15): 1-8.

Roach, Peter J. 1989. Report on the 1989 Kiel Convention: International Phonetic Association. *Journal of the International Phonetic Association* 19(2): 67-80.

Tafti, Ahmad P., Ahmadreza Baghaie, Mehdi Assefi, Hamid Arabnia, Zeyun Yu and Peggy Peissig. 2016. OCR as a service: An experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In George Bebis, Richard Boyle, Bahram Parvin, et al. (eds.), *Advances in Visual Computing: 12th International Symposium, ISVC 2016*, pp. 735-746. Las Vegas, NV, USA, December 12-14.

Optical Character Recognition of the International Phonetic Alphabet: Research and Applications

LONG Congjun and GUO Chengyu

[Abstract] The International Phonetic Alphabet (IPA) is a widely used system for phonetic notation and transcription. Optical character recognition (OCR) technology of IPA can enable bulk IPA recognition, significantly improving the efficiency of IPA processing. This technology is of high value for the digitization of linguistic resources and research on ethnic minority languages and Chinese dialects. This paper presents "IPAOCR-IEA", an innovative IPA recognition model developed by the authors. The model, trained on a dataset of over 630,000 IPA image-text pairs, utilizes a convolutional neural network and can efficiently identify IPA characters. It achieves top performance across several metrics, including word accuracy, character precision, and recall, with all metrics exceeding 98%. Additionally, the lightweight "Longshui IPA OCR Software" based on this model has been launched, enabling users to automatically capture images and extract IPA characters from the selected regions.

[Keywords] International Phonetic Alphabet optical character recognition model automatic IPA recognition

（通信地址：龙从军　100081　北京　中国社会科学院民族学与人类学研究所
郭承禹　519085　珠海　北京师范大学文理学院）

【本文责编　胡鸿雁】