医学文本图像字符识别校正技术研究与应用

李 琴 杨 斌 郇宝贵 江俊龙

（青岛百洋智能科技股份有限公司 青岛 266042）

〔摘要〕 针对医学文本图像字符识别的后处理技术进行研究，通过建立医学常用专业词库，基于中文汉字图像的 Hog 特征和相关系数计算字符之间的文字相似度，从而对识别后的文本进行拼写检查校正，提高字符识别准确率。

〔关键词〕 医学词库；Hog 特征；字符识别；校正

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.06.013

Study and Application of the Recognition and Correction Technology of Medical Text Image Characters

LI Qin, YANG Bin, HUAN Bao－gui, JIANG Jun－long, Qingdao Baheal Intelligent Technology Co., LTD, Qingdao 266042, China

〔**Abstract**〕 The paper studies on the post－processing technology of recognition and correction technology of medical text image characters. By building the common medical lexicon, it checks and corrects spelling of the recognized text on the basis of the Hog feature of Chinese character images and word similarity between correlation coefficient calculation characters, so as to enhance the accuracy rate of character recognition.

〔**Keywords**〕 Medical lexicon; Hog character; Optical Character Recognition (OCR); Correction

1 引言

随着医疗改革和信息化的推进，越来越多的医疗信息系统已建立并为医疗健康产业服务，这些信息化系统在整合医疗机制、更好地为患者服务的同时也提高医生的工作效率。但是由于医疗资源的独立性特点，不同医疗信息系统之间还没有建立统一的数据共享机制，一些医疗科研信息系统没有或很难与医院信息系统建立对接，不同医疗信息系统之间的信息传输成为难题，同样的医学数据可能需要在不同的信息系统上人工录入数次，耗费人力。字符识别技术（Optical Character Recognition，OCR）的进步使其在医疗信息系统上的应用成为可能。医学数据必须保证其数据的准确性和一致性，由于目前字符识别技术特别是在医学术语的识别上还不能达到直接应用的水平，需要人工复核来保证数据的准确性，因此从各方面完善医学文本图像字符识别技术具有必要性和现实意义。

字符识别技术在国内外多个行业有一定的应用，如名片、身份证识别、银行卡识别等。但在医疗行业的应用较少[1－2]，且限于对特定检验检查单的识别，对于医学字符识别校正技术的研究更是不多[3－5]。本文以字符识别技术应用于医疗信息系统为出发点，深入研究医学文本的特点，结合字符识

〔收稿日期〕 2018－03－05

〔作者简介〕 李琴，高级工程师；通讯作者：杨斌，助理研究员。

别技术，建立一种医学专业字符识别校正方法，可以提高字符识别结果的准确性。

2 字符识别技术和结果错误分析

2.1 字符识别技术

2.1.1 概述 光学字符识别技术是利用光学技术和计算机技术将印在或写在纸上的文字读取出来，转换为文本。其过程的输入是文本图像，输出为计算机可直接识别的文本数据。文字识别是计算机视觉研究领域的分支之一，目前的研究已取得一定的成果，在相关生产领域应用。

2.1.2 识别形式 分为印刷体识别和手写体识别，目前较为成熟的应用为印刷体识别，本文的医学文本识别对象也是印刷体。即便是印刷体的识别也存在一定的困难，如字符的字体、字号大小以及在印刷过程中字体很可能出现断裂或墨水粘连，使识别异常困难，或者扫描和拍照生成图像质量较差问题等都会影响识别效果。

2.1.3 识别内容 按文本识别的语言分类，识别内容将是人类的所有语言（汉语、英语、德语、法语等）。按识别的内容分类又有所不同，如医学行业识别的内容包括：汉字、英文字母、阿拉伯数字、希腊字母等类型。根据识别内容的不同识别难度也不相同。在所有字符识别中难度最大的是中文，字符高达数千个，结构非常复杂，因此要将中英文、其他字符等混合一起的字符准确识别出来具有一定的挑战。

2.1.4 识别过程 包括图像预处理、数据降维和特征选择、分类算法和结果后处理等过程。在深度学习未兴起之前一般的字符识别过程多采用支持向量机（Support Vector Machine，SVM）等传统机器学习分类方法。随着深度学习技术的进步循环神经网络（Recurrent Neural Nemork，RNN），长短期记忆神经网络（Long Short Term Memory，LSTM）等技术为字符识别提供新的思路，提高识别准确率。目前谷歌开源字符识别项目 Tesseract[6] 也开始转向基于深度学习算法的研究和应用。

2.2 结果错误分析

字符识别的效果除与算法有关外，与训练样本也有很大的关系。因此医学文本图像的字符识别需要建立在大量医学常用词汇和符号的训练样本之上。虽然经过大量的训练字符识别技术仍然存在一定错误率，不同的错误形式可以采用不同的技术来改进，如文本图像质量改进预处理和识别结果拼写检查后处理等。字符识别的错误类型分为识别缺失、多字、少字、错字等问题，其中出现较多的还是识别错字问题。本文提出的方法主要解决识别错字的问题，通过校正提高识别结果准确率，属于字符识别后处理。

3 医学文本图像字符识别校正技术

3.1 识别过程

文本图像的字符识别在医学信息化推进中有重要应用价值，将医学信息文件从不同的数据介质通过移动图像设备转换为文本图像，或原本是文本图像的数据直接转换为信息系统之间读取应用的文本数据，加快医学数据的信息传输效率。医学文本图像的识别过程包括数据源的整理、文本图像预处理、OCR 识别、识别结果后处理和医疗信息化应用 5 个环节，见图 1。其中识别结果后处理即应用本文提出的医学文本图像识别校正技术基于医学词库进行识别校正。应用拼写检查校正技术后的医学文本数据可以通过数据提取技术将医学图片数据电子化，或将对应的医学信息数据直接应用到医疗信息化系统中，为医疗信息化服务。

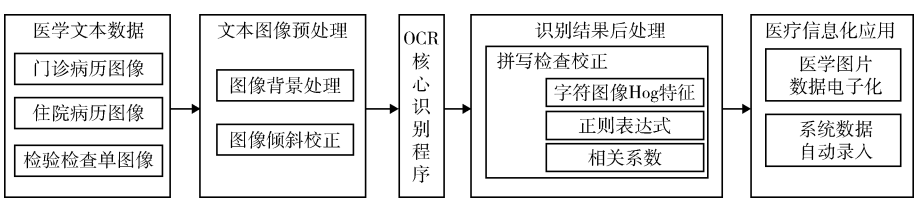


图 1 医学文本图像字符识别过程

3.2 一般拼写校正技术及其局限性

拼写检查校正技术一般应用于办公软件和搜索引擎中，如微软 Word、谷歌浏览器和百度搜索引擎拼写检查等。拼写检查校正技术的研究对象多为西方语言特别是英语，而对于中文的拼写检查，由于其语法复杂，实现更为困难。目前拼写检查的主流算法是根据词语的编辑距离来计算词与词之间的相似度，从而对错误词语选择最相似的正确词语来校正。但是这种方法对于 OCR 识别结果的文本纠错来说并不合适。OCR 在识别过程中的词语错误是字体形状造成的，而拼写检查校正的错误词语往往是字母或拼音的拼写错误或其他方式造成，不完全是字体形状的原因，因此传统基于编辑距离的拼写检查校正对于 OCR 识别结果的字符校正并不适用。

3.3 医学词库建立

字符识别结果的准确性不仅依赖于算法本身，更重要的是词库的建立。根据应用场景建立相应的识别词库，无论是对于字符字集的训练还是识别后结果的校正都有重要作用。医学词汇有明显的规范性和专业性特点，特别是门诊病历和医学检验检查单在行业内有一些相关标准可以参考，语言描述形式相对固定。建立准确的医学专用词库不仅可以作为医学字符识别训练集来训练字符，也可以用于医学识别校正算法。医学文本图像的数据资源以门诊病历、检验检查单、药品库等为主，根据医学字符识别的专业性，以上述几类数据类型为主，通过参考行业标准、医学词典、医学论文等资源建立一个专业的用于医学文本图像字符识别的词库，用于对识别结果专业词汇的纠错，同时也可以补充医学字符识别训练集。部分医学专业词语，见图 2。本文中校正算法是根据词语的长度从长到短的顺序逐词校正，因此应注意词库中的词语包含问题。如“嗜酸性粒细胞计数”和“白细胞计数”都存在于医学词库中，当待校正文本中出现“嗜酸性白细胞计数”时，校正算法会优先按照词库中的“嗜酸性粒细胞计数”进行校正。因此在建立医学专业校正词库时要特别注意包含词语的添加和删除。另外本文中的医学专用词语选择 3 个字符以上的词语，因为低于 3 个字符的医学词语不再具有特异性，可能会造成词语纠错的错误。

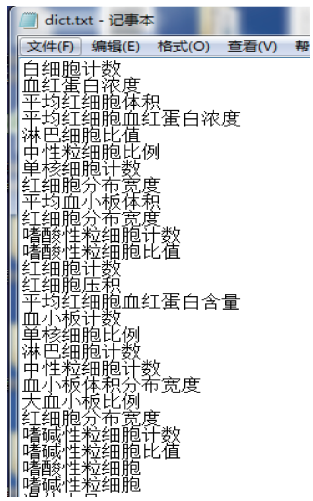


图 2 部分医学专业词语

3.4 基于方向梯度直方图特征的文字字符相似度计算

在进行拼写检查校正时，检索到的待纠错词语可能会出现几个对应的正确词汇，如识别结果为“嗜皮性粒细胞计数”，该词语在医学词库中有两个可以校正的词语，分别为“嗜酸性粒细胞计数”和“嗜碱性粒细胞计数”。因此需要结合具体错字和正确词库中的待纠错字进行字符相似度的计算，取最相似的字对应的正确词语进行校正。采用字符图像方向梯度直方图（Histogram of Oriented Gradient，HOG）特征的相关系数计算方法，是一种在计算机视觉和图像处理中用来进行物体检测的特征描述符，通过计算和统计图像局部区域的梯度方向直方图来构成特征。首先运用文本图像生成技术建立每个字符的字符图像（约 5 130 个字符），每个字符图像大小为 64 × 64 像素。部分识别字符图像集，见图 3。之后运用图像分析学中的特征提取算法和相似度计算方法计算各字符之间的相似度。图像特征

图 3 部分识别字符图像集

提取方法如 Hog 特征码、LBP 特征和 Haar 特征等，相似度计算方法有 4 种，分别为 Correlation 相关性、Chi - Square 卡方、Intersection 交集法、Bhattacharyya 距离法。本文选择基于 Hog 特征和 Correlation 相关性来计算字符图像之间的相似度。

首先将图像分成小的连通区域，然后采集细胞单元中各像素点的梯度的或边缘的方向直方图。最后将这些直方图组合就可以构成特征描述算子。由于 Hog 是在图像的局部方格单元上操作，所以对图像几何和光学形变都能保持很好的不变性。Correlation 相关系数则是通过以下公式计算 H_1 和 H_2 之间的相似度，计算公式如下[3]：

$$d(H_1,H_2) = \frac{\sum_I (H_1(I) - \overline{H_1})(H_2(I) - \overline{H_2})}{\sqrt{\sum_I (H_1(I) - \overline{H_1})^2 \sum_I (H_2(I) - \overline{H_2})^2}}$$

$$\overline{H_k} = \frac{1}{N}\sum_J H_k(J)$$

其中 N 是直方图中 bin 的数目。其结果范围为 [-1, 1]。不考虑负相关的情况，则值越大表示两个字符的相似度越大。基于 Opencv - python 软件包开发识别字符图像的 Hog 特征，将特征值存为文件形式。由于字符数量大，所有字符之间的相似度计算结果可以采用实时计算的方式，只需在校正时对待校正的错误字符和备选校正的正确字符之间计算相似度即可，可以大大降低计算时间和空间复杂度。

3.5 医学字符识别拼写检查校正算法

字符识别的技术限制和图像质量造成识别结果总会存在一定的错误率，但可以通过词语校正实现字符识别的后处理降低错误率。拼写检查校正是对识别算法识别后的文本结果进行错词校正，通过对比医学专业词库、计算错字与正确字之间的相似度选择最终正确的词汇进行纠正。该算法的流程，见图 4。具体实现过程为：（1）建立医学专业术语词库。包括门诊病历、检验检查单和药品常用词，为避免校正错误，一般词语为 3 个字符以上。（2）基于医学专业术语词库建立正则表达词语库。本文以单字符模糊匹配为准则，如“骨质疏松”的对应正则表达模糊匹配词语为“. 质疏松”、“骨. 疏松”、“骨质. 松”、“骨质疏.”。（3）输入字符识别后的待校正文本。遍历正则表达匹配词集，匹配出待校正的错误词语。（4）提取相似度校正。将错误词语的错字与对应待选择的正确医学专业术语词语的字计算相似度，取相似度最高的字进行替换校正。

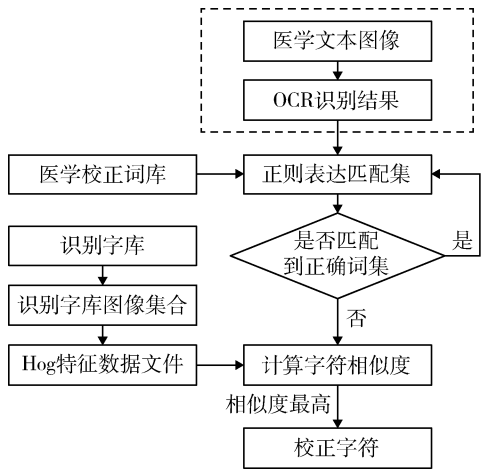


图 4 医学字符识别拼写检查校正流程

4 实验验证

基于开源 OCR 技术和拼写检查校正技术建立医学专业 OCR 系统，分别选择门诊病历和检查单文本图像进行识别，门诊病历字符识别结果校正前后对比，见图 5。通过对比发现运用拼写检查校正方法可以较好地将识别的错字按照医学词库的词语校正，从而提高准确率。医学检验检查单识别效果校正前后对比，见图 6。该检验单由于图片存在倾斜、模糊等情况，质量较差，识别效果不好，但是应用该方法仍然能够将医学专用词汇校正准确，提高识别效果。正确的医学专业词汇识别为下一步从文本中自动识别出对应的数值填写到医疗信息系统对应的表单中提供支持，实现从医学文本图像到医疗信息系统医学数据的自动识别上传。

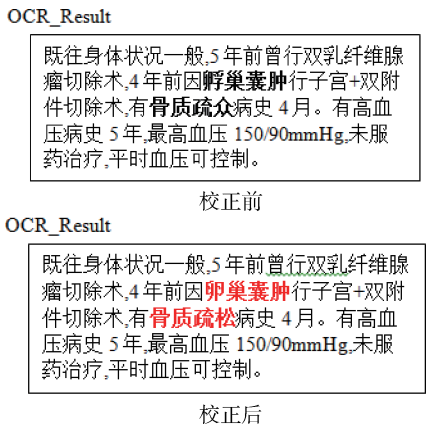
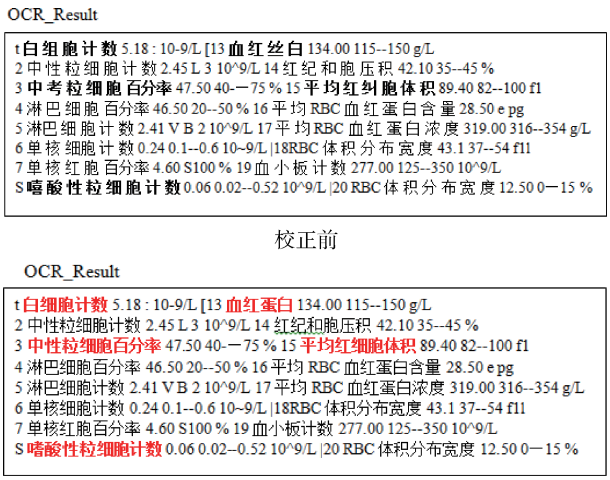


图 5 门诊病历字符识别结果校正前后对比



校正前

校正后

图 6　医学检验单字符识别结果校正前后对比

5　结语

应用字符识别校正算法，建立专业的医学词库，实现对医学文本图像的字符识别校正，提高医学字符识别准确率。该方法可以直接应用于有文本数据录入需求的医疗信息系统，用于辅助用户自动、高效、准确地将医学数据输入到信息系统中，降低人力成本。将医学文本图像识别技术与校正方法结合，开发相应的医学数据拍照自动录入模块嵌入到医疗信息系统中，可以有效地提高医学检验数据的录入效率和准确率，具有一定的实际应用价值。

参考文献

1　向明华，向国华. 基于 OCR 技术的医疗档案管理系统研究与构建［J］. 中国医疗设备，2015，30（10）：106－107.
2　郭世雄. 医疗仪器中的数字识别技术研究［D］. 西安：西安建筑科技大学，2009.
3　Shaohua Yang，Hai Zhao，Xiaolin Wang，et al. Spell Checking for Chinese［C］. Istanbul：Proceedings of the Eight Interational Conference on Language Resources Evaluation（LERC'12），2012：730－736.
4　王宸敏. 基于 OCR 技术的化验单识别方法研究［D］. 杭州：浙江大学，2016.
5　Opencv Dev TeamHistogram Comparison［EB/OL］.［2017－02－14］. https：//docs. opencv. org/2. 4/doc/tutorials/imgproc/histograms/histogram _ comparison/histogram _ comparison. html.
6　Ray Smith，Daria Antonova，Dar－Shyang Lee. Adapting the Tesseract Open Source OCR Engine for Multilingual ORC［C］. Barcelona：Procedings of the Interational Workshop on Multiligual OCR 2009，2009.

（上接第 46 页）

结构，解决相同疾病监测在不同部门、单位之间“信息孤岛”的问题，使新建系统不仅限于满足本单位、部门的需要[6]。本研究中的概念模型是基础也是关键，最终目标是增强数据利用，促进数据共享。使用该概念模型，必然在数据库设计中重复使用相应构成，这种重复使用会为独立开发的各个软件系统在数据上展现一致性。数据的一致性是非常重要的特征，其提高将使疾病监测应用在不同信息系统之间共享数据更加便捷，使复杂数据映射和转换过程变得简单。数据的一致性还将允许数据跨越多重系统进行比较和连接，同时有利于分析、发现趋势，促进公共卫生大数据平台的建立。提供可重复使用的数据分析及数据库设计，最终开发出通用平台，将大幅减少开发时间和费用。本研究模型还不够完善，能否代表登革热疾病防控的所有业务细节尚有待验证，检验模型最好的方法是通过将模型不断应用在系统建设中来发现不足，后续将通过逻辑、物理建模来分析数据模型的优劣，进一步完善模型。

参考文献

1　孟凤霞，王义冠，冯磊，等. 我国登革热疫情防控与媒介伊蚊的综合治理［J］. 中国媒介生物学及控制杂志，2015，（1）：4－10.
2　胡大权. 数据库概念模型的分析与应用［J］. 计算机工程与应用，2002，（22）：211－214.
3　刘丹红，徐勇勇，王霞，等. 卫生信息概念数据模型与数据元标准研究［J］. 中国卫生质量管理，2005，（6）：1－3.
4　张先波，金水高，刘丽华. 公共卫生实验室检测活动类的提取与泛化［J］. 中国医疗器械杂志，2007，（4）：248－252.
5　郭赟，金水高，刘丽华. 场景分析在公共卫生信息概念模型构建中的应用研究［J］. 中国医院，2007，（7）：32－34.
6　Boudreaux E D，Cydulka R，Bock B，et al. Conceptual Models of Health Behavior：Research in the Emergency Care Settings［J］. Academic emergency medicine?：official journal of the Society for Academic Emergency Medicine，2009，16（11）：1120－1123.