光学字符识别对卒中结构化随访表格识别效果的对比研究

王驰 王伟 邱玉发 唐冬梅 舒张(通信作者)
215400太仓市第一人民医院卒中中心，江苏 太仓

doi:10.3969/j.issn.1007-614x.2022.29.055

摘 要 目的：研究不同光学字符识别(OCR)方案在真实环境中对纸质卒中随访表格的识别效果，探索OCR技术在结构化卒中随访表格电子化中的可行性。方法：收集太仓市第一人民医院2019-2020年社区、乡镇人群心脑血管病危险因素纸质随访表，根据图像采集质量分为正常、角度不佳、光线不佳。文字识别分别采用通用OCR和百度自定义模板文字识别(IOCR)，输出文字经人工校对与原始数据对比，计算错误率。结果：通用OCR的正常、角度不佳、光线不佳识别率分别为99.3%、86.1%、97.1%，IOCR的正常、角度不佳、光线不佳识别率分别为99.0%、93.7%、98.1%。通用OCR与IOCR总体识别率比较，差异有统计学意义($P<0.05$)。正常、角度不佳、光线不佳图片的通用OCR识别率比较，差异有统计学意义($P<0.05$)；正常、角度不佳、光线不佳图片的IOCR识别率比较，差异有统计学意义($P<0.05$)。结论：结构化表格通过模板构建的IOCR识别方案可以提高识别率，在正常采集角度下能够基本满足临床需求，极大地提高医生的录入效率，形成电子档案，说明图像采集角度是影响识别率的重要因素，为数据的进一步利用提供基础。

关键词 光学字符识别；卒中；结构化表格

Comparative Study on Recognition Effect of Optical Character Recognition on Structured Follow-Up Form of Stroke

Wang Chi, Wang Wei, Qiu Yu-fa, Tang Dong-mei, Shu Zhang(corresponding author)
Stroke Center, Taicang First People's Hospital, Taicang 215400, Jiangsu Province, China

Abstract Objective: To explore the recognition effect of different optical character recognition (OCR) schemes on paper stroke follow-up forms, and to explore the feasibility of OCR technology in the electronic identification of structured stroke follow-up forms. Methods: The paper follow-up forms of community and township population with risk factors for cardiovascular and cerebrovascular diseases in the Taicang First People's Hospital from 2019 to 2020 were collected and divided into normal group, bad angle group and bad light group according to the quality of image acquisition. Text recognition adopted general OCR and Baidu OCR Text Recognition (IOCR). The output text is manually proofread and compared with the original data to calculate the error rate. Results: The recognition rates of normal, bad angle and bad light images were 99.3%, 86.1% and 97.1% by the general OCR, and were 99.0%, 93.7% and 98.1% by the IOCR, with statistical difference ($P<0.05$). There was a statistical difference in the overall recognition rate between the general OCR and the IOCR ($P<0.05$). There was a statistical difference in the recognition rate of normal, bad angle and bad light images by IOCR ($P<0.05$). Conclusion: Structured forms can improve the recognition rate through the template construction IOCR recognition scheme, basically meet the clinical needs under the normal acquisition angle, greatly improve the input efficiency of the physicians, and formulate electronic archives, suggesting that image acquisition angle is an important factor affecting the recognition rate, which provide the basis for further application of data.

Key words Optical character recognition; Stroke; Structured forms

目前随着卒中中心建设的持续推进，将结构化随访表格用于患者随访的情况日益普遍，标准化的卒中随访与改善患者出院后的生活自理能力和降低再住院率有关[1-2]。其中积累的大数据是进行分析提高卒中防治效果的宝贵资源，目前国内大型临床试验数据管理逐渐从纸质化不断向电子化数据采集(electronic data capture，EDC)转变。但在客观层面，电子病历方案建设成本高，在卒中随访中纸质储存仍是普遍的储存形式，但其不利于保存或进行大数据分析[3]。同时结构化表格具有冗余性、多样性的特点，有研究显示，人工录入的数字化方式其平均时间成本每个字段为87.62 s[4]。因此高效的电子化技术对卒中治疗发展

基金项目 苏州市青年科技项目(kjxw2018064)

具有重要意义。

光学字符识别(optical character recognition，OCR)是一种较为低成本的数字化方式，但识别率是整个OCR工作的核心问题。有专家认为OCR识别率若＜90%则后期人工校正工作将抵消OCR所带来的效率。现实中由于图像采集环境和方式的不稳定性，OCR在实际场景中的识别率常低于标准，也是限制其实际应用的核心问题。随着现代化技术的发展，对于结构化表格，可以通过模型构建实现进一步提高在真实环境下的识别率。本文旨在分析不同OCR方案在真实环境中对纸质卒中随访表格的识别效果，探索OCR技术在结构化卒中随访表格电子化中的可行性。

资料与方法

收集太仓市第一人民医院2019-2020年社区、乡镇人群心脑血管病危险因素纸质随访表，图像采集均为手机拍摄，摄像头1 200万像素，图片分辨率统一为1 080×1 440，图片大小2.8 MB，共计60张。根据图像采集质量分为正常(图片采集角度＜30°且无光线阴影)、角度不佳(图片采集角度30°～45°)、光线不佳(采集时图片内容中存在明显的光线阴影)。

方法：①通用文字识别：60张图片均采用通用OCR和百度自定义模板文字识别(IOCR)分别识别1次。社区、乡镇人群心脑血管病危险因素随访表数字化后共计1 236字符，将正常、角度不佳、光线不佳图片分别输入OCR，文字识别调用百度通用文字识别API进行识别，输出文字经人工校对与原始数据对比，计算错误率。②模板构建及识别：采用IOCR功能，以电子版社区、乡镇人群心脑血管病危险因素随访表作为上传模板，以版式中位置和内容固定不变的字段作为参照字段，对识别图片进行校正。将正常、角度不佳、光线不佳图片分别输入OCR，文字识别调用构建的IOCR模板API进行识别，输出文字经人工校对与原始数据对比，计算错误率。

统计学方法：数据采用SPSS 22.0统计学软件分析，经*Shapiro-Wilk*法进行正态性检验，识别正确率不服从正态分布，采用中位数(最小值，最大值)表示。通用OCR和IOCR的组间比较采用*Wilcoxon*符号秩检验。三组总体比较采用*Kruskal-Wallis H*检验，两两比较采用*Bonferroni*法。

结　果

通用OCR与IOCR的识别率比较：通用OCR与IOCR识别率比较，差异有统计学意义($P<0.05$)。正常、角度不佳、光线不佳图片的通用OCR识别率比较，差异有统计学意义($P<0.05$)；正常、角度不佳、光线不佳图片的IOCR识别率比较，差异有统计学意义($P<0.05$)。见表1。

讨　论

本研究中，通用OCR和IOCR在正常情况下识别率均＞99%，与理论值相仿。但在角度不佳或光线不佳情况下总体识别率明显低于正常情况，差异有统计学意义($P<0.05$)。角度不佳情况下通用OCR和IOCR识别率分别为86.1%和93.7%，这与OCR系统的图像预处理有关，倾斜角度＞3°就可能导致在矫正过程中对字符产生切割和识别错误[5]。因此病历拍照存档多采用高拍仪以保证拍摄角度，而卒中相关软件多为手机端应用，所以在图像采集时应注意拍摄角度问题[6]。光线不佳对识别率影响较小，通用OCR及IOCR识别率分别为97.1%及98.1%，需要校对的文字较少。

随着卒中建设的不断推进，特别是向县域基层医疗机构推进的过程中信息化发展不平衡更为明显，经济发达地区信息化水平明显高于欠发达地区和农村基层医院，一、二级医院由于受资金和人才因素的制约，信息化建设整体水平偏弱，高成本的电子化建设方案无法普遍适用[7]。在实际工作中电子化档案管理仍存在如技术发展不成熟、档案整合性差、信息储存方式不合理等问题[8]。同时国内卒中相关的大数据研究逐渐增多，电子化要求日益突出。尹芳等[9]研究显示，与纸质材料相比，电子化数据研究成本减少≥60%，同时电子化数据的临床试验完成时间较纸质材料提前1～3个月。因此加强档案电

表1　通用OCR与IOCR的识别率比较[%(最小值，最大值)]

识别方法	总体	正常	角度不佳	光线不佳
通用OCR	97.1(70.1，99.5)	99.3(99.0，99.5)	86.1(70.1，96.8)	97.1(72.5，98.8)
IOCR	98.2(79.2，99.6)	99.0(98.2，99.6)	93.7(79.2，98.9)	98.1(91.4，98.7)

子化进程，优化电子化管理模式有利于更好地提高卒中救治推进及整体效果。

综上所述，不同OCR识别方案在真实环境中对卒中随访表格的识别效果有显著差异，结构化表格可通过模板构建的IOCR识别方案提高识别率，在正常采集角度下能够基本满足临床需求，极大地提高医生的录入效率，形成电子档案，为数据的进一步利用提供基础。

参考文献

[1] Terman SW,Reeves MJ,Skolarus LE,et al.Association Between Early Outpatient Visits and Readmissions After Ischemic Stroke[J].Circ Cardiovasc Qual Outcomes,2018,11(4):e004024.

[2] Condon C, Lycan S, Duncan P, et al. Reducing Readmissions After Stroke With a Structured Nurse Practitioner/Registered Nurse Transitional Stroke Program[J].Stroke,2016,47(6):1599-1604.

[3] 肖豪杰,温嘉悦.纸质病案无纸化方案基于成本效益的对比分析[J].劳动保障世界,2019(30): 74-75.

[4] Dorr DA, Phillips WF, Phansalkar S, et al. Assessing the difficulty and time cost of de-identification in clinical narratives[J].Methods Inf Med,2006,45(3):246-252.

[5] 卜飞宇,刘长松,丁晓青.灰度名片图像快速倾斜检测和校正方法[J].中文信息学报,2004,18(1):62-69.

[6] 周广清,丁苏青.纸质病历扫描拍照存档系统的研发与应用[J].医疗卫生装备,2015(1):38-39.

[7] 毛园园.医院信息化建设现状与发展对策研究[J].中国管理信息化,2018,21(8):43-44.

[8] 王莉.大数据环境下高校档案电子化管理研究[J].中国多媒体与网络教学学报(电子版),2019(5):62-63.

[9] 尹芳,陈君超,刘红霞,等.临床试验纸质与电子化数据管理的比较研究[J].药学学报,2015,50(11):1461-1463.

(上接第163页)

更准确地掌握饮食搭配原则，同时借助饮食模型可指导患者养成健康的饮食习惯[8-9]。不良生活习惯是影响2型糖尿病患者病情控制的重要原因，因此对患者普及健康生活习惯的重要性，嘱其遵循健康科学的生活方式，戒烟、限酒、适当运动有利于提高病情的控制效果。根据患者病情控制效果，制定运动计划，可在患者耐受的基础上开展有效运动，从而养成健康的生活习惯。在上述疾病管理干预基础上，对患者进行随访，了解其病情控制情况及遇到的问题，并及时给予针对性解决方案。在患者复诊时引导其进行全科门诊治疗，以本院多个科室组成的诊疗小组对其进行综合性诊治，在发挥治疗作用的同时还能够及时发现是否出现其他系统的损伤。药物治疗结合辅助检查能够给予患者专业性的指导，帮助其更好、更规范地用药，从而保证病情控制效果[10]。

综上所述，社区糖尿病团队管理与全科门诊治疗应用于2型糖尿病患者中有显著效果，有利于患者控制血糖、血压、血脂水平，降低并发症发生情况，改善患者生活质量。

参考文献

[1] 王雪芳.2型糖尿病社区团队管理与全科门诊治疗的疗效[J].名医,2020,84(5):90.

[2] 徐戟.家庭医生团队管理对社区2型糖尿病患者的影响[J].深圳中西医结合杂志,2019,29(10):135-137.

[3] 叶静雪,曾庆秋,钱宁,等.基层医疗机构2型糖尿病患者家庭医生团队中医药健康管理服务现状调查分析[J].中国全科医学,2019,22(11):22-26.

[4] 刘玉,朱清艳.全科-专科联合模式管理社区2型糖尿病患者的效果评价[J].上海医药,2020,41(6):45-47.

[5] 周恩飞,郭翔廷,吕文晴,等.家庭医生“1+1+1”签约服务联合全科团队管理对社区2型糖尿病患者达标的影响研究[J].贵州医药,2019,43(11):1759-1760.

[6] 张国荣,林庆.家庭医师团队综合管理对社区2型糖尿病预后的影响[J].临床合理用药杂志,2020,13(8):151-153.

[7] 中华医学会,中华医学会杂志社,中华医学会全科医学分会,等.2型糖尿病基层诊疗指南(实践版·2019)[J].中华全科医师杂志,2019,18(9):810-818.

[8] 齐玫玫.家庭医生团队对社区老年2型糖尿病患者的管理效果评价[J].慢性病学杂志,2020,21(11):140-142.

[9] 周蒨,董萍,陈真,等.全专团队模式下中西医结合康复对社区2型糖尿病认知功能障碍的效果评价[J].海南医学,2020,31(24):3188-3191.

[10] 陆惠珍,夏元旦,戚玉勤,等.家庭医生签约服务在农村社区2型糖尿病管理中的作用探讨[J].中国卫生产业,2020,17(11):115-117.