移动端体检报告影像识别及数据分析应用研究

孟彩霞,魏荣娟

(西安邮电大学 计算机学院,陕西 西安 710000)

摘 要:提出了移动端体检报告影像识别及数据分析方法。该方法通过移动端设备拍照的方式获取用户的体检报告图像,采用图像预处理、字符切分和识别等智能图像处理技术以及医学符号分析、数据库和数据可视化等技术,设计与实现了基于移动端的个人医疗健康管理原型系统。在对移动端体检报告影像识别方法的研究中,采用平均法进行彩色图像灰度化,利用中值滤波法去噪以及最大类间方差法实现二值化,改进了表格线检测及消除算法;字符切分法在传统字符切割前增加了字段切割,利用投影法加边界阈值判断法实现字段切割,对字符段再次利用投影法并结合基于识别的切分方法实现单字符分割;采用三次卷积插值算法进行归一化,最后利用模板匹配法完成识别。在对体检数据分析及可视化方法的研究中,受限于实验数据资源的不足,主要对识别结果的个体历史体检数据进行了简单对比分析,画出可视化图,最后结合实验结果给出健康建议。初步的实验结果证实该方法可以有效地提取体检结果文字信息,并将体检信息以图形化的方式呈现给用户,便于用户查看自己的健康数据变化,及时了解自身健康状况。

关键词:移动应用开发;影像文字识别;字符分割;数据可视化

中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2018)03-0187-05

doi:10.3969/j.issn.1673-629X.2018.03.040

Research on Application of Image Recognition and Data Analysis of Physical Examination Reports in Mobile Terminal

MENG Cai-xia, WEI Rong-juan

(School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710000, China)

Abstract: A method of image identification and data analysis for physical examination report on mobile terminal is proposed, which acquires the physical examination report image of users by taking pictures of mobile terminal device, and uses the intelligent image processing technologies like image preprocessing, character segmentation and recognition, as well as others like medical symbol analysis, database and data visualization for design and implementation of a prototype system for personal health management based on mobile terminal. In the study of the image identification methods for physical examination reports, we improve a table line detection and elimination algorithm through the mean method to transform the color image to gray one, denoising by the median filtering and binaryzation according to Otsu algorithm. Before traditional character segmentation, the character fields segmentation implemented by methods of projection and boundary threshold is added in character segmentation, and the single character segmentation is achieved by projection algorithm again in combination with segmentation algorithm based on recognition. We adopt the three convolution interpolation algorithm for normalization. Finally, the template matching is used to complete the recognition. In the study of physical data analysis and visualization, due to the limitation of finite data resources, we mainly analyze the individual historical physical examination data of the identified results and draw their visual maps. Finally the health advice is given by experiment. The preliminary results show that the proposed method can effectively extract the examination results of text information, and will present medical information graphically to the users, which is convenient for them to understand their own health data changes and know their own health status in time.

Key words: mobile application development; image character recognition; character segmentation; data visualization

收稿日期:2017-03-02 **修回日期**:2017-07-12 **网络出版时间**:2017-11-15

基金项目:陕西省自然科学基金资助项目(2014JM8303);陕西省教育专项科研计划资助项目(11JK0988);西安邮电大学研究生创新基金项目(CXL2015-39)

作者简介:孟彩霞(1966-),女,教授,研究方向为数据仓库和数据挖掘、算法分析与设计;魏荣娟(1989-),女,硕士,研究方向为移动视觉智能计算。

0 引 言

随着经济的发展及人民生活水平的提高,人们对自身健康越来越重视。当前我国人口老龄化和亚健康问题是受到广泛关注的社会热点问题。有报告显示,截至 2015 年底,我国 60 岁以上老年人口已达 2.22 亿,占总人口的 16.15%[1];另一方面,中国内地城市白领中有 76% 处于亚健康状态[2]。健康体检是了解受检者健康状况、早期发现疾病线索和健康隐患的重要方式,对老龄化和亚健康人群是非常必要的。目前城市中体检的理念已经深入人心,很多单位和个人每年都会进行相应的体检,目前已经出现很多运营成功的商业化健康体检机构,如爱康国宾体检中心、百岁啦、普惠体检中心等。

体检数据的管理与应用具有重要的社会和经济价值。体检结果的电子数据储存在体检机构,交给客户的是纸质版的打印报告。显然纸质报告具有不易保管、易丢失和破损等缺点,常年累积的体检报告还占用空间,不便于管理。另一方面,用户体检是一个长期行为,如果采用数据分析技术从用户多年体检数据中发掘出有用的信息用于指导个体健康极具应用价值。

移动互联网和移动智能终端的发展与普及推动了基于移动端的个人医疗健康管理应用的开发。文中对移动端进行健康体检报告影像识别及数据分析的方法进行了研究,将文字识别功能移植于智能手机上,利用手机便捷的照相机功能,获取病历报告的图像文件并识别,快速存储纸质介质信息;针对个人历史体检数据进行管理及数据分析和可视化并给用户提供相关的建议,指导个人健康生活。该研究工作具有重要的社会意义和应用价值。

1 系统需求与设计

1.1 需求分析

文字识别功能的主要技术为 OCR(optical character recognition)[3]。目前 OCR 技术应用已经相当成熟,市场上出现了不少相关 OCR 文字识别的应用软件,著名的有国外的 Abbyy、国内的汉王等,但大部分软件都只应用在计算机平台上或嵌入到扫描设备中,便捷性太低,加上扫描图像的繁琐,不能满足随时识别文字并记录的需求。OCR 技术在移动医疗领域的应用更是处于探索阶段,针对病历报告识别的软件屈指可数,仅有的几款软件如珍立拍、病历夹等,均存在识别版面不全、字符识别率低、识别结果易受环境影响、软件功能单一等问题。基于移动端的个人医疗健康管理系统可以解决传统大型扫描设备使用步骤繁杂和移动不便等窘境,及时存储个人医疗数据,发现潜在疾病风险,给移动生活和健康管理提供了极大的便利。

1.2 功能模块设计

基于移动端个人医疗健康管理系统框架如图 1 所示。

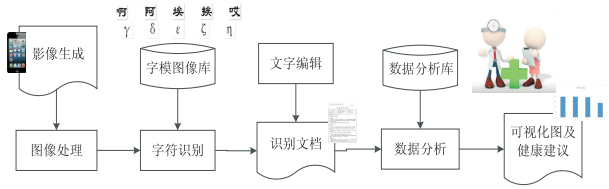


图 1 系统框架

系统主要分为 5 个模块:影像生成模块、图像处理模块、字符识别模块、文字编辑模块和数据分析模块,具体如下:

(1)影像生成模块。用户通过注册登录客户端进行图像采集,运用打开手机上的摄像头,对各种纸质病历报告、体检报告进行拍照,或是通过在已有的手机图片库中进行选择,获取需要处理的图片。

(2)图像处理模块。对获取到的图像文件进行图像处理,主要具有去噪、图像增强、图像旋转等功能,目的是提高文字识别率。

(3)字符识别模块。是系统的关键部分,对后续数据分析结果有直接影响。字符识别模块主要对获取的字符进行“翻译”,根据病历报告的分布特点首先进行行分割的字段“阅读”,再对每个字段内的单个字符进行列切割的逐字辨认、单字“翻译”。

(4)文字编辑模块。主要对 OCR 识别后的字符进行修改和编辑。系统自动查找可疑字,对认为有误的字符可以由用户进行文字编辑,实现人工校正功能。

(5)数据分析模块。使用 OCR 算法识别关键信息并储存到数据库中,通过识别结果对用户历年的个体数据进行管理和分析,给出直观的可视化图形,帮助用户发现疾病潜在风险并提出健康小建议。

2 OCR 技术的实现

2.1 图像预处理

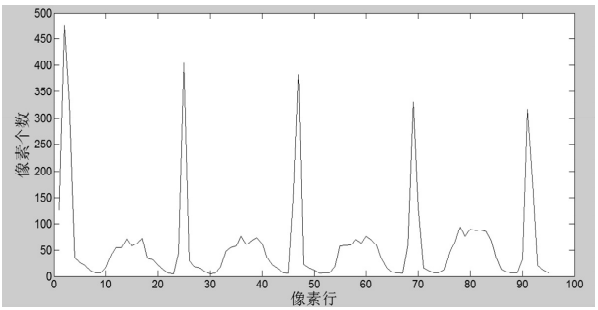
文中方法的第一步是移动端拍摄体检报告的预处理。预处理步骤包括采用平均法[4]进行彩色图像灰度化、中值滤波[5]去噪、最大类间方差法(Otsu 算法[6])进行二值化。病历报告中存在着大量表格线(见图 2(a)),影响文字识别结果。对于有表格内容的识别处理操作有两种:一是表格线中内容先提取,再进行识别;二是对表格线进行消隐,再对纯文本进行识别[7]。文中使用的体检报告表格线和表中字符基本无粘连,字符跨表现象少,所以采用第二种方法进行处理。

要实现表格线检测,首先要进行直线检测,传统的直线检测法有 Hough 直线检测算法[8]、投影法[9]等。由于 Hough 直线检测过程中容易丢失某些线段的端

点和长度信息,投影法又难以分割开表格线交叉处的直线,因此文中采用投影法结合经验阈值判断的方法进行表格线检测,采用二值化图像赋值法进行表格线的消隐。算法思路是首先对图像进行行投影,得到该图的水平投影直方图(见图 2(b)),图中有数个明显的波峰分别对应了报告单中的数条横向表格线,但由于横竖表格线之间有交叉,造成尖峰波峰之间有较小较宽像素的连接,因此,需要找个合适的分割位置对波峰进行切断操作,也即对表格线边界进行判断。

总胆红素	10.1	μmol/L		3.4-20.5
直接胆红素	4.7	μmol/L		0-6.8
间接胆红素	5.4	μmol/L		3.4-13.7
丙氨酸氨基移换酶	22	U/L		5-40

(a)体检报告表格线



(b)行投影图

图 2　表格线及其投影

表格线边界判断步骤如下:

Step1:求出所有波峰(包括尖峰和较宽峰)的平均高度 avgRow,寻找到最高波峰高度值 maxRow;

Step2:对平均高度 avgRow 放大一定的倍数 T_1,当表格线像素点的高度大于 T_1 倍平均高度时,即为波峰的一个极大点,记录其位置,找到所有尖峰位置 Row(i)= T_1 * avgRow。其中 T_1 为经验阈值,初始值为 T_1 = maxRow/avgRow。实验中可以对 T_1 进行适当调整,直到找到所有的尖峰高度值;

Step3:根据实验及经验知识可知,体检报告表格线交叉处像素点累计个数一般为个(十)位数,所以沿着每个尖峰最高位置 Row(i)分别向上、向下遍历寻找到最后一个个(十)位数为止,即为潜在表格线边界位置 topR(i)和 bottomR(i);

Step4:以 topR(i)和 bottomR(i)为行扫描上下边界,进行二值图像的赋值,将所有像素赋值 0,使目标像素点变为背景像素点,即实现行表格线的消除。

经过行表格线消除后的列表格线已经清晰地分开,可直接利用投影赋值法进行消除。

2.2　字符识别

2.2.1　字段切割

体检报告单中文字信息排版不规则,文字块相隔间距较大,为了便于将识别结果的相应体检项目、测量结果及单位存入数据分析数据库中,首先做一个字段分割处理,步骤如下:

Step1:通过行投影法对图像进行投影及行分割。

Step2:通过垂直投影法对每行进行列投影,以每个波峰的左右边界为起始点,分别向左、右 T_2 个长度进行遍历,若 T_2 个连续位置均无像素点,则此波峰左、右边界即为字段左、右边界,根据边界进行字段列切割;若左(右)边界 T_2 个长度内有任意一个像素点存在,则跳到下一个波峰的左(右)边界,继续遍历寻找,直到找到要求的边界为止。这里 T_2 为经验值,文中取 T_2 = 15,切割结果如图 3 所示。

体检项目	测量结果	单位
总胆红素	10.1	μmol/L

(a)生化检验(总胆红素)项目

(b)字段切割结果

图 3　字段切割示例图

2.2.2　字符切割及归一化

当前字符切分技术主要有以下几种方法[10-11]:基于图像分析的切分、基于识别的切分、综合切分及整体识别切分,文中采用投影法结合经验值的试切分方法对文字进行切割。针对切割后的每一个字段重新进行投影分割,对于少量不能被正确分割出来的字符,如图 4 中的“mo”,采用基于识别的切分方法,给予 1/2,2/3……一系列经验值的切分位置进行试切分,送入模板库进行匹配,最终找到切分位置并识别出结果。

由于切割后的字符大小不一,为了便于和字模库中的字模进行匹配,需要进行归一化处理。常用的归一化方法主要有两种:分裂合并归一化与插值变换归一化[12]。文中的字符经过分割后像素偏小,需要对图像进行放大,所以选用第二种方法,采用三次卷积插值算法[13],将字符尺寸归一化到 48×48 点阵,便于与字模数据库进行比对,提高识别率。

图 4　字符归一化结果

2.2.3　字符识别

目前,用于字符识别的算法主要有基于模板匹配的字符识别算法[14]、特征统计匹配法[15]和基于神经网络的字符识别算法[16]。特征统计匹配法是提取待识别模式的一组统计特征,通过按一定准则确定的决策函数进行分类判别。在字符识别中常用的特征方法是网格特征匹配法。但是实际应用中,由于外部原因常会出现字符模糊、倾斜等情况,影响识别效果。因而,此方法实际应用效果不理想,鲁棒性不强。模板匹配方法体现的是字符的整体特征,它比特征统计匹配

法更有效。由于医疗体检报告字体是印刷体,结构标准,所以采用模板匹配法进行识别。文中实验方案暂时不考虑分类器方法[17],原因是分类器方法涉及到大量的训练,针对手写字符、变体字符及多字体识别时比较有效。

为了保证识别结果的正确性,系统做了两方面的保证:一是识别结果提供人工校正的编辑接口;二是对比数据库的建立。系统中预先已经建立了一个常用医学符号单位库,数据库中的内容包含了所有预先识别的医学名称、单位、符号等字集,方便识别结果进入库中进行字符串比对,确保识别结果的正确性及权威性。

2.3 数据分析及可视化

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析,提取有用信息并形成结论而对数据加以详细研究和概括总结的过程。在实用中,数据分析可帮助人们做出判断,以便采取适当行动。实验中,数据分析模块采用描述性统计方法[18]进行,对不同用户的历史体检报告进行了识别,对同一用户不同时期的体检数据进行了统计和简单分析,画出不同项目的识别结果对比可视化图形,根据分析结果给出健康建议。

3 系统测试

3.1 测试环境

(1)数据来源。

选用了某大学 10 位教师在某商业体检公司中近三年体检纸质报告作为研究对象(已获得本人同意)。实验中分析了体检报告中的基本信息表、身高体重血压表、血常规表和生化检验表,分不同时期(以近期 5 个时间段为准)共计 160(10×3×5+10)张测试样本数据集。

(2)实验数据。

(A)图像数据库。

收集 160 张标准体检报告单,对每张报告单采用高清像素(苹果 800w)、普通像素(华为 1 300w)和劣质像素(天语 800w)的三级拍照模式,在每一级模式下又采用自然光、白炽灯两种不同的光照进行拍照,样本总库达到 320(160×2)张样本图像,完成图像数据库的建立。

(B)字模数据库。

采集国家标准汉字 6 763 个(国标一、二级字库),英文字母 52 个(大、小写),医学特殊符号 100 个,数字 10 个(0 ~9),累计 6 925 个字符,采用自己编写的字模提取工具软件,构建英文、数字、汉字及医学特殊符号字模数据库。

(C)数据分析库。

统计常用体检报告医学单位符号 50 个,存入标准单位库 D_Stand 表中;从手机客户端录入用户的 ID 号、姓名、密码,从基本信息表识别后的结果中提取用户的姓名、性别、年龄、手机号码、工作单位、体检日期,构建客户信息表 D_Users;从身高体重血压表识别结果中提取身高、体重、血压等构建身高体重血压表 D_HWB;从生化检验表识别结果中提取总胆红素、血蛋白等数据,构建生化检验表 D_Bioc;从血常规报告识别结果中提取白细胞数、淋巴细胞比值等数据,构建血常规表 D_Blood,完成数据分析库的建立。

(3)实验环境。

所有实验使用的系统软硬件环境均相同。CPU 为 Intel(R)® Core(TM)® i5-4210M 双核(64 位处理器),2.59 GHz,内存 4 G;操作系统为 Windows® 10 中文版(64 位系统);实验程序用 Matlab® R2010b 编写;原型 APP 系统的开发系统是 Android 4.4.2 版本。

3.2 实验结果与分析

实验结果分为两部分,一是识别正确率:计算每一幅图像中文字和符号的识别正确率,最后求出平均值,即为体检报告整体识别率。具体做法是:针对每幅图像,事先编辑一个文本文件,储存各个字符正确的识别结果,最后将经过算法识别的结果和这个文件相比较,计算出最终的识别正确率,如表 1 所示。二是数据分析结果,画出不同项目的识别结果对比可视化图形(见图 5),给出参考健康建议。

表 1 体检报告识别结果

模板匹配	识别率	平均识别率
汉字	94.20%(3 768/4 000)	94.67%
英文字母	95.66%(3 348/3 500)	
数字	94.83%(1 138/1 200)	
医学符号	94.00%(47/50)	

(a)登录界面 (b)识别界面 (c)可视化界面

图 5 APP 部分界面展示

此外,在上述算法基础上,文中的一个重要工作是开发了原型 APP 系统(图 5),系统界面简洁、使用方便,帮助用户及时存储自己的疾病信息,了解自身健康状态,使用户的生活更健康、有质量。

4　结束语

基于移动端个人医疗健康管理系统,用户只需要一款简单的手机,就能轻松实现快速存储病历信息,了解健康状况,通过系统建议引导自身健康生活。系统使用方便,应用前景广阔。

该研究虽取得了一些有益的成果,但仍存在以下缺点和不足:针对倾斜角度太大的病历报告识别效果不好;实验用户数据库数据太少,能挖掘的信息较少。下一步的研究方向是多种因素干扰下的病历报告识别及通过此 APP 累积大量用户的数据,可以采用数据挖掘以及机器学习的算法建立用户健康模型,对个体健康提供相应的建议与指导。

参考文献:

[1] 张锦莉. 当前农村养老保障问题探析[J]. 人大建设,2016(12):48-50.

[2] 许鲁平. 健康类电视节目的公信力研究[D]. 济南:山东师范大学,2014.

[3] SHAH P,KARAMCHANDANI S,NADKAR T,et al. OCR-based chassis-number recognition using artificial neural networks[C]//IEEE international conference on vehicular electronics and safety. [s. l.]:IEEE,2009:31-34.

[4] 张　凤. 街景影像的文字识别[D]. 北京:北京建筑工程学院,2012.

[5] 洪　涛,梁伟建,卢玉凤. 标牌粘连字符自适应定位分割重建与识别[J]. 中国图象图形学报,2014,19(6):886-895.

[6] OTSU N. A threshold selection method from gray level histograms[J]. IEEE Transactions on Systems Man & Cybernetics,1979,9(1):62-66.

[7] 谢　亮. 表格识别预处理技术与表格字符提取算法的研究[D]. 广州:中山大学,2005.

[8] 滕今朝,邱　杰. 利用 Hough 变换实现直线的快速精确检测[J]. 中国图象图形学报,2008,13(2):234-237.

[9] 刘　昱. 印刷体表格识别的研究[D]. 哈尔滨:哈尔滨工程大学,2013.

[10] CASEY R G,LECOLINET E. A survey of methods and strategies in character segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,1996,18(7):690-706.

[11] LU Y. On the segmentation of touching characters[C]//International conference on document analysis and recognition. [s. l.]:IEEE,1993:440-443.

[12] 万金娥,袁保社,谷　朝,等. 基于字符归一化双投影互相关性匹配识别算法[J]. 计算机应用,2013,33(3):645-647.

[13] 王　帅,冯　晋. 基于三次卷积插值的贝叶斯滤波方法研究[J]. 系统科学与数学,2015,35(2):170-180.

[14] 陈丽芳,刘　渊,须文波. 改进的归一互相关法的灰度图像模板匹配方法[J]. 计算机工程与应用,2011,47(26):181-183.

[15] 罗辉武,唐远炎,王　翊,等. 基于结构特征和灰度特征的车牌字符识别方法[J]. 计算机科学,2011,38(11):267-270.

[16] 杨天长. 基于神经网络的文字识别技术研究及应用[D]. 北京:北方工业大学,2013.

[17] 陈　文,张恩阳,赵　勇. 基于多分类器协同学习的卷积神经网络训练算法[J]. 计算机科学,2016,43(9):223-226.

[18] 王　曼. 医学论文统计描述性数据审核的问题与方法[J]. 中国科技期刊研究,2015,26(4):359-362.

(上接第 186 页)

参考文献:

[1] 魏　莉. 报文传输业务中的常见问题及解决方法[J]. 气象研究与应用,2007,28:111-112.

[2] TAI S,MIKALSEN T A,ROUVELLOU I. Using message-oriented middleware for reliable web services messaging[C]//International workshop on web services,e-business,and the semantic web. Berlin:Spring,2003:89-104.

[3] 黄美林,马建华,李　东. 基于 SSH 框架与泛型的通用分页方法设计与实现[J]. 计算机技术与发展,2012,22(1):67-71.

[4] 付更丽,曹宝香. SOA-SSH 分层架构的设计与应用[J]. 计算机技术与发展,2010,20(1):74-77.

[5] 周杨川,孙淑霞,丁照宇. 基于 Spring+JPA 框架的电子政务基础平台[J]. 计算机技术与发展,2008,18(4):98-100.

[6] 鲍婷婷,陈　鹏,殷笑茹. 省级气象资料监控业务系统设计与实现[J]. 气象水文海洋仪器,2014,31(3):104-106.

[7] 曹　威,刘　洁,杨维发,等. 湖北省气象信息传输监控与处理平台的设计与实现[J]. 电子技术与软件工程,2016(18):68-69.

[8] 朱　勍,唐　民. 民航气象观测报文监控和反馈系统研究[J]. 中国民航飞行学院学报,2015(2):77-80.

[9] 孙周军,肖文名,宋远清,等. 气象信息实时监视系统改进设计与实现[J]. 成都信息工程学院学报,2012,27(2):168-173.

[10] 朱　璇,马少妆. 常规气象观测报文传输监控操作过程分析[J]. 科技创新导报,2010(7):119.

[11] 叶汶华. 气象数据监控系统的设计与实现[J]. 电子技术与软件工程,2016(6):185.

[12] 钟　静,李　赟,陈海涛,等. 基于 SMS 技术的气象报文监控设计与实现[J]. 贵州气象,2011,35(4):42-43.

[13] FERNANDEZ I. Beginning oracle database 11g administration:from novice to professional[M]. [s. l.]:Dreamtech Press,2009:53-59.

[14] 王荣生,杨际祥,王　凡. 负载均衡策略研究综述[J]. 小型微型计算机系统,2010(8):1681-1686.

[15] VUKOTIC A,GOODWILL J. Apache Tomcat 7[M]. [s. l.]:Apress,2011.