

一种基于连通域搜索的英文单词切分方法

尹芳^{1,2}, 李全通¹, 周昕¹, 金飞虎¹

(1. 哈尔滨理工大学 计算机科学与技术学院, 黑龙江 哈尔滨 150040;

2. 哈尔滨理工大学 仪器科学与技术博士后科研流动站, 黑龙江 哈尔滨 150040)

摘要: 英文文本识别系统中单词切分效果直接影响系统识别准确率. 提出一种有效利用单词中的连通域以实现单词切分的方法, 该方法首先对给定的图像文件进行扫描, 搜索图像中所存在的所有连通域, 然后根据连通域的坐标、尺寸等影响因素, 对有效的连通域进行合并, 并过滤掉噪声等因素所产生的连通域碎片, 以实现有效的英文单词区域的正确分割, 从而实现英文单词的正确切分. 实验结果表明, 针对相同的英文文本图像, 该方法对英文单词切分的准确率达 95% 以上, 与垂直投影切分法最高准确率 27% 相比有较大幅度的提高, 证明该方法可以很好解决垂直投影法无法解决的字母之间相互交迭的单词切分问题.

关键词: 连通域搜索; 单词切分; 文本识别

中图分类号: TP391 **文献标志码:** A **文章编号:** 1007-2683(2014)05-0109-04

A Connected Domain Research Method for English Word Segmentation

YIN Fang^{1,2}, LI Quan-tong¹, ZHOU Xin¹, JIN Fei-hu¹

(1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150040, China;

2. Instrument Science and Technology Postdoctoral Research Station, Harbin University of Science and Technology, Harbin 150040, China)

Abstract: The word segmentation effect of the English text recognition system directly affects the system identification accuracy. This paper puts forward a connected domain search based word segmentation method, which firstly search whole image to record all the connected domain, then merge connected domain according to the location, size of connected domain. The experimental result shows that the method can better solve the word segmentation with overlapping between the letters than vertical projection method

Key words: connected domain; word segmentation; text recognition

0 前言

文档识别目前已广泛应用于办公自动化、文档处理、文档排版、信息类型转换等应用领域, 为人们的工作生活带来极大便利, 是模式识别领域的一个

主要研究方向. 文档识别技术的应用极大提高了工作效率和准确率, 同时节省了大量人力资源, 可以带来巨大的经济、社会效益, 因此文档识别技术的研究是一个具有重要现实意义的工作内容.

文档识别系统的处理流程可大致描述为^[1]: 首先获取待识别文档对象图像, 对其进行二值化、去白

收稿日期: 2014-04-15

基金项目: 黑龙江省教育厅科学技术研究项目(125411119).

作者简介: 尹芳(1978—), 女, 副教授, 博士, E-mail: yinfang@hrbust.edu.cn;

周昕(1968—), 女, 副教授;

金飞虎(1973—), 男, 副教授.

边、去噪声的预处理;然后进行切分,得到独立识别对象,中文文档以汉字为识别对象,英文数字则要切分得到单独的字母和数字作为识别对象;然后对其进行去白边、去噪、归一化等预处理;处理后提取特征,送入分类器分类得到初识结果;最后对其进行后处理校正初识结果,得到最终识别答案^[2]。

英文文档识别系统以字母为研究对象,因此文档切分步骤需要首先将文档进行行列切分得到单词,然后对单词进行切分得到字母,才可继续后续的识别处理工作。单词切分算法效果好坏将直接影响切分得到字母的独立性和完整性,是英文文档识别系统中的一个重要步骤,因此要使用高效高准确率的单词切分算法,以保证后面的字符识别可以顺利进行。

目前研究者常用的文本切分方法包括垂直投影法及其改进方法^[3-5],曲线切分路径法^[6],集成切分法^[7],聚类法^[8-10]及识别反馈法^[11]。其中垂直投影法利用水平方向投影直方图实现文本切分,算法简单,实现方便,但是对于单词中字母之间存在的上下搭接现象,直方图无法正确处理;曲线切分路径法在垂直投影法直线切分的基础上进行改进,采用探索曲线路径实现切分,可以有效解决存在字母搭接的单词切分问题,但是探索曲线路径的过程复杂度高,效率较低;集成切分法主要适用于粘连字符切分,首先搜索连通域得到字母或字母组合的子模块,然后分析各连通域的宽高比,对确认为字母组合的连通域进行进一步的切分,该方法可以有效处理粘连字符识别,但算法复杂度过高,计算量大;聚类法主要适用于无结构、背景复杂的彩色场景文本识别;识别反馈法基于前端预测识别来分割粘连字符,根据粘连字符图像的特征初步确定前端字符的候选字符集,然后根据候选答案与前端字符图像匹配的必要条件做进一步筛选,该方法算法复杂,效率较低,适用于低质量的粘连字符识别。总之以上方法或者无法处理搭接字符切分问题,或者算法复杂、执行效率低,对于排版规范、质量较好的印刷体文本识别均不很适用。

印刷体英文文本大多字形规范,排列工整,切分比较容易实现。但是由于英文字符的形状特点决定了字母之间存在大量搭接现象,而统计获知英文OCR系统中识别错误的原因多源自字符搭接引起的误切分^[12-13]。所谓字符搭接^[14]是指两个或两个独立的字符虽然没有直接粘连,但是在空间上存在空间交叠的部分,如图1所示。把一个独立的单词切

分得到几个独立的字符因误切分而合在一起送往识别器,势必造成识别结果的不正确,如对图2(a)单词进行切分得到图2(b)的切分结果,其中fe两个字母由于存在字母搭接,被切分成一个字母。因此搭接字符切分的正确性将直接影响识别系统的识别准确率。

ff tr tu
leftwards from
opposite physical
feasible

图1 存在搭接的单词及字母组合样本示例

feasible

a)feasible 原始图像

feasible

b)feasible 切分结果

图2 垂直投影法切分效果示意图

连通域搜索方法目前也被广泛应用于图像分割领域,比如利用连通域进行文档切分^[15],去除噪声处理^[16],在车牌识别中作为描述特征^[17-18],图像标记^[19-20]等,均取得较好效果。本文在传统基于垂直投影直方图的切分方法基础之上,提出一种有效的利用单词中的连通域实现切分的方法。该方法首先扫描图像,搜索存在的连通域,然后根据坐标、尺寸等影响因素对连通域进行合并,可有效解决存在字母搭接的单词切分问题。

1 连通域

连通域搜索是进行图像分割的主要工具,定义可描述为:若任选一个集合A为连通域则其中任意两个点,都可以在A中找到一条折线将其连接。连通域又可分为单连通区域和多连通区域^[21]。在连通区域D中,若D中的任意一条封闭曲线,都可以不经过D外的点而连续收缩于D中的某一点,则称D为单连通区域,如图3(a)所示;否则,则称D为多连通区域,如图3(b)所示。多连通域也称为复连通区域。

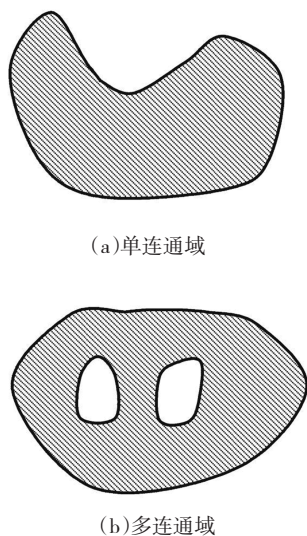


图3 连通域示意图

数字图像处理的处理对象为离散的点阵信息,对于离散信息而言,根据搜索方向的不同,可分为四连通域和八连通域两种类型. 对于一个图像区域,若从其中的任意一点出发,通过上、下、左、右四个方向能到达图像的任意一点,则称该图像区域为四连通域;若从图像的任意一点出发,通过上、下、左、右、左上、右上、左下、右下八个方向能到达图像的任意一点,则称该连通域为八连通域. 如图4,以图像中的像素点P为研究对象,则与P上、下、左、右四个方向相邻的四个像素P0, P2, P4, P6称为P的四邻域;与P上、下、左、右、左上、右上、左下、右下八个方向相邻的8个像素P0~P7称为P的八邻域.

P3	P2	P1
P4	P	P0
P5	P6	P7

图4 邻域示意图

2 基于连通域搜索的英文单词切分算法

本文采用生长算法实现连通域搜索,生长算法主要考虑像素及其空间相邻像素之间的关系,确定一个或多个作为种子的像素点后以像素点的灰度值为相似性度量准则增长区域,逐步生成具有某种均匀性的空间区域,直至没有可以归并的点为止.

算法实现

假设待处理图像的宽度、高度分别为 $mWidth$ 、 $mHeight$, 图像中每个点的处理状态用数组 $image$ $[mWidth]$ $[mHeight]$ 标记, 0 表示对应像素点未处理, 1 表示对应像素点已处理, 初始值全部为 0. 变量 $domainnum$ 表示连通域的序号, 初始值为 0. 设置结构体用来存储切分结果, 结构如下:

```
typedef struct CutProperty{
    int initposx;
    int width;
    int initposy;
    int height;
    BYTE * Matrix;
    struct CutProperty * pLink;
} CUTPROPERTY;
```

其中 $initposx$ 和 $initposy$ 分别用来存储连通域的起始坐标, $width$ 和 $height$ 分别用来存储连通域的宽度和高度, $Matrix$ 用来存储每个连通域中的图像像素信息. 连通域搜索过程描述如下:

选择图像中任意一个像素点, 设其坐标为 (x, y) , 首先根据式(1)判断该点的状态.

若判定该点为未处理前景像素点, 则将其入栈, 然后令

$domainnum = domainnum + 1$
 $domain[x * mWidth + y] = domainnum$
 否则, 不做任何处理.

$$(x, y) = \begin{cases} \text{背景点} \\ \left\{ \begin{array}{l} domain[x * mWidth + y] = 0 \wedge \\ image[x * mWidth + y] = 0 \end{array} \right. \\ \text{已处理前景点} \\ \left\{ \begin{array}{l} domain[x * mWidth + y] \neq 0 \end{array} \right. \\ \text{未处理前景点} \\ \left\{ \begin{array}{l} domain[x * mWidth + y] = 0 \wedge \\ image[x * mWidth + y] \neq 0 \end{array} \right. \end{cases} \quad (1)$$

依次对所有进入堆栈的像素点进行下述处理:

取栈顶元素 a , 搜索该元素的八邻域点 x_i ($i = 1, 2, \dots, 8$), 若 $image[x_i] = 0$, 则为背景, 不处理; 若 $image[x_i] = 1$, 则 x_i 与 a 属于同一个连通域, 同上处理, 令

$domain[x * mWidth + y] = domainnum$, 然后将 x_i 入栈.

循环执行上述操作, 直到图像中所有点都处理完毕.

以上算法的处理流程逻辑描述如图5所示.

初次搜索可能会将一个字母分成多个连通域,因此需要对其进行后处理,完成连通域的合并. 设图像高度为 $mHeight$, 连通域高度为 $height$, 根据单词中字母之间高度差的关系, 若 $height < \frac{mHeight}{3}$, 则考虑将它与附近连通域合并或者认定为噪声, 忽略处理. 根据该连通域与其距离最近的连通域之间的交叠距离 $Distance$ 判定具体处理方式. 假设两个连通域的横坐标分别为 (x_1, x_2) 、 (y_1, y_2) , 则 $Distance$ 定义如式(2)所示.

$Distance$ 标示两连通域的搭接程度. 设定一个阈值 r , 若 $Distance > r$ 则将两连通域合并, 否则认为该连通域为噪声, 忽略.

由于印刷质量问题有可能出现字母竖直断裂的现象, 这样会产生与字母 i 、 l 、 l 非常相似的、宽高比非常小的碎片. 这时可以通过分析它占据位置来区分: 碎片一般占据位置比较居中, 介于上下基线之间, 而字母占据位置靠中上, 介于顶部边线和基线之间. 若判断为碎片则进行合并.

$$Distance = \begin{cases} |x_2 - y_1| (y_1 > x_1) \wedge (y_2 > x_2) \\ |y_2 - y_1| (y_1 > x_1) \wedge (y_2 < x_2) \\ |y_2 - x_2| (y_1 < x_1) \wedge (y_2 < x_2) \\ |x_2 - x_1| (y_1 < x_1) \wedge (y_2 > x_2) \end{cases} \quad (2)$$

使用本文提出的基于连通域搜索的英文单词切分方法对图 5(a) 中的 *feasible* 单词进行切分, 得到结果如图 5(b) 所示. 从图中可以看出, 基于连通域的切分方法可以将存在上下交迭的 *f*、*e* 两字母切分开, 较之垂直投影切分法有效.

feasible

(a) *feasible* 原始图像

feasible

(b) *feasible* 连通域切分结果

图 5 连通域切分示意图

3 实验结果及分析

随机选择 300 个单词及字母组合作为实验对象, 尽量包含存在字符搭接现象的试验样本, 将其平均分成 3 组, 每组 100 个单词及字母组合样本, 分别

采用基于垂直投影直方图的切分方法和本文提出的基于连通域搜索的切分方法两种方法对其进行单词切分, 然后送入同一识别引擎进行识别, 以检验本文提出方法的切分效果.

该实验将本文方法与垂直投影切分法在同一识别平台上进行比较, 二者处理流程中仅单词切分算法不同, 其他预处理、特征提取及分类器分类等步骤均采用相同方法, 以比较两种切分方法的效果. 最终识别正确率如表 1 所示:

表 1 切分结果比较

单词总数/个	投影法/%	本文方法/%
100	18	97
100	27	98
100	21	95

从表 1 中数据可以看出, 各组样本中基于本文提出的切分方法的系统识别准确率要远远高于基于投影切分方法的系统识别准确率. 之所以存在如此大的差距, 主要原因在于本文方法对于英文单词中普遍存在的字母搭接现象处理的有效性, 同时证明了本文提出的基于连通域搜索的单词切分方法是有效的.

4 结 语

本文所提出的基于连通域搜索的英文单词切分方法, 解决了垂直投影方法无法解决的单词上下交迭的问题, 可以得到更准确的切分结果. 通过图像文件的扫描, 搜索图像中所存在的所有连通域, 然后根据连通域的坐标、尺寸等影响因素, 对连通域进行合并和过滤, 以实现有效的英文单词区域的正确分割, 从而实现英文单词的正确切分. 实验结果表明, 针对相同的英文文本图像, 该方法对英文单词切分的准确率达 95% 以上, 与垂直投影切分法最高准确率 27% 相比有较大幅度的提高, 证明该方法可以很好解决垂直投影法无法解决的字母之间相互交迭的单词切分问题, 是一种切实可行的单词切分方法, 为文本识别的后续工作奠定了良好的基础.

参 考 文 献:

- [1] 尹芳, 王卫兵, 陈德运. 印刷体英文文档识别技术 [J]. 哈尔滨理工大学学报, 2008, 13(6): 9-12.
- [2] 尹芳, 陈德运, 吴锐. 改进的 Gabor 小波变换特征提取方法 [J]. 计算机工程, 2012, 38(15): 145-147.

(下转第 119 页)

- Graw-Hill Book Company, 1986: 133 – 161.
- [4] 屈宁, 姚守献. 电机工程中的坐标变换 [J]. 哈尔滨电工学院学报, 1995, 18(3): 291 – 297.
- [5] LEE R J, PILLAY D, HARLEY R G. Reference Frame for The Simulation of Induction Motor [J]. Elect Power Sys Research, 1984(8): 233 – 242.
- [6] 田铭兴, 励庆孚, 王曙鸿. 交流电机坐标变换理论的研究 [J]. 西安交通大学学报, 2002, 36(6): 568 – 572.
- [7] 蔡骏, 邓智泉. 基于相电感综合矢量法的开关磁阻电机初始位置估计 [J]. 中国电机工程学报, 2013, 33(12): 145 – 151.
- [8] 刘剑, 杨贵杰, 高宏伟, 等. 双三相永磁同步发电机的矢量控制与数字实现 [J]. 电机与控制学报, 2013, 17(4): 50 – 57.
- [9] PELLEGRINO G, BOJOI R, GUGLIELMI P. Unified Direct – flux Vector Control for AC Motor Drives [J]. IEEE Transactions on Industry Applications, 2011, 47(5): 2093 – 2102.
- [10] TICHIT P H, BUROKUR S N, WU X, et al. Application of Co-ordinate Transformation for Novel Antenna Design Techniques [C] // 2012 European Conference on Antennas and Propagation, 2012: 1711 – 1715.
- [11] 衣晓, 何友, 关欣. 一种新的坐标变换方法 [J]. 武汉大学学报, 2006, 31(3): 237 – 239.
- [12] 张华强, 王新生, 魏鹏飞, 等. 基于空间矢量调制的直接转矩控制算法研究 [J]. 电机与控制学报, 2012, 16(6): 13 – 18.
- [13] 王晶鑫, 姜建国. 矩阵变换器驱动感应电机转矩和转子磁链控制 [J]. 电机与控制学报, 2010, 14(7): 59 – 65.
- [14] 王松, 谈龙成, 李耀华, 等. 两相坐标系下链式 STATCOM 电流跟踪策略 [J]. 电机与控制学报, 2014, 18(1): 19 – 25.
- [15] 王德成, 林辉. 一种基于支持向量机的矢量控制方法及应用 [J]. 电机与控制学报, 2012, 16(5): 52 – 56.
- [16] 陶兴华, 李永东, 孙敏. 一种基于同步旋转坐标变换的单相锁相环新算法 [J]. 电工技术学报, 2012, 27(6): 147 – 152.
- [17] 唐酿, 肖湘宁, 盛超等. 矩阵变换器新型间接空间矢量调制算法 [J]. 电工技术学报, 2013, 28(2): 262 – 269.
- [18] 宋文胜, 冯晓云, 谢望玉. 单相三电平整流器 d – q 坐标系下的控制与 SVPWM 方法 [J]. 电机与控制学报, 2012, 16(4): 56 – 63.
- [19] 刘锦波, 陈永校. 电机统一理论中功率不变的坐标变换系统 [J]. 浙江大学学报, 1999, 33(6): 679 – 686.
- [20] 张永德, 姜金刚, 唐伟, 等. 基于遗传算法的直角坐标式排牙机器人路径规划 [J]. 哈尔滨理工大学学报, 2013, 18(1): 22 – 26.
- [21] 高景德, 王祥珩, 李发海. 交流电机及其系统的分析 [M]. 北京: 清华大学出版社, 2005: 318 – 327.

(编辑: 关 毅)

(上接第 112 页)

- [1] 陈涛, 杨晨晖, 青波. 基于投影和固有特征结合的车牌字符分割方法 [J]. 计算机技术与发展, 2009, 19(5): 45 – 47.
- [2] 王江晴, 曹卫. 基于极大阈值和曲线拟合的垂直投影汉字切分 [J]. 中南民族大学学报: 自然科学版, 2011, 30(4): 83 – 85.
- [3] 焦蓬蓬, 郭依正. 一种基于数学形态学的车牌定位与分割方法 [J]. 自动化技术与应用, 2013, 32(6): 57 – 59.
- [4] WANG J, JEAN J. Segmentation of Merged Characters by Neural Networks and Shortest Path [J]. Pattern Recognition, 1994, 27(5): 649 – 658.
- [5] 杨武夷, 张树武. 一种视频中字符的集成型切分与识别算法 [J]. 自动化学报, 2010, 36(10): 1468 – 1476.
- [6] 吴锐, 尹芳, 唐降龙等. 模糊逻辑和谱聚类的字符图像分割 [J]. 哈尔滨工业大学学报, 2010, 42(2): 268 – 276.
- [7] 尹芳, 陈德运, 吴锐. 改进的谱聚类图像分割方法 [J]. 计算机工程与应用, 2011, 47(21): 185 – 187.
- [8] 尹芳, 吴锐, 陈德运, 于晓洋. 基于谱聚类的改进的文本图像分割方法 [J]. 高技术通讯, 2013, 23(10): 1024 – 1029.
- [9] 张振绘, 刘赛. 女书文字切分算法的设计与实现 [J]. 中国科技信息, 2010, 12: 119 – 120.
- [10] 李佐, 王妹华, 蔡士杰. 一种基于前端预测识别的粘连字符分割方法 [J]. 计算机研究与发展, 2001, 38(11): 1337 – 1344.
- [11] LU Yi. Machine Printed Character Segmentation: an Overview [J]. Pattern Recognition, 1995, 28(1): 67 – 80.
- [12] 刘阳兴. 基于折线切分路径的粘连搭接字符切分算法的研究 [J]. 计算机应用研究, 2001, 28(10): 3998 – 4000.
- [13] 于明, 郭金, 王栋壮, 等. 改进的基于连通域的版面分割方法 [J]. 计算机工程与应用, 2013, 49(17): 195 – 198.
- [14] 刘渝, 张彦铎, 鲁统伟. OCR 中一种基于最小一乘法的连通域去噪方法 [J]. 武汉工程大学学报, 2011, 33(1): 84 – 87.
- [15] 王渊, 付勇, 盛翊智. 复杂图像中基于综合特征的车牌定位方法 [J]. 自动化技术与应用, 2006, 25(6): 54 – 56.
- [16] 孟庆远, 白艳萍, 胡红萍. 基于字符连通域特征的车牌汉字字符识别技术 [J]. 测试技术学报, 2011, 25(1): 87 – 92.
- [17] 谢宜壮, 谭许彬, 陈禾. 一种新的连通域标记算法 [J]. 北京理工大学学报, 2012, 32(12): 1273 – 1278.
- [18] 梁成. 基于像素二值化法的连通域标记 [J]. 计算机与数字工程, 2013, 41(5): 823 – 826.

(编辑: 关 毅)