工程图纸图像图文自动分割工具 SegChar*

江 早 刘积仁 刘晋军

(东北大学软件中心 沈阳 110006)

摘要 文章分析了工程图纸图像图文分割的技术特点、关键步骤和基本框架,着重介绍了图文自动分割工具 SegChar 采用的技术,如:(1) 自动字符尺寸阈值过滤技术,可使图文分割过程自动化和智能化;(2) 任意方向、任意长度字符串检测技术,通过精确 HOUGH 空间寓求、松弛共线、基于字符串的 HOUGH 域更新等策略,提高了字符分割的处理速度,降低了处理的空间复杂度,能够使复杂的中西文字符串得以完整提取.文章最后给出了性能评价.

关键词 图纸图像处理,图文分割,HOUGH 变换,自动阈值.

中图法分类号 TP391

工程图纸中的图形和文字在逻辑上具有不同含义和功能,通常图形部分采用矢量化方法进行处理,而字符部分采用 OCR(optical character recognition)技术进行识别.图文分割可以提供基本的待识别单元,是进一步进行文字符号识别、图形矢量化、图纸检索乃至图纸理解的基础.

图文分割与字符识别是密切相关的.通常的做法是:首先基于字符的某些共同特征进行字符的提取,然后送入 OCR 识别器,最后利用识别的结果进行校正.其中的字符提取可以称为一种粗分割.由于字符具有某些公共特征,即使脱离 OCR 过程,粗分割仍然可以获得较好的分离结果,本文的图文分割也是以此为依据的.其目的在于:对于西文,找到图纸中用于注释的字符串、组成字符串的单个字符;对于中文,找到单个的汉字及汉字所在的短语或句子.关于字符的识别,留给 OCR 部分处理.

工程图纸的处理属于光栅文档处理的一个子领域.对于普通光栅文档的分析,已经存在大量方法进行文档的结构分割、字符提取,典型的方法有 RLSA 等[1].这类方法的特点在于,基于已知的文档编排特性,自顶向下进行分析.先提出文字块、图形块,进一步得到字符行、字符串直至单个字符.

工程图纸中的字符通常存在以下几个特点[2~4]:(1) 属于手写体范畴,在同一张图纸中,大小不统一;(2) 位置分散、方向不一;(3) 部分字符与图形在尺寸及拓扑模式上相似;(4) 部分与图形粘连;(5) 经常存在噪声干扰.因此,字符提取和识别的难度比普通光栅文档更大,不能采用传统的处理方法和步骤.必须考虑处理方法对字符尺寸变化的适应性、处理任意字符串方向的能力、抗噪声干扰性能以及区分字符与特殊图形的能力.根据工程图纸的特点,通常采用一种与普通光栅文档处理过程相反、自底向上的处理方法,即首先找到单个字符,甚至字符的某些部分,然后通过组合生成字符及字符串,进一步生成文字段等[3].

对于工程图纸中字符的处理,目前主要有基于轮廓的方法[5]、基于连通检测的方法[2]、基于区域增长的方法[6]和基于聚类的方法[7],其中文献[2]提出的方法揭示了图纸字符处理的一种典型过程.该方法有以下几个关键步骤:连通体的标识、字符过滤器的设计、字符串的成组以及后处理.

本文采用了文献[2]中提出的处理基本框架.为了实现系统整体较高的性能,SegChar 特别采用了如下几项技术:(1) 快速连通体标识技术;(2) 连通目标区域树存储技术;(3) 字符尺寸自动阈值技术;(4) 噪声自动过滤

* 本文研究得到国家"九五"科技攻关项目基金和中国博士后基金资助.作者江早,1965年生,博士后,副教授,主要研究领域为图像图形处理,CAD/CAM 及数控技术.刘积仁,1955年生,博士,教授,博士生导师,主要研究领域为分布式多媒体信息处理,图像图形处理,协议工程.刘晋军,1970年生,工程师,主要研究领域为工程图纸自动输入处理.

本文联系人:江早,沈阳 110006,东北大学软件中心 308 信箱

本文 1997-10-20 收到原稿,1998-09-02 收到修改稿

技术;(5) 任意方向字符串检测技术.其特点是,能够全自动地完成字符分割,并获得满意的字符串提取效果和处理速度.

连通体标识是在分离字符与图形之前而采用的基本处理技术,其算法性能是关键.而且在实用过程中,交互增删也十分必要.我们采用了基于区域树的存储结构来方便、快速、准确地实现这种功能,关于(1)、(2)两项,详细内容可参见文献[8].

本文将主要介绍图文分割过滤器的尺寸阈值设计和任意方向字符串检测技术,重点阐述字符过滤的自动阈值技术以及基于HOUGH变换的字符串检测技术.

1　字符分割自动尺寸阈值

字符的重要特征在于相同尺寸单体的聚类效应,这种尺寸门阈的确定,国外类似软件采用了手动输入的方法.我们在提供手动输入功能的同时,还提供了一种自动阈值技术.利用该技术,可以在不设初始参数的条件下,全自动地完成字符的分离过程.

通常,线状图中字符的尺寸分布是规律的,因此,采用自动分割技术省去了试验选取门限参数的麻烦,智能性得以提高.手动门限的选取,通常是绝对值选取,这一绝对值往往适应性差,而自动门阈采用的是相对值,适应性良好.

该技术是通过统计字符尺寸特征来实现的.一个典型的机械零件图如图1所示.如不考虑颗粒噪声,连通体按区域面积大小出现的频率分布如图2所示(采用15点邻近平滑,图3与此相同),可以认为是一个正态分布.由于[9]

$$P\{|X-\mu|<3\sigma\}=0.9937,$$

其中X为某一大小连通体,μ为均值,σ为标准差,这表明,可以认为字符一定落在$(\mu-3\sigma,\mu+3\sigma)$范围内,因此,可以采用迭代方法计算自动阈值.即

(1) 先计算μ,σ;

(2) 去除落在$(\mu-3\sigma,\mu+3\sigma)$之外的连通体;

(3) 重新计算μ,σ;

(4) 重复步骤(2)、(3),直至没有新的连通体可以去除,

则最终的$(\mu-3\sigma,\mu+3\sigma)$即为自动筛选字符的阈值.

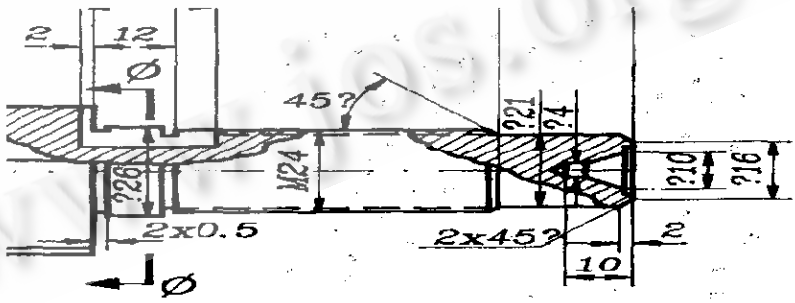

图1　典型机械工程光栅图

连通体的区域面积分布曲线的实际情况要复杂一些.经过实验观察,一种普遍的情况是:存在两个峰值,一个是点状噪声区,另一个是字符聚集区,如图3所示.两个区域是容易区别的.点噪声连通体区域面积较小,故位于坐标低端范围;字符聚集区的区域面积相对较高.如果单纯采用上述方法进行阈值估计,必然产生较大的偏差,使分割结果受到严重影响.为此,可以采用简单的固定阈值,首先过滤颗粒噪声,然后计算字符的阈值.由于事先有连通体检测作为基础,过滤该类噪声并不占用额外的计算.

本文考虑自动地同时统计噪声和字符阈值.设图3中两个区域均为正态分布,则可以采用下列步骤计算自动阈值:

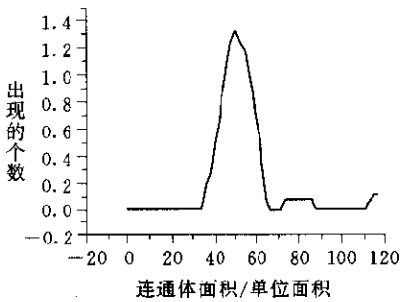

图2　连通体区域面积频率分布

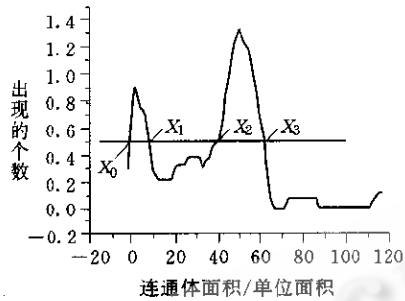

图3　存在颗粒噪声的连通体区域面积频率分布

(1) 计算连通体整体平均值 μ_{all};

(2) 以 $\mu_{all}/10$ 为步长由最大连通体向下搜索,确定 4 个分开的坐标值 X_0, X_1, X_2, X_3;

(3) 采用前文方法分别在(X_0, X_1)和(X_2, X_3)内估计阈值范围$(\mu_{noise}-3\sigma_{noise}, \mu_{noise}+3\sigma_{noise})$,$(\mu_{char}-3\sigma_{char}, \mu_{char}+3\sigma_{char})$.

2　任意方向、任意长度字符串检测

字符是以串的形式存在的.分析字符的字符串归属能够方便字符串整体的交互修改、增删;能够为字符识别提供方向信息、上下文信息.字符串成组通常采用的方法有两种,一种是相邻检测[3],另一种是利用 HOUGH 变换进行共线检测[2].相邻检测属于一种局部方法,HOUGH 变换是一种整体方法.本文采用的是后一种方法,主要涉及了 HOUGH 变换、基于字符串的 HOUGH 域更新等方法,能够满足任意方向、任意长度字符串的要求.其特点是,通过选取适当的 HOUGH 域分辨率,计算量较小,因此可以获得很快的检测速度,同时也符合人为审图习惯.在实现过程中,本文分析并解决了以下几个关键问题:

(1) HOUGH 变换的空间复杂度较大,在选择合适的分辨率的情况下,需要确定 HOUGH 空间的最大范围.本文提出了一个计算公式,合理解决了这一问题,使内存需求降低近 1 倍.

(2) 由于字符串并非严格的共线,因此,采用单纯的 HOUGH 变换无法获得正确的结果.本文提出一种称为松弛共线的策略,有效改善了字符的串归属.

(3) 由于字符串的多方向、非严格共线给 HOUGH 变换带来了问题,所以,HOUGH 空间必须反复刷新,而且更新过程中的次序也非常重要.为此,本文提出一种基于字符串的检测次序,使字符串的提取更趋精确,对于检测交叉字符串和处理中文字符十分有效.

2.1　HOUGH 变换

HOUGH 变换将笛卡尔域的直线变换成 HOUGH 域的点.笛卡尔域直线可以表示为

$$\rho = x\cos\theta + y\sin\theta, \tag{1}$$

则对应的 HOUGH 域的点为(ρ, θ),ρ, θ 分别表示极径和角度坐标.与此相类似,笛卡尔域的每个点(x, y)与 HOUGH 域的一条曲线相对应.图纸图像中某字符串中的字符通常位于一条直线上,对共线字符的中心分别作 HOUGH 变换,则 HOUGH 域相应的曲线必然交于一点(ρ, θ).在具体实施过程中,需将 HOUGH 域离散化,并将每一坐标点称为累加单元,在每个累加单元相交的曲线个数称为聚集数.

2.2　字符检测基本步骤

(1) 设置 HOUGH 域分辨率.在用于共线检测过程中,HOUGH 域的分辨率选取直接影响检测的结果.本文采用的分辨率为 1°.ρ 必须是字符尺寸的相对量,过大的极径将使不共线的部分产生混淆,过小的极径使同一字符串分成多组.极径最好是选为局部字符串尺寸的函数,但由于共线检测前该尺寸未知,因此,本文采用全局字符尺寸的综合平均值,尺寸选为字高 h(由于字符方向未知,字高为字符的纵向尺寸).本文的试验表明,极径选为$(0.2\sim0.4)*h$ 效果较好.

(2) 对给定区域进行 HOUGH 变换.根据人为字符书写的习惯,取角度变化范围为 0°～180°.

(3) 统计松弛因子,计算平均字符尺寸.统计的次序按照先大后小的原则,先取聚集数多的累加单元进行组合,主要是为了避免交叉共线字符的丢失[2].计算平均字符尺寸是为了进一步分割同一共线组上的字符串.

(4) 字符串成组并分离字符.该步骤只是简单地计算共线单元的字符串归属.

(5) 更新 HOUGH 域.将已经成组的字符串从 HOUGH 域中删除,避免下一循环重复统计.

2.3 HOUGH 空间确定

HOUGH 空间的存储要求由 ρ,θ 维数决定.θ 已选定为 180°,因此,需要适当选取 ρ 的范围,也就是获得 ρ 的最大值,从而请求需要的内存量.

式(1)为 HOUGH 变换的基本公式,其中(x,y)为某点的直角坐标.我们的问题可以表述为:已知式(1)中的 x,y 的取值范围和 θ 的极值,求 ρ 的最大值.

通常,因为 $\cos\theta\leqslant 1,\sin\theta\leqslant 1$,所以有

$$\rho_{max1}=(x_{max}+y_{max})/R, \tag{2}$$

这是一个比较粗略的估计.当图纸较大时,空间需求仍然很大.我们采用二元不等式求极值的方法,得到了一个较为精确的值,如式(3),

$$\left.\begin{aligned}\rho_{max2}&=(x_{max}\sqrt{1+\alpha^2})/R\\ \alpha&=y_{max}/x_{max}\end{aligned}\right\}, \tag{3}$$

其中 R 表示 ρ 的分辨率.

经过比较,式(3)的计算结果比式(2)的结果小近乎 1 倍.例如,当 $x_{max}=y_{max}$ 时,$\rho_{max2}/\rho_{max1}=1/\sqrt{2}\approx 0.707$.这对节约空间是很有效的.

2.4 松弛共线

由于字符串中单个尺寸和位置的复杂性,单纯统计单个累加单元的聚集数将会漏掉字符串中的某些字符.采用合适的分辨率 R,可以使共线具有一定的动态范围,但是对于类似上下标的处理是不够的.事实上,属于同一字符串的字符即使不是聚集在同一累加单元上,其所在单元一般在 θ 坐标上也是相邻的.为此,累加单元数的统计可以采用将相邻的累加单元合并的方法,以增加统计的松弛性.具体的调整方法是,找到一个松弛因子 δ,保持 ρ 不变,将 θ 的动态范围调整成$\{(1-\delta/2)*\theta,(1+\delta/2)*\theta\}$.很明显,分辨率越高,动态范围越小,因此 δ 应增大,以提高动态范围;如果字符平均高度较大,δ 也应相应增加.δ 可以描述成分辨率 R 和字符平均高度 h 的函数:

$$\delta=\lambda h/R, \tag{4}$$

其中 λ 是系数,$\lambda=1\sim 1.5$.由试验得到.

经过松弛共线处理,可以较好地处理具有上下标、特殊符号的字符串归并.

2.5 基于字符串的 HOUGH 域更新

文献[2]在字符串的提取过程中,采用了先长后短的原则,目的是防止误提取.在同一图纸中,字符串有长有短,处理的顺序非常重要.下面的例子表明了这一情况.

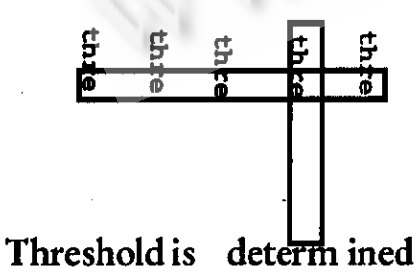

图 4　HOUGH 域检测顺序的影响

由图 4 可以看出,当不同方向的字符串出现交叉时,如果不考虑次序,则 thres+m 可能将被提取.而与之相交的横向字符串中的 m 也被取走,结果 determined 字将被识别成 deter 和 ined 两个字.虽然采用后处理可以解决一些问题,但如果采用先处理长字符串的方法,就可以避免这类问题.因此,在整个检测过程中,可以采用从大到小逐层筛选的策略.具体做法是:设定阈值,然后逐层减 1,先处理长串,后处理短串.

在实施这一算法的过程中,我们注意到一个新的问题,如图 4 所示,字串 thre 共有 4 个字符,而当一组字符形成文字段形式时,在段方向上字符累计数有时大于字符串本身的字符数.因此,依照先长后短

的原则,我们提取到的是一串字符 e.这表明,单纯采用先长后短的原则还不够,必须进行进一步的限定.

经过分析,我们认为以字符串大小作为限定是非常合适的.也就是说,在我们得到一个足够长的共线组后,还必须继续将其分解为字符串,只有被分解的字符串长度符合限定时,这个字符串才可以被提取.

具体实施过程是在长串分割成字的过程中,再次引进门阈 $R_{threshold}$,当最终字符串的长度大于 $R_{threshold}$ 时,进行最终提取.字符串的判断采用近邻检测方法,门阈依据局部字高判断.由于在一段字符中,行距通常大于字符间距,采用近邻检测将上面提到的一串字符 e 分割成只有一个字符 e 的串,因而先不被提取,保证了结果的正确性.

单纯采用先长后短的检测方法所存在的另一个问题是对中文字符的处理.中文字符有一个特点是,它常常由多个部分组成,当仅以连通体个数进行字符串统计时,有时一个中文字符的连通体个数就相当于一个英文字符串的连通体个数,因此必然引起误判.例如,对于上文的字符 e,若将其替换为中文字符"行",则与一个英文串"long"的连通体个数 4 相当,完全可能先被提取.本文的方法是:在统计字符串过程中引进连通体的字符属性归并,通过计算连通体的相对位置确定其字符归属,从而调整字符统计结果,形成正确的字符记数,进而使提取顺序正确.

3　性能评价

SegChar 能够有效地将工程图档中的文字与图形部分分离,将文字和图形分别进行存储.对于文字符号,可以提取任意大小、任意方向的字符,且不受盐粒噪声的干扰.能够提取单个字符并识别相应的字符串,保存字符的位置信息、方向信息、像素和尺寸信息以及字符串归属信息;识别结果存储在区域树结构中,可以方便地实现交互增删,满足用户对分割结果进行修整的要求.其中,字符串的归属准确程度高,对上下角标、符号,由多个部分组成的字符归属问题处理得当,能较好地处理典型的 i 问题以及中文中的点划等问题.

SegChar 采用 C 语言编程,在 UNIX 工作站上实现,整体性能较好,灵活性较强.下面仅提供自动阈值条件下的分离正确率计算速度指标.分离的正确率不包括与图形粘连字符的情况.关于字符粘连的处理,我们已经提供一定的自动分离策略,但对于大图而言,效率尚有待改进,因此,该工具中尚未包括.经过对多幅图像的识别测试,字符分割平均正确率可达 94%以上.图 5 表示了原型系统分割结果的一个场面,其中左上角倾斜的字符串 R2.5 也得以正确分割,而且得到了相应的方向参数.图中为了显示方便,仍按字符所占矩形框画出.

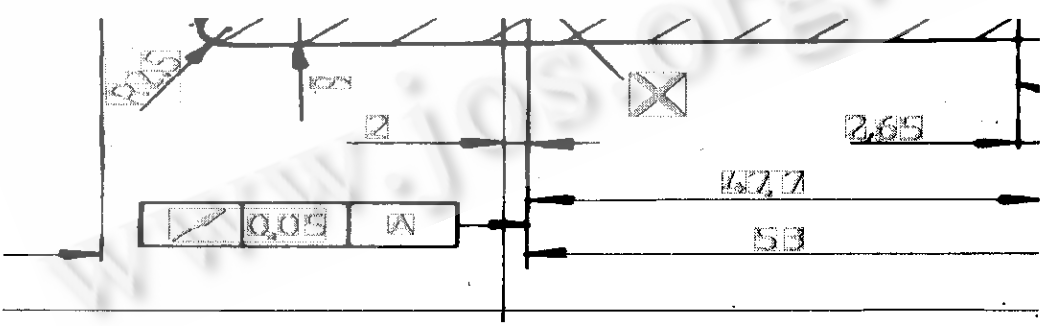

图 5　字符分割例子

SegChar 的计算速度在一台 SGI Indy 上进行测试.其主要指标为 CPU:MIPS R4600,132MHz;RAM:32MB.测试样图的局部如图 1 所示.整图大小为 4420×3836 pixles(300dpi,374.65×328.87mm^2);字符单体 156 个;字符串 55 个;计算时间为 12s.采用多幅样图的测试表明,字符多少并非影响速度的主要因素,而图的尺寸都起着决定作用.因此可以推算,处理一个 A0 尺寸的图大约需要 96s,实际上则往往低于此值.而处理一个选定为 1000×1000pixels 的窗口大小的图像仅需不足 1s.因此,SegChar 处理效率上是满足一般要求的.

4 结束语

图纸分层处理是一个基本步骤.典型处理就是字符与图形的分割.为了适应图纸中字符分割的特点,本文采用了自底向上的一整套处理方法,提出了有效的字符自动分割阈值技术.在采用HOUGH变换进行字符串检测的实践中,提出了精确HOUGH空间需求、松弛共线、基于字符串的HOUGH域更新等技术,使字符分割的处理速度显著提高,内存用量有效降低,处理结果明显改善,从而形成了实用工具SegChar.

参考文献

1 Wang P S P, Bunke H. Handbook on Optical Character Recognition and Document Image Analysis. Singapore: World Scientific Publishing Company, 1996

2 Fletcher L A, Kasturi R. A robust algorithm for text string separation from mixed text/graphics images. IEEE Transactions on PAMI, 1988,10(6):910～918

3 Oliver L. Automatic indexing of line drawings for content based information retrieval [Ph. D. Thesis]. ETH No. 11870, Dipl. Informatik-Ing. ETH. 1996

4 Filipski A J, Flandrena R. Automated conversion of engineering drawings to CAD form. Proceedings of IEEE, 1992,80(7):1195～1209

5 杜建强,陈月林,刘少媚等.工程图纸上的字符提取和识别系统.计算机工程,1995,21(1):62～65
(Du Jian-qiang, Chen Yue-lin, Liu Shao-mei *et al*. A system for character extraction of engineering drawings. Computer Engineering, 1995,21(1):62～65)

6 陈建国,罗伯朋,魏小鹏等.对扫描图像的一种新型图文分离方法.见:谭建荣主编.计算机工程图学的探索与实践.杭州:浙江大学出版社,1994. 340～343
(Chen Jian-guo, Luo Bo-peng, Wei Xiao-peng *et al*. A new method for text/graphics segmentation of scanning images. In, Tan Jiang-rong ed. Exploring and Practice of Computer Engineering Drawings. Hangzhou: Zhejiang University Press, 1994. 340～343)

7 Fan K, Lu J M, Wang L S *et al*. Extraction of characters from form documents by feature point clustering. Pattern Recognition Letters, 1995,16(9):963～970

8 江早,刘积仁,王冬等.一种可交互删改的二值图像快速连通标识方法研究与实现.东北大学学报(自然科学版),1998,19(3):251～254
(Jiang Zao, Liu Ji-ren, Wang Dong *et al*. Study and implementation of a fast interactive modifiable method for connected component labeling of binary image. Journal of Northeastern University (Natural Science edition), 1998,19(3):251～254)

9 王淑琴.概率论与数理统计.沈阳:东北工学院出版社,1988
(Wang Shu-qin. Theory of Probability and Statistics. Shenyang: Northeastern University Press, 1988)

SegChar——a Tool for Automatic Text/Graphics Segmentation of Engineering Drawing Images

JIANG Zao LIU Ji-ren LIU Jin-jun

(*Software Center Northeastern University Shenyang* 110006)

Abstract In this paper, the technical characteristics of text/graphics segmentation of engineering drawings, its critical steps and basic processing framework are analyzed. The SegChar, a practical tool for automatic text/graphics segmentation developed by the authors, is presented. The emphasis is put on two technical respects: (1) the automatic text size threshold method, which makes the processing procedure automatic and intelligent; (2) the probing strategy for extraction of text strings of arbitrary direction and length, which can extract Chinese/Western character strings integrally, and enhances its performance in processing speed, space complexity by introducing the concepts of accurate HOUGH space requirements, collinear relaxation and the string based HOUGH field refreshing method. The performance evaluation of SegChar is given finally.

Key words Drawing image processing, text/graphics segmentation, HOUGH transform, automatic threshold.