分割图形粘连字符的图元屏蔽技术研究*

江 早 刘积仁

(东北大学软件中心,沈阳 110006)

摘 要 在工程图纸计算机处理过程中,分割字符与图形是非常重要的步骤,但字符与图形粘连的问题很难处理。本文在分析了几种字符与图形分割技术的原理和实现方法的基础上,提出一种分割与图形粘连字符的方法——图元屏蔽技术。试验结果表明,这种方法对于分割与正交方向线段粘连的字符十分有效,特别适用于处理字符串中全部与图形粘连或多个粘连的现象。

关键词 工程图纸处理,图文分割,字符粘连

1 概 述

以图纸形式存放的工程技术档案,存在占用空间大、易损、难以长期保存、查询烦琐、不利于修改利用等致命缺点,直接影响到设计和生产的效率、质量以至于产品的市场竞争力。虽然 CAD 技术的迅速发展使工程设计在一定程度上实现了计算机化,但是,即使在美国这样的发达国家,仍有 55%的工程图以图纸档案形式存在[1],国内的图纸档案占有比例更是高达 90%[2],而且今后每年还以很高的比例不断增长。工程图纸计算机处理技术是近年来解决以上问题的一种有效途径。它采用光盘存储、图象处理等许多先进的计算机技术,能够实现图纸档案的计算机输入、存储归档、检索、修改设计、以至于向标准 CAD 文档格式的转化,具有广泛的应用价值和潜在市场;同时作为一种工程信息自动化输入的手段,是未来基于图纸草图设计的重要技术基础。

不同层次的图纸处理技术满足不同的应用需求,目前从处理形式上主要可以分为两方面:即光栅编辑技术和图纸 CAD 转化技术。光栅编辑技术指对图纸进行基于像素的编辑处理,而图纸 CAD 转化技术是指经过处理将光栅图象转化为 CAD 标准格式,这是目前图纸处理的中心内容。图纸 CAD 转化技术包括:图文分割、矢量化、字符识别、二维图纸理解、三维重建等技术。工程图纸主要由图形轮廓、各种标注符号、用于说明的标题栏等组成。字符与图形在含义、表示方法上完全不同,这种左异形成了图形元素间一个最主要的区别。从层次角度考虑,字符与图形应分放在两个不同的逻辑图层。这就是通常 CAD 格式转化进行的第一步处理,其目的在于,为下一步采用不同的处理方法提供原始素材,同时节省处理存储开销。

工程图纸中的字符有大小不一、方向不一、位置分散、部分与图形粘连、很多与图形相似几个特点。因此,字符提取和识别比处理普通纸介质文档的难度更大。其中,尤以粘连字符最难处理[1,3,4]。

造成图形字符粘连的原因主要是扫描质量差或者原图多次复制等。粘连可以分为字—字粘连和字—图粘连。字—字粘连问题一般放在 OCR 阶段解决,而且已有一些基本解决方法。本文主要研究字—图粘连的处理。在这类问题中,最常见的就是下划线类的粘连问题,如工程图纸中与尺寸线粘连的字符。考虑这一问题,Fletcher 和 Kasturi 曾经提出了基于字符上下文技术处理粘连字符的方法[3],由于该法基于至少由 3 个以上字符组成的字符串以及普通的连通检测技术,因此只适用于字符串中个别粘连字符的处理,对于分散性很强的小字符串或单个粘连

* 国家九五重点科技项目(攻关)计划 96-B08-3-3 资助
收稿日期:1997-01-21;收到修改稿日期:1997-03-26

字符以及全部有粘连现象的字符串，这种方法难以解决问题。朱林等提出一种基于尺寸线区域的字符搜索技术[5]，对于解决字符粘连有一定效果，但只限于在尺寸线周围局部区域搜索；Fan 为解决表格文档中字符分割问题，提出一种基于聚类分析的字符分割技术[6]，对于处理与表格粘连的字符也很有效果，但该法主要适用于中文，对于解决工程图纸中大量的数字与西文字符，由于类聚特征不显著，效果不佳。字符与图形粘连的形式多种多样，其粘连特征难以描述和提取，目前还没有功能较强的图形粘连字符分割方法，成为字符图形分割的主要障碍之一。本文提出一种分割与图形粘连字符的图元屏蔽技术，主要思想是对可能与字符粘连的图形部分首先进行标识，并在分割过程中进行屏蔽处理，结合范围、密度等门限技术，可以比较有效地处理工程图纸中字符与尺寸线及其他正交标志线的粘连问题。并具有如下几个显著特点：(1) 能够有效处理全部与线段粘连的字符串；(2) 能够处理少于 3 个字符的粘连小字符串；(3) 不受区域限制。因此具有一定的普遍意义。

2 字符分割技术

常用的字符分割方法有基于连通检测的分割方法；基于区域生长的分割方法；基于轮廓的分割方法等。所有这些分割技术，都是基于字符与图形的三个主要区别或假设，即字符与图形一般是分离的；单个字符比图形占有相对小的区域；字符一般成串组成。由于基于轮廓的方法不适合本文的屏敝技术，故不予叙述。

2.1 连通检测字符分割

所谓连通体，是指象素之间呈 4 或 8 邻接的一组象素集。基于连通检测的分割技术如图 1 所示。首先对要处理的区域进行连通体标识，区分不同的连通部分，再依据一组阈值过滤器，对连通体进行筛选，区分字符和图形。在图 1 中，不同的标识表示了不同的连通体，其中字符串“300”和“489”中字符和字符之间呈粘连状态，故只分别检测到一个连通体⑦、⑧。⑨表示了互相连通的图形部分。关于连通体检测，方法分为几种，具体可参见文献[3,8]。关于筛选的阈值，主要包括：

2.1.1 尺寸阈值包括长、宽及面积。

这是字符与图形的主要区别。这个阈值可以人为地输入，也可以采用自动阈值。手动阈值依赖于操作者的经验，但不影响分割速度。自动阈值的原理基于对全部生成的连通体尺寸的一种统计结果。自动阈值由于增加了部分计算量，速度比手动阈值稍慢。在给定阈值后，可以将大部分字符与图形分割开来，但是必然有部分非字符成分也被列入到字符集中来。如较小的图形部分、虚线、较小的断线、噪声点等。从阈值来看，只有上限还不够，需要补充相应下限，以限制噪声点。下限不能选得太大，否则由于字符尺寸的不均匀性，将遗漏某些字符。通常，经过尺寸筛选后，小图形和虚线等仍保留下来。

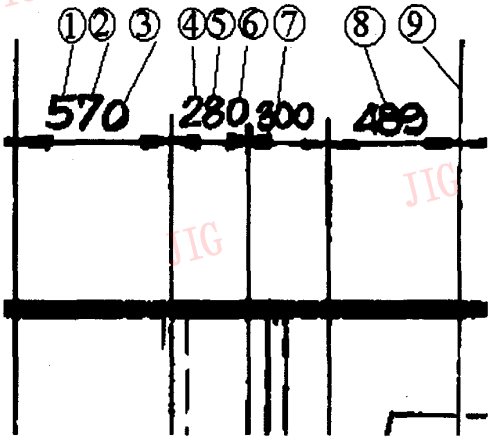

图 1 基于连通检测的字符图形分割

Fig. 1 Character segmentation based on connected components

2.1.2 特殊阈值

特殊阈值是为了区分字符和虚线一类特殊线型所选择的特殊阈值。如 Kasturi 等采用的竖直方向线宽域值淘汰虚线的方法[3]就是特殊阈值法。它首先采用了密度阈值筛选以同一字符为重复对象的字符串，然后判断字符串竖直方向的统计宽度，当这一宽度低于某一给定阈值时，则认为是虚线。

很明显，基于连通检测的方法，在字符与字符有粘连及字符与图形有粘连时，将遇到很大障碍。

2.2 区域生长字符分割

基于区域生长技术的字符图形分割技术[9]在本质上也是一种连通检测技术，但是表现形式有所不同，因此结果也不尽相同。他的基本原理是从一个最初包含前景象素的起始范围框开始，不断向与范围边缘象素邻接的周边拓延，在选定的尺寸阈值范围内，直至周边不再包含前景象素为止，确定此区域为

字符区域。图 2 中表示了字符“0”的搜索过程。

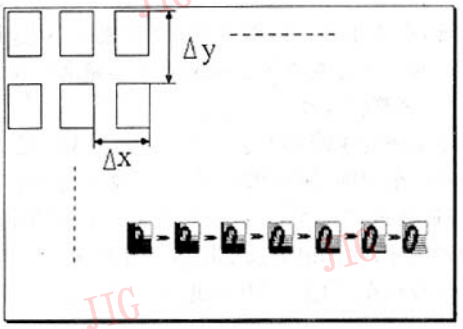

图 2　基于区域增长的字符图形分割

Fig. 2　Character segmentation based on area enlarging

具体实现这一技术,包括以下几个重要参数。

(1) 初始范围框,一般是一个矩形框。关于这一范围的选择,必须小于指定阈值。但也不能过小,否则扫描速度将大幅度下降。

(2) 扫描步长,指矩形框扫描图纸区域时在水平或垂直方向每一步移动的距离,如图 2 中 Δx, Δy。通常,扫描步长应小于初始范围的任一边长,以保证没有漏扫。

(3) 拓延步长,也称膨胀步长。指以初始范围框开始向周边每一步拓延的长度。这一长度通常是一个象素,否则将有误判和漏判。

(4) 字符阈值。这个值与连通检测技术采用的阈值是相同的。在扫描过程中,通常要先判阈值而后拓延,而不是在完全拓延以后再判阈值,目的在于减小计算量,加快计算速度。

3　粘连字符分割技术

由对基于区域增长字符分割方法的分析可知,区域生长与连通检测的重要区在于是连通体的生成方式,而区域生长的方式给分割图形粘连字符带来可能。如果在拓延的过程中可以识别与字符粘连的图形部分,处理粘连字符的难度就会大大减小,本文提出的图元屏蔽技术就是基于这种思想。

3.1　基本原理

图元屏蔽技术首先要对图形的有关部分加以标记,为拓延过程做准备。在进行区域增长的过程中,以标记过的象素为拓延结束信号;而在处理标记过的象素本身的过程中,又完全按普通象素对待。因此,对于拓延过程而言,标记部分相当于被屏蔽掉了,又因为屏蔽以图元为单位,所以称之为屏蔽技术。

屏蔽技术目前用于分割与水平或垂直的尺寸线或其他线段粘连的字符。如图 3 所示,首先对图象水平和竖直方向进行游长编码[4,10],当某一游长大于一定阈值后,认为它 是直线的一部分,加以标记,如图 3 中与字符粘连的尺寸线中最上一层象素(连续的前景象素,实箭头所指)。当拓延过程遇到该标记部分时,则停止拓延过程,可以使粘连的字符完整的分离开来。在具体的实现过程中,采用不断拓延边上屏蔽标记点个数的方法,并规定了点数的门限。图 3 空心箭头表示了拓延的方向。由于字符的搜索并没有象文[5]那样进行区域限制,因此,与尺寸线距离较远的字符仍可以得到有效分离。这种思想可以拓延到几种常见的几何图元的标记,如圆,圆弧等。

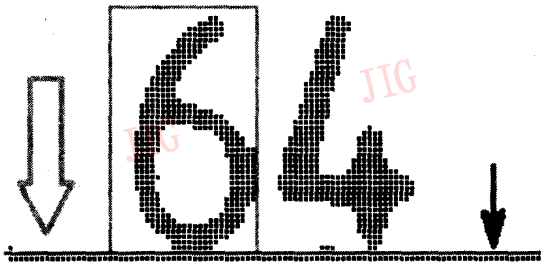

图 3　图元屏蔽方法

Fig. 3　Graphic element shielding method

屏蔽技术本质上是为粘连字符的分割提供了预备知识。因此,预备知识的完备性是分割效果好坏的前提。事实上,字符分割之前,图元的处理不可能完善,而对图元完全不处理,字符分割也受到很大限制,原则上,这是一种交叉精化的过程。理想的条件下,一个图纸合理的处理过程应该是各个部分处理过程的并行化过程,处理结果通过某种通讯机制互相传递,使处理过程在新的特征信息得到不断补充、中间结果不断精化中最终获得正确的处理结果。普通串行的处理方式受到很大限制,但对于不太复杂的图纸处理而言,必须分清主次,详略得当。本文采用了相对简单的检测技术,由于图形识别本身是很复杂的过程,而我们采取了先提取字符而不是先进行矢量化的过程,因此,此过程尽量与后续的图形处理分开。当然,根据需要的不同,也可采取较为复杂的标记预处理,如先进行 RG 矢量化[4],或者采用

3×3模板进行直线匹配等[10]。

另外，采用这种标记方法，而不是把它直接置成背景象素。原因在于：如果置成背景象素，由于不是完整的直线识别，因此，残余的未标识部分将部分加入字符集合。试验结果表明，这将使字符集合产生很大膨胀，给后续的识别过程带来很大困难。顺便指出，字符的分割与后续的识别是紧密相关的。这种关系体现在两个阶段结果的互相补充，也就是说，是一种逐步求精的过程。比如，字符分割中的大部分伪字符处理要依赖于 OCR 识别过程的拒识效应。

3.2 方法应用的策略

基于区域生长的字符分割技术，无法处理孤立性良好但在水平或垂直方向投影有交叉的一组字符的分割这一缺陷。尺寸阈值的选择必须允许狭长形区域的分割。这种情况属于字符与字符粘连的情况，主要依赖于后续识别处理的 OCR 技术中粘连字符分割技术[11]。如果是因为倾斜而无法分割，也可以采用连通检测方法进一步进行细分。本文采用上文提到的两种分割方法的有机结合，其中关键是方法运用的次序。

本文采用如下策略：首先对图元进行标记，判别字符粘连的可能性。而后基于区域生长方法进行字符分割，并形成字符串；最后对已分割字符进行基于连通检测的细分割。如果反之而行，则效率低许多。字符分割流程如图 4 所示。

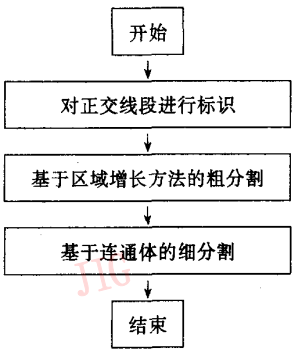

图 4 字符分割流程图

Fig. 4 Diagram of segmentation process

在处理过程中，还注意到以下几个问题：

3.1.1 效率问题

以往采用此方法的效率主要受到重复检测的影响，因此本文在检测到字符后，均进行了区域标识，以防重复检测，大大提高了检测速度。

3.1.2 特征问题

矩形框的位置与扫描步长有关，最终膨胀后的区域往往比所包容的字符大。为了便于后续特征抽取处理，本文采用了收缩处理，使区域最小化，其结果与连通区域检测的包容框结果大小完全一致。

3.1.3 字符串处理

本文研究了两种方法。方法之一是只考虑水平和垂直两个方向，原理相对简单。当检测到一个字符后，即搜索该字符的左右两侧，如有字符存在，则记录排列顺序，并标记为水平排列字符串。如果左右均没有字符存在，则上下搜索，此时，如果搜索到字符，则记录排列顺序，并记录该字符串为垂直排列字符串。方法之二的基本原理引自文献[3]，可称为共线法。采用 Hough 变换技术对连通项进行共线检测，从而使共线的字符聚类并形成字符串。采用该法可以检测任意方向的字符串，但是字符大小受到限制，由于共线检测的要求，字符串至少由 3 个字符组成。

3.3 实例

在 IBM RS6000 上采用 C 语言和 motif 程序对以上算法进行了试验。样图如图 5。由图可以看出，这是一份质量较差的机械图的一部分，其中的字符书写很不规则，字符与图形在多处粘连，因此更增加了字符分离的难度。这是未做屏蔽处理之前，采用基于区域增长技术进行字符分离的结果，图中显示了终止拓延的各个矩形框。结果表明，箭头所指处的粘连字符未得到分离。图 6 表示首先采用图元屏蔽技术对垂直方向的直线段加以标记，如图中空心箭头所示线段，分割算法因此对这部分进行屏蔽处理。结果表明，如此处理之后，原来未能分割的粘连字符均得到了很好的分割。图中的“K6”和“H9”这样的字符串均只由两个字符构成，如采用 kasturi 提出的方法显然不能分割[3]，因为他的方法必须基于 3 个以上字符构成的字符串。而本文提出的方法则没有该限制。另外，图中由汉字组成的字符串基本属于都粘连的情况，采用 kasturi 的方法也受到限制，因为它是基于上下文的，必须在大部分字符得以分离后，才能根据上下文关系分离个别粘连的字符，而本文提出的方法则不需要这种条件，即使字符全部粘连，也可以得到良好的分割。

图中某些个别未能包括的字符，主要是由于初始范围框尺寸选择的问题，如还可以选得再小一些，但得到的扫描效率也低一些。另外，还可以发现屏蔽算法的一些副效应。如图 6 中实体箭头所示，某中心线的一端被误判为字符。这种情况，可以交给 OCR

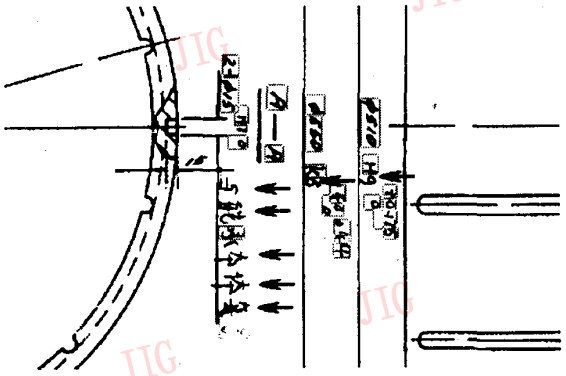

图 5　进行屏蔽之前的字符分割结果

Fig. 5　Segmentation result before shielding

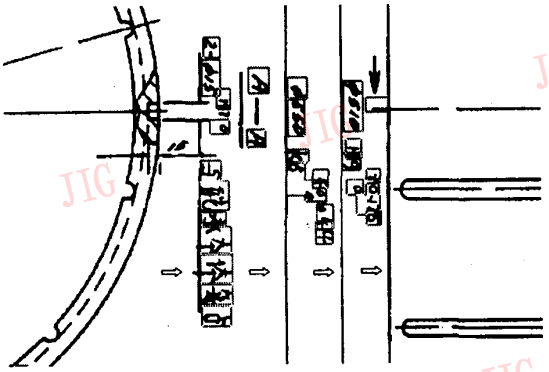

图 6　进行屏蔽之后的字符分割结果

Fig. 6　Segmentation result after shielding

部分处理,通过上下文或手动等方法排除在外。

4　结　论

本文提出了一种工程图纸处理中分割与图形粘连字符的图元屏蔽算法。能够有效解决与正交线段(包括尺寸线等)粘连字符的分割问题。特别适用于处理字符串中全部与图形粘连或多个粘连的现象,而且对非粘连字符处理没有任何影响,从而显著提高了自动分割字符图形的效果。该方法核心思想的延伸适用于与任何规则图元粘连的字符分割问题,具有普遍意义。

参 考 文 献

1　Filipski A J,Flandrena R. Automated Conversion of Engineering Drawings to CAD Form. Proceedings of IEEE, 1992, 80(7): 1195～1209.

2　东大阿尔派,SEAS 工程图纸自动处理设计与管理系统. 技术资料, 1994.

3　Fletcher L A, Kasturi R. A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1988, 10(6): 910 ～ 918.

4　Monagan G, Roosli M. Appropriate Base Representation Using a Run Graph. Proceedings of the Second International Conf. on Document Analysis and Recognition, Tsukuba, Japan, 1994,623 ～626.

5　朱林,常明,李小涛. 工程图形中的图素识别方法. 华中理工大学学报, 1995,23(2):120 ～ 124.

6　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.

7　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.

8　张树生. 一种基于线的标号传播二值图象连通体快速检测方法. 计算机研究与发展,1994,31(10):51～54.

9　陈建国,罗伯朋,魏小鹏等. 对扫描图象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. 浙江:浙江大学出版社,1994. 340 ～ 343.

10　陈廷标,夏良正. 数字图象处理. 北京:人民邮电出版社,1990.

11　Liang S, Shridhar M, Ahmad M. Segmentation of Touching Characters in Printed Document Recognition. Pattern Recognition, 1994,27(6):825～840.

江早,1989 年毕业于北京清华大学。1995 年在东北大学获博士学位。现为东北大学软件中心博士后研究人员,主要研究方向为图象处理及工程图纸的计算机处理技术。

刘积仁，东北大学教授，计算机应用专业博士生导师，东北大学副校长，东北大学软件中心主任。主要研究方向为分布式多媒体信息处理、数字图象处理、软件组件技术等。

On the shielding technique for segmentation of touched characters with graphics

Jiang Zao, Liu Jiren

(Software Center of Northeastern University, Shenyang 110006)

Abstract The segmentation of character and graphics is a fundamental step when processing engineering drawings using a computer. However, it is hard to deal when the character touches the graphics. This paper focus on this problem and gives a solving method. The method is called "the technique by shield". It shields lines in a graphics before segmentation. Some tests have shown good results of segmentation compared to which does not use this method in a graphics with a lot of touching points between the characters and graphics contours.

Keywords Engineering drawing processing, Segmentation of character and graphics, Touching characters

---

新书推介

《OpenGL 编程指南》

孙绍麟 费月娥 编译

全书共十二章，三个附录，16 开，300 页。全面、深入、详细地讨论了 OpenGL 编程中的实际问题，对 OpenGL 编程人员极具参考价值。

内容包括：绪论，绘几何对象，坐标变换，显示表，颜色，光照，混合、反走样和雾，绘象素、位图、字体和图象，纹理映射，帧缓存，鉴别器和 NURBS，选择和反馈，附录 1，2，3。定价：48 元

邮购：《中国图象图形学报》读者服务部（100088，北京海淀区花园路 6 号，
电话：62378784，联系人：李如珍）