字符粘连及字线相交的分割与识别方法*

邹荣金 蔡士杰 张福炎 苏 丰 陈冀兵

(南京大学计算机科学与技术系 南京 210093)

(南京大学计算机软件新技术国家重点实验室 南京 210093)

E-mail: zjzrj@public.zj.js.cn

摘要 描述了工程图纸矢量化中多向粘连字符及字线相交的分割算法与识别方法.提出不同情况下字串的定向计算方法,通过粘连字块的特征矢量计算和迭代计算实现字块的分割.运用波形投影方法解决了粘连字符及字线相交情况下的字间切割问题,使工程图多向字符识别精度显著提高,该算法对局部退化状态下的字符识别具有良好的抗干扰性.

关键词 计算机图形学,工程图纸识别,图像分割,识别.

中图法分类号 TP391

工程图纸的自动识别技术是目前学术界研究的一个热点.随着大幅面工程扫描仪的出现,研究领域和应用领域都对此十分重视,同时,多媒体技术的发展越来越要求计算机能够自动读入各种文字、图表、声音等信息,以减轻人们在计算机前的输入工作量.由于近年来计算机DPU和GUI的高速发展,文本OCR技术在走向应用和普及,而手写输入目前还未达到实用阶段,尤其是手写文稿的自动识别还有相当一段路要走,使用手写输入板对非计算机人员有一定帮助,但对一般会操作计算机的人员并不方便.工程图及图表的计算机自动识别、理解这方面的研究有很大潜在的应用价值,①其表现形式较为客观和准确,与计算机三维视觉相比难度要小;②以单色图像表示,数据量相对较小,分析识别的精度比较高;③直接面向产业界,对于工程设计、图档管理和GIS应用等领域有直接的经济价值.

目前,工程图自动输入处理技术的研究[1~4]主要是使用细化的算法.虽然细化在一定程度上可获取骨骼线,但会导致大量信息的丢失,尤其对于图纸上各种不同对象形态的整体理解较为困难.这些文献中对工程图字串的识别未作深入研究,文献[1]使用连通域进行了字串方向定位,这种使用连通域中心线定向的方法在字体形状变化较大时会产生较大的方向误差,而且不能处理和识别字串粘连.字线粘连及字线相交的情况.据此,我们作了多种试验来分割字的粘连情况,提出从连通群的特征矢量出发,逐次逼近解求字串的方向,并采用纵向投影技术实施字符的分割,对字线粘连、字线相交的分割算法作了详细描述.

1 工程图纸扫描图像中字符的分类与定向

字符的提取是以连通域为基础的,首先采用八连通算法将字块提取出来,依据连通域的大小进行分割和分类,一般先分为图形、字符和噪声3类.对于大字符及小图形分类上的交叉情况,其分类方法需要依据局部图像的矩的数学特性.

---

* 本文研究得到香港有利集团资金资助.作者邹荣金,1962年生,在职博士生,副教授,主要研究领域为计算机图形学,图像识别.蔡士杰,1945年生,教授,博士生导师,主要研究领域为计算机图形学,CAD技术.张福炎,1939年生,教授,博士生导师,主要研究领域为计算机多媒体技术,计算机图形学.苏丰,1974年生,硕士生,主要研究领域为工程图纸识别,CAD技术应用.陈冀兵,1974年生,硕士生,主要研究领域为工程图纸识别,CAD技术应用.

本文通讯联系人:邹荣金,镇江 212001,江苏省镇江市京河路13号301室

本文1997-10-14收到原稿,1998-03-19收到修改稿

1.1 相关大小字符、图形的数学特征与分类

对于一幅二值化图像$\{f(x,y);x,y=0,1,2,\ldots,N-1\}$,可定义一组相关数字特征量作为局部图像几何形状的判定因子与类别所属:

(1) 质心

$$(\bar{x},\bar{y})=(M_{10}/M_{00},M_{01}/M_{00}),\text{其中 } M_{pq}=\sum\sum f(x,y)x^p y^q. \tag{1}$$

(2) 中心矩

$$m_{pq}=\sum\sum f(x,y)(x-\bar{x})^p(y-\bar{y})^q,(p=0,1,2;q=0,1,2). \tag{2}$$

(3) 密度

$$D=m_{00}/2S(\text{其中 } S \text{ 为连通区域中的像素总点数}). \tag{3}$$

(4) 扁度

$$e=\frac{m_{20}+m_{02}+\sqrt{(m_{20}+m_{02})^2-4m_{20}m_{02}+4m_{11}^2}}{m_{20}+m_{02}-\sqrt{(m_{20}+m_{02})^2-4m_{20}m_{02}+4m_{11}^2}}. \tag{4}$$

这里,扁度e定义为连通区域的长短轴之比.各阶矩m_{pq}反映了字块的内部特征,不同形态的字符、图形其连通域的质心所处的相对位置$(\bar{x},\bar{y})$、密度D及扁度e不尽相同.一般情况下,在搜索连通域时将上述几项关键因子计算出来作为分类依据.例如,字符的质心通常位于连通域中心,图形则不然;而密度高的一般为字符连通区域,密度低的则为图形连通区域;字符的扁度通常在0.5~1.5之间,图形则相对变化较大.

1.2 字串的定向与方向纠正迭代算法

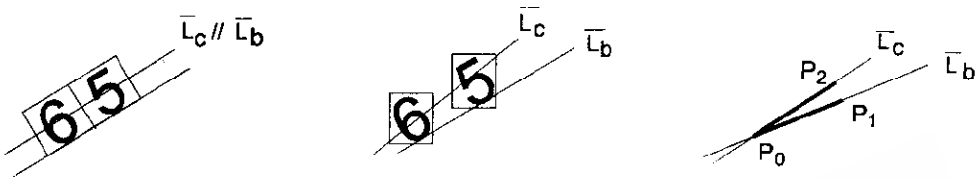
工程图中字符和数字的标识呈现多种不同的方向,要识别这些字符串,首先要从整个图像中将有意义的字符提取出来.工程图中分类出来的字符主要表现为两种状态:独立连通区和粘连.对于独立字符,它们具有各自独立的连通域,利用八连通递归算法可以取出每一个字符连通区的点阵,利用连通域的相邻关系,使其组成一个连通域链,链的每一个节点代表一个字符,并且每一个节点具有指向前一个字符节点的指针和指向后一个字符节点的指针,在确定前后相邻字符时需要考虑工程图中规定标识的可能存在位置.这样,在分析单个字符连通区特性时,就自然形成了串链.

工程图中的多向非粘连字串的识别,其定向问题是影响字串识别精度的关键因素.为此,首先建立字符串方向的严格定义:字符串的方向为工程图中字串书写的基线方向.这种基线的定义方式与AutoCAD中的DTEXT或TEXT命令绘图方式指定基线方向并沿基线方向书写是一致的,由于不同的字符其包围盒会不相同,特别是大小写字符.尽管如此,它们的基线却是一致的.遗憾的是,在图像中通过连通群分析一般只能获得连通字符图像块的连接关系,无法直接获取字符串的基线方向.由于字符集图像中各个字符的形状变化较大,使用不同的连通域参数(如轮廓、质心等)进行定向都会产生一定的定向误差.

本文首先是使用字符图像块的包围盒中心$\vec{C}((x_{\min}+x_{\max})/2,(y_{\min}+y_{\max})/2)$作为字符的定位点,利用首尾字符的中心点连线作为字符串的近似或初始方向.

对于长字串($n>10$),由于第一个字符的中心与最后1个字符的中心相距较远,利用其首尾字符的中心连线确定字符串的方向来取代字符的基线方向误差较小,设字符的中心点连线的方向为$\vec{L}_c$,字符串的基线方向为$\vec{L}_b$,则在这种情况下,$\vec{L}_b\cong\vec{L}_c$.对于中等长度字符串($3<n\leqslant 10$),采用最小二乘法确定字符串的方向一般具有较高的精度.

对于短字符串($n\leqslant 3$)使用上面两种方法都会产生较大的误差.如图1所示,如果已知每个字符的理想串包围盒,就很容易确定串的方向(如图1(a)所示).这时包围盒中心连线方向$\vec{L}_c$就是该串基线的方向$\vec{L}_b$,即$\vec{L}_c=\vec{L}_b$.而在实际计算中,一般只能获得串中每个字符的连通域表示及其链接关系,使用连通域包围盒的中心连线所确定的串方向与字的形态有很大关系.换言之,也就是用这种方法表示的方向只能确定一个串的初步方向$\vec{L}_c$,要获得字串基线的精确方向$\vec{L}_b$,需旋转$\vec{L}_c$,本文采用逐次迭代的直线移动方法快速确定字串的基线方向$\vec{L}_b$.设字串中各字连通域的包围盒为



(a) 理想串包围盒与基线方向$\vec{L}_b$　　(b) 包围盒中心连线方向$\vec{L}_c$　　(c) 方向误差$\vec{\varepsilon}=\vec{L}_c-\vec{L}_b$

图 1　短字串用首尾字包围盒中心连线表示串方向产生的方向误差

$$\vec{B}_k=(x_{\min},y_{\min},x_{\max},y_{\max})_k,\quad k=1,2,\ldots,m. \tag{5}$$

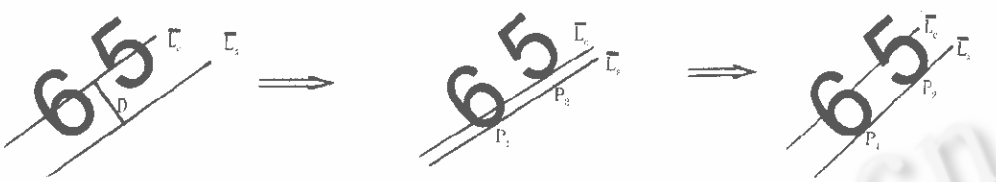
由此可以进一步计算出字块包围盒的高度和宽度：

$$h_k=(y_{\max}-y_{\min})_k,\quad w_k=(x_{\max}-x_{\min})_k. \tag{6}$$

取当前字串中所有字块的包围盒高度和宽度的最大值作为迭代初始直线$\vec{L}_s$到$\vec{L}_c$的距离

$$D=\max[(h_k,w_k),k=1,2,\ldots,m]. \tag{7}$$

如图 2 所示，首先依据上述方法得到初步方向$\vec{L}_s(\vec{L}_s=\vec{L}_c)$，将此直线向连通域靠近(每次推进的步距为 1～2 个像素)，分别获得第 1 次接触的点坐标 P_1,P_2，连接 P_1P_2 并使$\vec{L}_s=\overrightarrow{P_1P_2}$，继续上述向连通域靠近所求的新的接触点坐标 P_1,P_2，直到$\vec{L}_s$同时与两个连通域第 1 次相交或接触，这时取$\vec{L}_c=\overrightarrow{P_1P_2}$，即为基线方向，因此有$\vec{L}_c=\vec{L}_b$.



(a) $\vec{L}_s=\vec{L}_c$　　(b) $\vec{L}_s // \vec{L}_c$　　(c) $\vec{L}_s=\overrightarrow{P_1P_2}$

图 2　短字串的方向迭代逼近算法

2　完全粘连字串连通域的方向定位

字符粘连可分为两种情况：一是相邻字符间只有一处粘连；二是两相邻字符间有多处粘连. 首先从图像中用八连通算法将字符串提取出来，若粘连，则其宽度必明显大于其他独立字符，一个简单的处理方法是用一根直线与整个连通域求交，按照增量方法旋转直线，逐次逼近最大相交长度方向，即确定字符串的方向，然后进行垂直分割. 这种方法的缺点是定向精度不高，而且计算量较大.

2.1　粘连字符串的特征矢量

单个字符一般表现为纵向较长而横向较短的特点，而两个或两个以上粘连的字符则表现为横向较长而纵向较短的特点，因此只要找出连通域主轴方向，也就是连通域的特征矢量方向，则此方向可作为粘连字符串分割的参考方向.

设字符串图像为 $f(x,y)(0\leqslant x\leqslant m,0\leqslant y\leqslant n)$，则其协方差阵即二阶中心矩为

$$C=\begin{pmatrix}A & B\\ B & D\end{pmatrix} \tag{8}$$

其中

$$A=\frac{1}{M}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}(x-m_x)^2f(x,y),\qquad B=\frac{1}{M}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}(x-m_x)(y-m_y)f(x,y),$$

$$D=\frac{1}{M}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}(y-m_y)^2 f(x,y),\qquad M=\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}f(x,y),$$

$$m_x=\frac{1}{M}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}xf(x,y),\qquad m_y=\frac{1}{M}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}yf(x,y).$$

令 C 的较大特征值 λ_1 对应的归一化特征向量为 $(x_1,x_2)^T$,则

$$x_1=B\Big/\sqrt{B_2+(\lambda_1-A)^2},\qquad x_2=(\lambda_1-A)\Big/\sqrt{B_2+(\lambda_1-A)^2}. \tag{9}$$

这样便确定了主轴方向,因为参考字符主轴方向对应特征向量为 $(x_1,x_2)^T$,这样就可依据此方向作为字串定向的初始方向.对于部分或完全粘连字符串再进一步使用第 1.2 节描述的迭代算法,利用该特征矢量方向所确定的初始方向进行迭代计算,精确逼近整体字串的方向.

2.2 定向字符串的投影分割

设连通域图像函数为 $f(x,y)$,s 为投影方向,t 为与其垂直的方向,t 与 x 轴夹角为 θ,则 $f(x,y)$ 沿着 s 的投影定义为

$$p(t,\theta)=\int_{-\infty}^{+\infty}f(t\cos\theta-s\sin\theta,t\sin\theta+s\cos\theta)\mathrm{d}s. \tag{10}$$

当 θ 固定时,$p(t,\theta)$ 为 t 的函数,亦即一个一维波形.当 θ 从 0～2π 变化时,可得到 θ 在不同方向上 $f(x,y)$ 的投影.一般是将连通域图像按照特征矢量方向旋转到水平方向,这样,θ 方向的投影就变成 x 方向上的投影.在 x 轴上的投影定义为

$$p_x=p(t,0)=\int_{-\infty}^{+\infty}f(t,s)\mathrm{d}s=\int_{-\infty}^{+\infty}f(x,y)\mathrm{d}y. \tag{11}$$

粘连字符串水平方向的分割以投影波形为依据,实现时将倾斜方向的粘连字串旋转到水平方向,再沿垂直方向进行投影,将投影累计像素个数作为投影波形的幅度值,搜索波谷作为分割点(如图 3 所示).

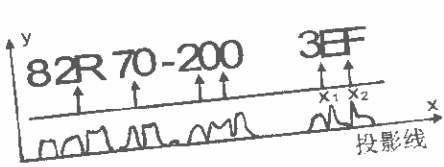


图 3 定向字符串的投影分割

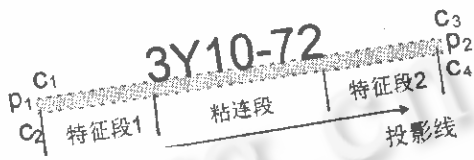


图 4 字线粘连与由特征线产生的切割线

3 字线粘连时字块图像切割方法

在工程图纸表示中,由于为表示字串与某线素的关系,往往把字串紧放(靠)在相关的线上方,这样,在工程图纸扫描时,由于圆形区域采样的原因,就出现了字串与底线相粘连的情况.而工程图中字串的识别的前提条件是字串及字符已构成独立的连通区域,因此,进行字线分离是字串准确识别的先决条件.

在以特征段为基础的直线矢量化方法中,一条直线上一般只有局部与字串粘连,因此,至少存在部分行段可用来构造、提取这条或几条特征段.如图 4 所示,利用特征段的定向延伸算法[5],可准确计算直线 $\overline{p_1p_2}$,利用特征行段分析获得的直线宽度,求得直线 $\overline{p_1p_2}$ 的包围矩形 $c_1c_2c_3c_4$,这样,位于直线 $\overline{p_1p_2}$ 两侧的切割线 $\overline{c_1c_4}$ 和 $\overline{c_2c_3}$ 即可求出.

$$\begin{cases}\vec{c}_1=\vec{p}_1+\vec{N}_{p_1p_2}\cdot width/2\\ \vec{c}_2=\vec{p}_1-\vec{N}_{p_1p_2}\cdot width/2\end{cases}\quad\text{及}\quad\begin{cases}\vec{c}_3=\vec{p}_2+\vec{N}_{p_1p_2}\cdot width/2\\ \vec{c}_4=\vec{p}_2-\vec{N}_{p_1p_2}\cdot width/2\end{cases}, \tag{12}$$

其中 $\vec{N}_{p_1p_2}$ 为垂直于直线 $\overline{p_1p_2}$ 的单位矢量($\overline{p_1p_2}\perp\vec{N}_{p_1p_2}$),$width$ 为线宽.

利用切割线 $\overline{c_1c_4}$ 和 $\overline{c_2c_3}$ 分别在直线 $\overline{p_1p_2}$ 两侧对直线边缘像素进行切割,采用 Bresenham 算法[6]将切割线 $\overline{c_1c_3}$ 上的像素改为背景色,即实现了字符与线的分离.

4　粘连字符的纵向切分算法

在工程图纸扫描图像中,由于字符串中的字符相距较近,某些字符在一起时会产生彼此粘连的现象,粘连的位置和程度与字符的形态有关,如图 5 所示.

图 5　字符粘连与投影特征

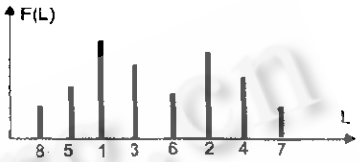


图 6　投影波形与波谷的逐级确定

造成字符粘连的原因来自 3 个方面:① 书写时或机绘时,字的间距太小;② 图纸的比例太小,也就是字太小;③ 扫描采样分辨率不够.这 3 个方面的因素在具体应用时都是无法避免的,前面两种因素是由于客观现存图纸及人们的绘图方式造成的,后一种因素虽然通过适当提高分辨率可获得缓解,但会造成数据量的急剧增加,使处理的时间增加很多,这也是不实际的,一般 A0 幅面的工程图纸扫描分辨率在 300dpi～500dpi 之间.因此,字符粘连在扫描图像中的出现不可避免.

在字符串中,一般是串中部分字符粘连,对短字串也可能会出现完全粘连的情况,对于前者仍可用上述一般方法进行字串定向,而对于后者则需要使用连通区特征向量分析和迭代逼近的方法确定字串方向.

在字串方向确定的基础上,如何实现粘连字符的分割是另一项复杂而又困难的工作.考虑到在一般情况下,字与字之间一般只有个别处粘连,因此在垂直于字串的投影方向上具有较少的厚度,使用第 2.2 节中提到的投影的理论与方法,建立投影图中的波形,把投影的波谷找出来,则波谷的位置就是切割线的位置.需要指出的是,由于字形的不同,可能呈现多个波谷,所以应当采用逐次细分的方法,逐级搜索波谷.如图 6 所示,其中 L 为投影位置,$F(L)$为投影像素累计值,序号 1～8 表示了逐次细分的位置编号.

在逐级搜索波谷位置时,应当考虑两个谷点(i,j)之间可能存在的距离 D_{ij},即

$$D_{\min}<D_{ij}<D_{\max}. \tag{13}$$

其中 $D_{\min}$为最小字距,$D_{\max}$为最大字距.当 $D_{ij}<D_{\min}$,则该谷点不成立;反之,若 $D_{ij}>D_{\max}$,则存在两个字符,即有一个分割谷点存在.此外,对于"1","l","i"这样的窄竖形字,其分割间距较小,其特征是在相应谷点的这一边有一个很大的谷峰存在,且满足

$$F(L_1)>F(L_2),\quad \text{当 } d_1<d_2 \text{ 时}(d_i \text{ 为某字符宽度}). \tag{14}$$

因此,在判定两个谷点之间是否还存在一次分割的谷点时,需要判别中间谷点的一侧或两侧是否有强谷峰存在,若有两个谷峰连续在一起,则判定的间距阈值还要进一步减少,因此需要使用自适应的方法来调整阈值.

5　部分字线相交时的分离与识别算法

当字串中的个别字符与外部直线相交时,往往很难分割和识别出被相交的字符,当连续 n 个字符被直线穿过时,有可能使本来完整的字串被分离为两个孤立的子串.

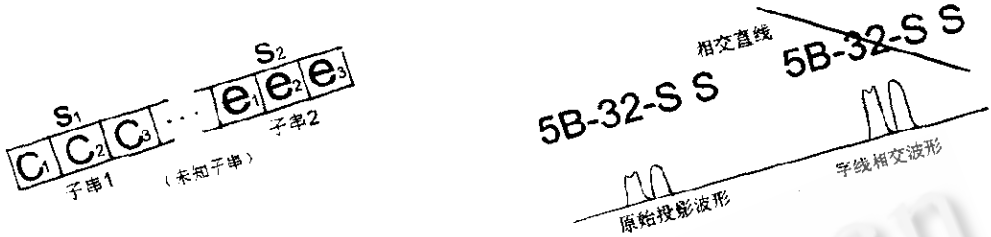
设存在两个相邻的字串为 s_1 和 s_2(如图 7(a)所示),$s_1=\{c_1,c_2,c_3\}$,$s_2=\{e_1,e_2\}$,那么这两个字串(或多个字串)是否属于同一个字串,需用以下约束规则进行判断:

(1) s_1,s_2 应该具有相同的字串方向.它们的中心延长线应当重合.

(2) s_1,s_2 两个字串的首尾最近距离应小于两个字符宽度(假定两个字串中间最多有两个字符被直线穿过).

(3) s_1,s_2 字串的平均字符大小(高、宽)基本相等.

满足上述条件,则表明 s_1,s_2,可能属于同一个字串,并且在其间可能存在未识别出的字符.取 s_1,s_2 两个字串的平均方向作为整个串的方向,依据两个串字符的平均高度和 s_1 到 s_2 的长度建立完整的矩形包围盒,使用投影的方法寻找字符分割的谷点,将字符逐个分割开来.



(a) 字串的相关性检测 (b) 字线相交与非相交的投影波形的比较

图 7 字线相交时的切割算法

由于直线宽度的一致性,在投影时,通常使投影累加值同时增长(如图 7(b)所示),只影响投影波形的幅度值而不会显著影响波形谷点的位置,因此仍然通过搜索波谷可较好地将字符分割开来.这样,使用标准模板进行中心位置的匹配识别将不会受到相交直线的影响.有了字串的方向及字符的大小,使用窗口探测技术,求出字串中间及两侧方向上单个字符的矩形窗口位置,同样采用标准模板进行匹配、识别,直到可能存在的字符检测完成为止.

6 实验识别与结论

在获取每个字串的方向及单个字符的横向分割位置之后,就可以将每个字符切割出来,由于工程图上的标注字符较为工整,只要字符切割位置准确,使用标准模板匹配的方法[4],并利用最大相似度准则可取得较高的识别精度.这种分割和识别的方法对字线粘连、字线相交及毛刺等局部退化现象均具有较好的抗干扰性.本文针对建筑结构图自动识别作了多种试验性研究,目的在于实现钢筋图纸的自动量数计算(香港有利集团项目).图 8 是 A0 图纸上局部一根梁的扫描图像,图 9 是相应的识别结果.

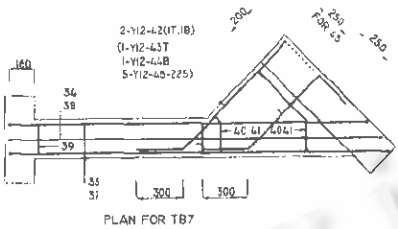


图 8 工程(结构)图扫描图像

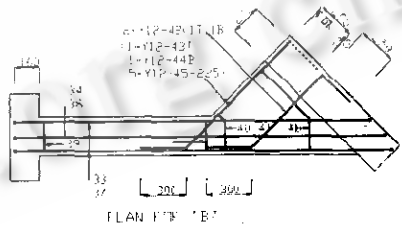


图 9 工程图识别结果

参考文献

1 Chan Pyng Lai *et al*. Recognition of dimension sets in engineering drawings. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994,16(8):848~855

2 Jairo Rocha *et al*. Character recognition without segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995,17(9):903~909

3 Nagasamy Vijay. Engineering drawing processing and vectorization system. Computer Vision, Graphics & Image Processing, 1990,49(4):379~397

4 Wei Dong-min *et al*. Using WEB grammar to recognize dimension in engineering drawings. Pattern Recognition, 1993,26(9):1407~1416

5 邹荣金,蔡士杰,张福炎.基于行程编码的直线拟合方法及其误差估计.软件学报,1997,8(增刊):404~410
(Zou Rong-jin, Cai Shi-jie, Zhang Fu-yan. Line interpolation method and error estimation based on run length coding.

Journal of Software, 1997, 8(supplement): 404～410)

6 唐荣锡，汪嘉业，彭群生. 计算机图形学教程. 北京：科学出版社，1994

(Tang Rong-xi, Wang Jia-ye, Peng Qun-sheng. Computer Graphics. Beijing: Science Press, 1994)

Segmentation and Recognition Methods of Adhesion and Intersection Character String with the Line

ZOU Rong-jin　CAI Shi-jie　ZHANG Fu-yan　SU Feng　CHEN Ji-bing

(*Department of Computer Science and Technology　Nanjing University　Nanjing*　210093)

(*State Key Laboratory for Novel Software Technology　Nanjing University　Nanjing*　210093)

Abstract　The segmentation and recognition methods of the omnirange character string which is adhered or intersected with the lines in the engineering drawing vectorization are described in this paper. The directional calculation methods of the string are also supplied on different drawing distribution cases, the segmentation of the adhered string between characters is also realized by calculating the feature vectors of the adhered local string block image and using the iteration approachment algorithm. The authors use the projection method to solve the segmenting problem of the characters in the case of character adhered or intersected with the lines, which made the recognition accuracy of the omnirange string increased. The algorithm has the antijamming property for the local degradation character recognition in the engineering drawings.

Key words　Computer graphics, engineering drawings vectorization, image