

一种字符图像线段提取及细化算法

刘峡壁 贾云得

(北京理工大学计算机科学与工程系, 北京 100081)

摘要 为了对字符图像进行稳定有效的细化和线段特征提取, 提出了一种同时完成字符图像线段提取和细化的算法。在定义字符图像的线段区域、交点区域和子段区域, 同时指出待细化字符图像是线段区域的集合的基础上, 该算法首先基于图像行(列)连通像素段来提取字符图像中所有的交点区域和子段区域, 然后对属于同一线段区域的所有子段区域和交点区域进行合并, 最后用 Bernstein-Bezier 多项式方法拟合每个线段区域的骨架。该算法的鲁棒性分析及实验结果表明, 该算法可取得稳定有效的细化及线段提取效果。

关键词 图像细化 线段提取 笔划提取 字符识别 区域分割

中图法分类号: TP391.41 文献标识码: A 文章编号: 1006-8961(2005)01-0048-06

An Algorithm of Line-segment Extraction and Thinning for Character Images

LIU Xia-bi, JIA Yun-de

(Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081)

Abstract This paper presents an algorithm for extracting line-segments from character images and thinning character images simultaneously. At first, the line-segment region, the sub-segment region and the junction region of an image are defined, and an image to be thinned is considered as a set of line-segment regions. Secondly, the method of extracting all junction regions and sub-segment regions is described based on connected pixel segments in rows and columns of an image, and the method of merging all sub-segment regions and junction regions that belong to the same line-segment region is presented. The Bernstein-Bezier polynomial method to fit every line-segment region's skeleton is also discussed. Finally stability analysis and experimental results of the proposed algorithm are reported.

Keywords character image thinning, line-segment extraction, stroke extraction, character recognition, region segmentation

1 引言

由于笔划是字符的重要结构特征, 因此笔划提取对于字符识别结构方法有关键性的影响, 脱机字符识别相对于联机字符识别, 其面临的主要困难之一就是笔划提取。笔划提取方法分为基于细化图像的方法^[1-3]和基于原始图像的方法^[4-7]两大类。由于现有的细化算法会在细化结果中引入畸变, 因此基于细化图像的笔划提取结果不够理想, 虽然人们试图对畸变进行修正^[1-3,8], 但作用也是有限的, 而更为可行的

方法还是基于原始图像的方法, 如行程长度编码法^[4,5], 轮廓信息法^[6,7]等, 这些方法虽都取得一定效果, 但目前此类方法仅局限于一种字符(比如只能用于汉字或只能用于字母), 仍缺乏适合于各种字符的通用方法。由于字符笔划特征是一种线段(直线段和曲线段)特征, 因此字符笔划提取问题可以上升为更一般的图像线段提取问题。同时, 由于图像骨架是单像素宽度的线段的集合, 且骨架中每个线段又是原始图像中某一区域的中轴, 所以待细化图像就是线段区域(简称线段)的集合。这样, 如果解决了一般图像线段提取问题, 就直接解决了像字符笔划提取这样的

收稿日期: 2003-05-23; 改回日期: 2004-08-20

第一作者简介: 刘峡壁(1972 ~), 男, 1998 年获华中理工大学计算机应用专业学士学位。现为北京理工大学计算机科学与工程系博士研究生。主要研究方向为计算机视觉、模式识别和人工智能。E-mail: yjia@bit.edu.cn; liuxiabi@bit.edu.cn

问题,同时也解决了字符图像细化问题。并且,如果首先提取字符图像线段,然后分别细化每个线段,最后组合线段骨架,那么由于在这一过程中避开了线段的相交,因而也就避免了结果的畸变。He 和 Yan 提出了基于上述思路的细化算法^[9],但由于该算法仅局限于汉字,因此针对性很强。

本文基于图像行(列)连通像素段,提出了一种新的字符图像线段提取及细化算法,该算法的特点是线段提取正确率高,细化结果无畸变,并具有良好的通用性和鲁棒性。该算法按执行先后顺序分为分割、合并和拟合 3 个步骤(如图 1 所示)。其中分割过程就是获取图像中所有的交点和子段,字符图像中线段相交形成的公共区域为交点区域,简称交点(图 1(a)),一个线段包含的全部交点将该线段分割为若干子段区域,简称子段(图 1(b)),没有交点的线段本身就是一个子段;合并过程是,首先单独细化每个子段(图 1(c)),并用多直线段拟合其骨架(图 1(d)),再利用多直线段对每个交点周围直接与之连通的子段进行配对,然后根据配对结果合并属于同一线段的所有子段和交点(图 1(e));拟合过程是用 Bernstein-Bezier 多项式描述合并后形成的每个线段(图 1(f))。

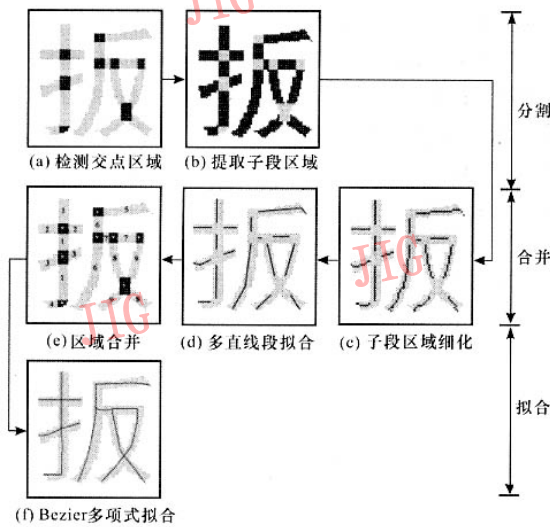


图 1 算法流程图及示例

Fig. 1 The flow chart and illustration of the algorithm

2 分割

图像行(列)上的连通像素段是本算法实现分割-合并的基础。设 $p(x,y)$ 表示图像中的第 x 行第

y 列像素。

定义 1 行段 $r(x_1, x_2, y)$ 是图像中某一行上的连通前景像素的集合,其满足 $x_1 \leq x_2$, 且 $p(x_1 - 1, y)$ 和 $p(x_2 + 1, y)$ 为背景像素值。

定义 2 行段中连通前景像素的个数称为行段的长度,简称段长。

定义 3 如果两个行段 $r(x_1, x_2, y)$ 和 $\hat{r}(\hat{x}_1, \hat{x}_2, \hat{y})$ 满足 $\hat{x}_1 - \hat{x}_2 \leq 1, \hat{x}_2 - x_1 \geq -1, \hat{y} - y = 1$, 则它们互为邻居,并且 \hat{r} 为 r 的下邻居, r 为 \hat{r} 的上邻居。

同样,可以定义列段、列段段长和列段之间的邻居、左邻居与右邻居。

图像交点区域中必然包含一些行(列)段,如果它们的某一侧邻居中包含不同线段的公共像素,而另一侧邻居中的全部像素仅仅属于某一个线段,则这样的行(列)段称为交点的边界行(列)段,简称交点边界。由于交点边界平行于坐标轴,因此每一个交点边界是在水平或垂直方向上施加于对应交点的一种限制。如果交点边界将对应的交点限制在它的下(右)部,则称为交点上(左)边界,否则称为交点下(右)边界。每个交点就是被它的全部边界围成的区域。

两个线段之间的相交关系划分为如图 2 所示的 3 种,但两个线段相交,其对应的交点边界只能是位于轮廓交点处的行(列)段。

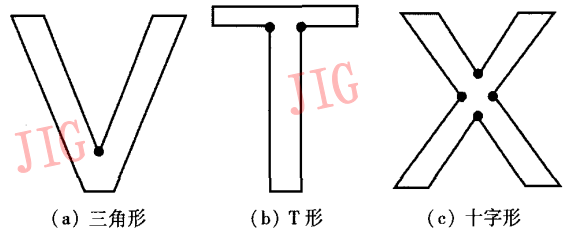


图 2 3 种线段相交关系示意图

(图中粗黑点表示线段轮廓交点)

Fig. 2 The illustration of three types of relationship between two line-segments which join or intersect with each other

图 3 是轮廓交点模型,其中向量表示轮廓,虚线表示行(列)段。在轮廓和轮廓交点处的行(列)段之间存在以下 3 种位置关系:

(1) 同侧关系 如图 3(a)所示,向量 a 与 b 位于行(列)段 c 的同侧,且 a 与 b 都不重合于 c 。由于 a 与 b 除了交点 o 以外,处处不交,因此 o 在 a 上的最近点 d_1 与 o 在 b 上的最近点 d_2 不连通,也就是,在与 a 对应的线段中存在一个以 d_1 为终点,且平行于 c 的行(列)段;在与 b 对应的线段中存在一

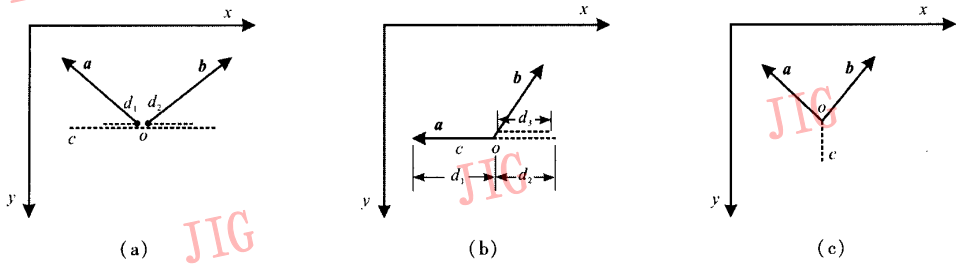


图 3 轮廓交点模型图

Fig.3 The modeling of junction points of line-segments' contours

个以 d_2 为起点,且平行于 c 的行(列)段。这说明,交点边界 c 必然在某一侧有两个邻居。

(2) 重合关系 如图 3(b)所示,向量 a 或 b 重合于行(列)段 c 。设 c 在交点 o 左边部分的长度为 d_1 ,在交点 o 右边部分的长度为 d_2 。在与 b 对应的线段中,存在一个距离行(列)段 c 最近,且平行于 c 的行(列)段,设它的长度为 d_3 。由于线段轮廓近似光滑,因此 $d_2 \approx d_3$ 。同时,若设 $d_1 = \lambda d_3$, λ 为缩放系数,则有 $\lambda \gg 0$, 否则难以感知与 a 对应的线段,于是 $d_1 + d_2 = (1 + \lambda) d_3$, $\lambda \gg 0$ 。这说明,交点边界 c 必然在某一侧有一个显著短于它的邻居;

(3) 异侧关系 如图 3(c)所示,向量 a 与 b 位于行(列)段 c 的异侧,且 a 与 b 都不重合于 c ,此时, c 没有明显的特征。如果忽视这种位置关系,那么有可能遗漏呈三角形相交关系的对应交点,但不会遗漏呈 T 形或十字形相交关系的对应交点,若将两个具有三角形相交关系的线段作为一个线段处理,则不会对细化结果造成影响,如果需要,还可以在细化完成以后,根据骨架的曲率变化实施分裂。

以上分析虽然仅针对两个线段的相交,但由于多个线段的相交实质上是两两相交的组合,因此该

分析也适用于多个线段的相交,只不过由于在图 3(a)所示的第 1 种位置关系下,交点边界在某一侧有两个以上的邻居,因此,结合以上分析结果,交点边界需要满足如下规则:

规则 1(交点边界规则) 如果一个行(列)段有两个或两个以上的上(左)邻居,或者长度显著长于它的一个上(左)邻居,则该行(列)段就是可能的交点上(左)边界;如果一个行(列)段有两个或两个以上的下(右)邻居,或者长度显著长于它的一个下(右)邻居,则该行(列)段就是可能的交点下(右)边界。

2.1 寻找可能的交点边界

规则 1 中有一个模糊条件,即一个行(列)段“显著长于”另一个行(列)段。该条件反映了图 3(b)所示情况,其可以被表示为:设 l_A, l_B 分别为行(列)段 A 与 B 的段长, α 为最小段长限制, β 为放大系数。如果 $l_A > \alpha \wedge l_A > \beta l_B$, $\beta > 1$, 则线段 A “显著长于” B。本文提出并采用段长直方图方法来计算图像行(列)段的 α 和 β 。段长直方图类似于灰度直方图,但横轴表示图像中不同的段长,纵轴表示图像中具有某种段长的行(列)段的个数(见图 4)。

图 4 中,不同线段包含的行(列)段可能具有不

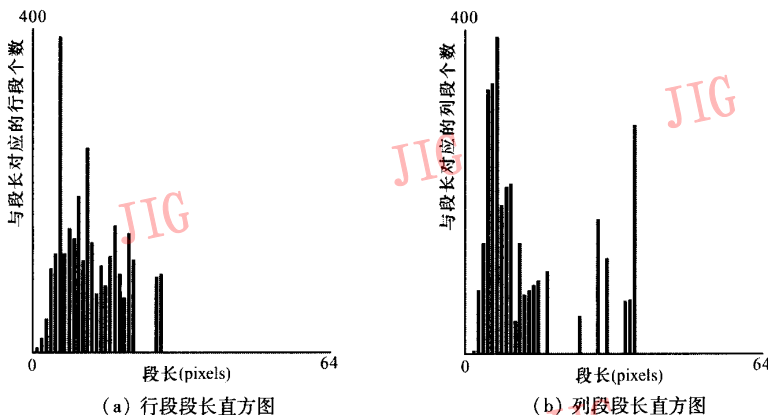


图 4 隶书“锈”字的行段段长直方图和列段段长直方图(图像尺寸为 64×64)

Fig.4 The HV-Run lenth-level histogram of Chinese character 'Xiu' with Li font(the image size is 64×64)

同的段长,但同一线段中包含的所有行段(列段)的段长是基本一致的,并且一般而言,线段中段长近似于其长度的行(列)段的个数远远小于段长近似于其宽度的行(列)段的个数。在图3(b)所示情况中,由于 $d_1 + d_2$ 近似于对应线段的长度, d_3 近似于对应线段的宽度,因此可以用段长直方图中的波峰与波谷来计算 α 和 β 。

在 α 和 β 确定以后,再检查图像中每一个行(列)段,如果它满足规则1中的条件,则记录该行(列)段是可能的交点边界,并以其对应交点边界的类型作为它的类型。

2.2 检测交点

设全部可能的交点边界构成的集合为 E 。检测交点的过程分为以下3步:第1步在 E 中找出所有的子集 e_i , e_i 是第 i 个交点全部对应边界的集合,它满足:① e_i 中至少有1个行段和1个列段,且二者有公共点;② e_i 中不包含两个同类型的元素;③ $e_i \cap e_j, j \neq i$;第2步根据每个 e_i 中包含的所有元素确定与每个交点对应的全部前景像素;第3步合并连通的交点。

在检测出全部交点之后,设 x 为横坐标, y 为纵坐标, $P(x,y)$ 为 (x,y) 点处的像素值,则前景像素 $P(x,y)=1$,背景像素 $P(x,y)=0$,可按下式计算各个交点的质心 (\hat{x},\hat{y}) :

$$\hat{x} = \frac{\sum \sum xP(x,y)}{\sum \sum P(x,y)}, \hat{y} = \frac{\sum \sum yP(x,y)}{\sum \sum P(x,y)} \quad (1)$$

2.3 提取子段

设 H 表示在某一侧没有邻居的行(列)段, \hat{H} 表示与交点边界互为邻居的行(列)段,则从交点的角度出发,图像中的子段可分为以下两种:一种是不与任何交点连通的子段,该子段中存在 H ,但不存在 \hat{H} ;另一种是与至少一个交点连通的子段,该子段中存在 \hat{H} ,也可能存在 H ,于是称子段中的 H 和 \hat{H} 为子段边界,并可得到子段边界需要满足的规则如下:

规则2(子段边界规则) 如果一个行(列)段没有上(左)邻居或者与一个交点的下(右)边界互为邻居,则该行(列)段是一个子段的上(左)边界;如果一个行(列)段没有下(右)邻居或者与一个交点上(左)边界互为邻居,则该行(列)段是一个子段的下(右)边界。

根据规则2找出图像中所有的子段边界之后,再从每一个子段上边界和每一个子段左边界出发,沿邻居作垂直或水平扫描,直到遇见一个子段下边界或子段右边界为止,扫描时,经过的所有行(列)段就构成

一个子段。最后,合并重复的子段,即完成子段提取。

3 合并

线段虽可被其中的交点分割为若干子段,但由于非封闭线段在同一交点处至多有两个子段,而封闭线段在同一交点处虽然可能有两个以上的子段,但与每一个子段直接连接的子段数至多为1,因此可以先以每个交点为中心,对直接与之连通的子段进行配对,再根据配对结果收集属于同一线段的所有子段和交点。例如,如子段 g_1 与子段 g_2 配对, g_2 与子段 g_3 配对,则 g_1, g_2, g_3 以及与它们连通的所有交点就共同构成一个线段,依此类推。配对的依据是子段之间的连续性。单独细化每个子段,并用多直线段拟合其骨架之后,即可用对应直线段之间的夹角来度量子段之间的连续性。本文采用的多直线段拟合方法为自顶向下的分裂方法。

3.1 子段细化

如果子段中存在行(列)段 w ,且 w 在某一侧有两个邻居,则可采用传统的迭代方法进行细化;否则需以行(列)段的端点轨迹为基础进行细化,步骤如下:

① 行段或列段的选取

如果子段中只有行段是交点边界的邻居,则选择行段进行细化;如果子段中只有列段是交点边界的邻居,则选择列段进行细化;否则需计算子段中所有行段的段长平均值 l_h (下角 h 代表horizontal)和所有列段的段长平均值 l_v (下角 v 代表vertical),如果 $l_h < l_v$,则选择行段进行细化,否则选择列段进行细化。

② 起点或终点的选取

设子段中所有行段(列段)构成的集合为 $\{k_i\}$, $i=1,2,\dots,n$, k_j 与 k_{j+1} 互为邻居, $1 \leq j < n$, $(x_{k_i}^{\text{start}}, y_{k_i}^{\text{start}}), (x_{k_i}^{\text{end}}, y_{k_i}^{\text{end}})$ 为 k_i 的起点和终点坐标(对于行段为横坐标 x ,对于列段为纵坐标 y),现以行段为例,其起点坐标的变化程度为

$$C_{\text{start}} = \sum_{i=1}^{n-1} |x_{k_{i+1}}^{\text{start}} - x_{k_i}^{\text{start}}| \quad (2)$$

终点坐标的变化程度为

$$C_{\text{end}} = \sum_{i=1}^{n-1} |x_{k_{i+1}}^{\text{end}} - x_{k_i}^{\text{end}}| \quad (3)$$

如果 $C_{\text{start}} < C_{\text{end}}$,则选择起点进行细化,否则选择终点进行细化。

③ 细化

$\{k_i\}, i=1,2,\dots,n, (x_{k_i}^{\text{start}}, y_{k_i}^{\text{start}}), (x_{k_i}^{\text{end}}, y_{k_i}^{\text{end}})$ 意义

同上。如选择起点进行细化,则细化过程为:设初始点 $o_1 = (x_{k_1}^{\text{start}}, y_{k_1}^{\text{start}})$, 依次从 $\{k_i\}, i=2, 3, \dots, n$ 上取出一 点 o_i, o_i 在满足与 o_{i-1} 连通的条件下, 距离 $(x_{k_i}^{\text{start}}, y_{k_i}^{\text{start}})$ 最近, 再平移 $\{o_i\}, i=1, 2, \dots, n$ 至于段的中轴后, 即形成细化结果。如选择终点进行细化, 则将上述过程中的 $(x_{k_i}^{\text{start}}, y_{k_i}^{\text{start}})$ 用 $(x_{k_i}^{\text{end}}, y_{k_i}^{\text{end}})$ 替换即可。

3.2 子段配对

设交点为 j , 由直接与之连通的所有子段构成的集合为 $\{g_i\}, i=1, 2, \dots, n$ 。对于每个 g_i , 可按如下方法确定一个对应的直线段 q_i : 如果 g_i 的细化结果仅包含一个点, 则 q_i 为该点与交点 j 质心的连线; 否则 q_i 为与 g_i 对应的多条直线段中直接连接交点 j 的直线段。若用 θ_{q_i, q_j} 表示 q_i 与 q_j 之间的夹角 (锐角), 则 θ_{q_i, q_j} 可以作为 g_i 与 g_j 之间连续性的度量。只有 $\theta_{q_i, q_j} < \mu$ (阈值) 时, g_i 与 g_j 才有可能配对, 并且 θ_{q_i, q_j} 越小, g_i 与 g_j 配对的合理性越大。这样求出所有的 θ_{q_i, q_j} 以后, 再设 G 为已经配对的所有子段构成的集合, 初始 $G = \emptyset$, 然后就可根据

$$\begin{cases} \min(\theta_{q_i, q_j}) \\ \text{s. t. } \theta_{q_i, q_j} < \mu, g_i \notin G, g_j \notin G \end{cases} \quad (4)$$

来确定配对结果 $\{g_i, g_j\}, G = G \cup \{g_i, g_j\}$, 直到配对结果为 \emptyset 。设定 g_i 与 g_j 之间夹角的阈值 μ 足够大 (30°), 以避免漏合并 (即同一线段的各子段没有合并在一起)。虽然这将使误合并 (即不同线段合并在一起) 的可能性增大, 但由于是采用角度最小化原则来进行配对, 因此只会造成具有三角形相交关系的线段的误合并。根据第 2 节的分析, 这种误合并对结果没有不利影响。

4 拟 合

Bernstein-Bezier 多项式是 3 维图形学中流行的

描述模型^[6], Liao 和 Huang, Lee 和 Wu 已将其应用在汉字笔划提取中, 用于度量分割后不同部分之间的连续性^[2,6]。本文采用 Bernstein-Bezier 多项式来描述合并以后形成的各个线段。设 P_0 与 P_m 为端点, P_1, P_2, \dots, P_{m-1} 为控制点, 则 Bernstein-Bezier 多项式为

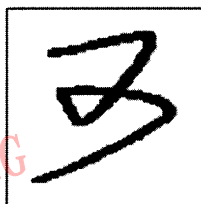
$$B(t) = \sum_{i=0}^m \binom{m}{i} P_i t^i (1-t)^{m-i} \quad \text{当 } 0 \leq t \leq 1 \quad (5)$$

线段包含交点和子段, 由于可以通过每个交点的质心和每个子段对应多直线的端点来确定其骨架, 因此这些点便是相应 Bernstein-Bezier 多项式中的 P_0, P_1, \dots, P_m 点。

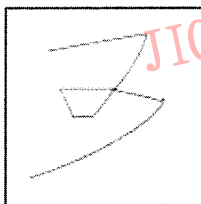
5 鲁棒性分析与实验

由前述可知, 分割是本算法稳定性的关键, 而第 2 节对交点边界特征的分析则是分割的基础。由于该分析充分讨论了线段相交的所有情况, 并找出了在各种不同情况下, 交点边界具有的特征, 因此整个分析是严格的。虽然根据分析结果提出的规则中有一个模糊条件, 但在第 2.1 节中已使其明确化, 并提出了段长直方图方法, 而且将其用于根据具体图像计算有关阈值时, 没有主观经验的影响。以上分析表明, 本算法是鲁棒的。

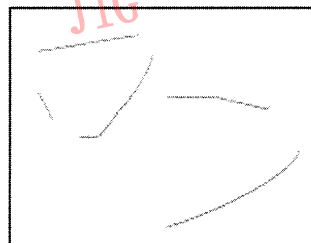
为了测试本文算法的有效性, 还针对手写体字符进行了实验。实验平台为 P4 1.6 型微机 (256M 内存), 实验对象为数字、英文和汉字各 100 个, 图像尺寸为 64×64 。本文算法对所有测试字符的细化正确率及线段提取正确率均为 98.5%, 对汉字的平均处理速度为 0.165s/字, 对数字和英文的平均处理速度为 0.047s/字。图 5~图 7 为 3 个实验示例及处理结果。基于前面的分析, 笔者认为实验结果中的误合并是正确的, 即细化正确率与线段提取正确率是一致的。



(a) 原始图像



(b) 细化结果



(c) 线段提取结果

图 5 手写数字“3”的实验结果

Fig. 5 The experimental result for handwritten numeral '3'

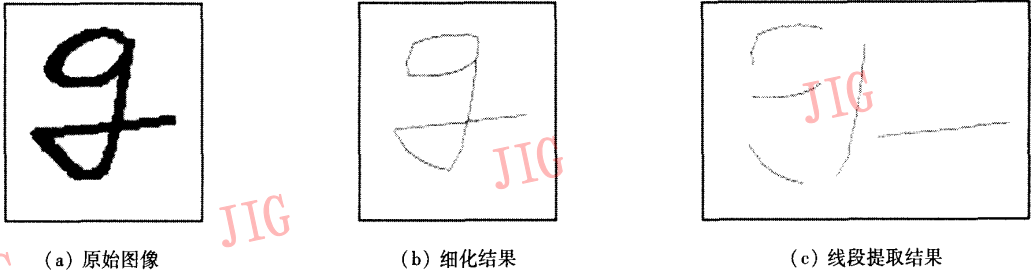


图 6 手写英文字母“g”的实验结果

Fig. 6 The experimental result for handwritten English letter 'g'

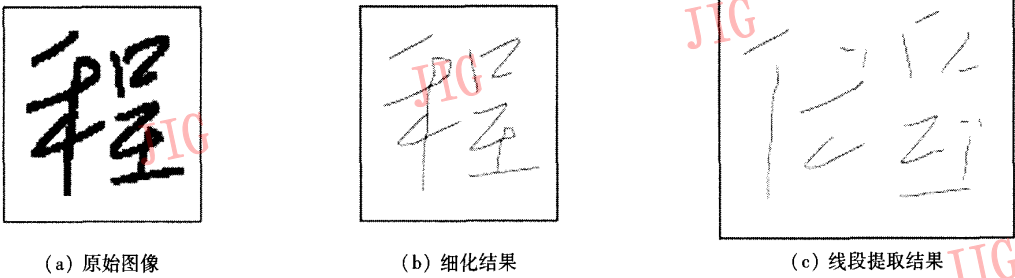


图 7 手写汉字“程”的实验结果

Fig. 7 The experimental result for handwritten Chinese character 'Cheng'

6 结 论

本文首先分析了待细化图像的特点,同时定义了待细化图像中的线段区域、交点区域和子段区域,并指出待细化图像是线段区域的集合;然后在此基础上,提出了同时进行图像线段提取和细化的算法。该算法首先基于图像行(列)连通像素段来提取图像中所有的交点区域和子段区域;然后根据子段区域之间的连通性和连续性将属于同一线段区域的所有子段区域和交点区域进行合并;最后用 Bernstein-Bezier 多项式拟合每个线段区域的骨架。鲁棒性分析及实验结果表明,这一算法是稳定有效的,可应用于图像处理、分析及理解的有关领域,而字符识别中的笔划提取可视为其典型应用。

参考文献 (Reference)

- 1 Liu K, Huang Y S, Suen C Y. Identification of fork points on the skeletons of handwritten Chinese characters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21 (10): 1095 ~ 1100.
- 2 Liao C W, Huang J S. Stroke segmentation by Bernstein-Bezier curve fitting [J]. Pattern Recognition, 1990, 23 (5): 475 ~ 484.
- 3 Liu C L, Kim I J, Kim J H. Model-based stroke extraction and matching for handwritten Chinese character recognition[J]. Pattern Recognition, 2001, 34 (11): 2339 ~ 2352.
- 4 Fan K C, Wu W H. A run-length-coding-based approach to stroke extraction of Chinese characters [J]. Pattern Recognition, 2000, 33 (11): 1881 ~ 1895.
- 5 Sun X M, Yang M J, Liu G H, et al. A totally knowledge-based stroke extraction method for Chinese characters [J]. Journal of Computer Research & Development, 2000, 37(5): 543 ~ 550. [孙星明, 杨茂江, 刘国华等. 完全基于结构知识的汉字笔画抽取方法 [J]. 计算机研究与发展, 2000, 37 (5): 543 ~ 550.]
- 6 Lee C N, Wu B. Chinese-character-stroke-extraction algorithm based on contour information[J]. Pattern Recognition, 1998, 31 (6): 651 ~ 663.
- 7 L' Homer E. Extraction of strokes in handwritten characters [J]. Pattern Recognition, 2000, 33 (7): 1147 ~ 1160.
- 8 Kegl B, Krzyzak A. Piecewise linear skeletonization using principal curves [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24 (1): 59 ~ 74.
- 9 He R, Yan H. Stroke extraction as pre-processing step to improve thinning results of Chinese characters [J]. Pattern Recognition Letters, 2000, 21 (9): 817 ~ 825.