

文章编号:1006-2467(2005)08-1288-04

一种通用的去除文字图像中干扰线的算法

程治国, 刘允才

(上海交通大学 图像处理与模式识别研究所, 上海 200030)

摘 要: 提出一种去除文字图像中干扰线的通用算法. 通过将图像转换为便于操作的基本图形及高级图形, 去除了图像的冗余同时获得了图像的几何信息, 然后将对干扰线的检测看作对图像主曲线的检测, 并将检测环节作为去除干扰线的关键步骤. 检测过程中采用改进的最短路径算法和方向偏移算法, 最后将检测到的干扰线与原图分离. 大量的实验结果证明, 该算法对文字图像中干扰线的检测和去除高效实用.

关键词: 图像识别; 干扰线; 主曲线; 基本图形; 超级图形

中图分类号: TP 391.4 **文献标识码:** A

A General Algorithm for Eliminating Interferential Curve of Text in Images

CHENG Zhi-guo, LIU Yun-cai

(Inst. of Image Processing & Pattern Recognition, Shanghai Jiaotong Univ., Shanghai 200030, China)

Abstract: The algorithm converts the images into basic graph and super graph, and then treats the interferential curve as principal curve to detect principal curve in images. In the detection, an improved shortest path algorithm and orientation offset algorithm are used, and finally, the detected curve is removed from the original image. The experiments conducted with a variety of text images show that this algorithm is effective for eliminating interferential curve of text in images.

Key words: image recognition; interferential curve; principal curve; basic graph; super graph

在计算机视觉、模式识别等领域中, 准确检测和识别图像中的文字是一项重要的研究内容. 例如, 光学字符识别(OCR)作为一项较为成熟的技术, 已得到多方面的应用^[1,2]; 同样, 在工程图、地图等对图像内容的识别研究中, 对文字符号的准确识别和提取也是十分必要的^[3~6]; 此外, 对视频中字幕的定位和提取^[7,8], 可以为基于内容的图像检索提供简洁方法.

然而, 在上述对文字图像识别的情况中, 很多图像中的文字会与其他无用图像重叠交织在一起, 给

文字的正确识别带来困难, 特别是在扫描的地图、工程图等图像的识别中. 干扰线是最常见的对文字符号的干扰. 本文中干扰线是指具有一定长度的光滑曲线, 允许平缓或尖锐的拐弯, 从另一个角度也可以称之为曲线. 在地图图像中, 主曲线可指道路边沿线、房屋轮廓线等, 而在文档性图像中, 主曲线可指与文字或字母接触的下划线或上划线, 也可以指文字的删除线, 或其他与文字叠合的曲线, 并且这些线型可以是直线, 也可以是具有一定平滑方向变化性的曲线.

收稿日期: 2004-09-02

基金项目: 国家高技术研究发展计划(863)项目(2002AA135210)

作者简介: 程治国(1977-), 男, 河南安阳市人, 博士生, 主要研究方向为计算机视觉、图像处理与模式识别.

刘允才(联系人), 男, 教授, 博士生导师, 电话(Tel.): 021-62933723; E-mail: whomliu@sjtu.edu.cn.

万方数据

在去除图像文字中的干扰线方面,许多研究文本分析、地图、工程图等学者都给出了一些方法,如 Xu 等^[1]采用自适应性合并原形提取算法和自由分割算法去改进 OCR; Reiher 等^[3]在地图符号识别中提出基于知识的方法,采用 Hausdoff 距离和神经网络方法进行识别; Whichello 等^[5]采用聚类 and 图像处理对信件邮政编码进行自动 OCR 识别。

本文提出一种通用的去除文档中干扰线的方法。该方法的思想是通过成功检测出文字图像中的干扰线,即主曲线,将其与图像中的文字部分分离,从而得到图像中的文字部分。其最主要的特点是不需要对文字符号的干扰线有任何先验知识,而是仅利用本身构造的数据结构和算法来进行检测和分割。

1 文字图像的图形表达

假设已对原始文字图像进行处理,得到二值特征图像,并且,文字和干扰线都经过细化,具有单一像素宽度。

对于给定的文本图像,无论是其中的文字部分,还是干扰线部分,都是以像素的形式存在,它们之间没有任何区分。因此,成功分离图像中文字和干扰线的第一步是将文字图像转化为图形,然后利用丰富的图形几何信息,在图形域中对干扰线检测。在图形域部分,又分为基本图形和超级图形。

1.1 基本图形

对于经过预处理和细化后的文本图像,其中的每一个像素都是从原始图像中得到的,保持了原始图像的几何结构。图像到图形的转化过程即为其中的每一个像素都转换为图形中的一个节点。本文定义基本图形节点具有 6 个参数用以描述图形节点的几何特征和拓扑属性,分别为节点编号、节点所属曲线号、节点位置、节点类型、节点连接链表、操作标志。其中,节点位置定义了节点在图像中的坐标;类型数表示节点与相邻节点的连接数目;连接链表描述节点和相邻节点的拓扑关系。在这些参数中,类型数参数特别有用,它给出了节点的拓扑连接数目,能够被用来区分曲线节点的类别。例如,类型数为 0 表明是孤立节点,类型数为 1 时表示曲线的端点,类型数为 2 时表示曲线的内部节点,类型数大于 2 时表示曲线的交叉节点,而且交叉节点的类型数参数等于该节点上的曲线分支数目。

基本图形在图像中通过 8 邻域聚类方法获得。采用 8 邻域聚类可以保证图形曲线有较好的连接性和光滑度,但是,也产生了图形表达的冗余。在定义万方数据

严谨的图形结构中,如果 2 个节点间的距离小于或等于像素栅格的 $\sqrt{2}$ 倍,则从其中一个节点到另一个节点的路径应该是唯一的。8 邻域聚类方法生成的基本图形会在相邻节点间产生多条路径,这种多路径连接称为图形的冗余。在后续的图形处理中会出现严重的问题。

基本图形中如果存在两两具有多条路径的相邻节点的集合,那么此图形就是冗余的,这些图形内部节点的数目被定义为维数,而与这些互连的内部节点相连的外部节点的个数,被定义为阶数。基本图形中的冗余采用如下办法消除:

(1) 如果一个图形的阶数小于等于 8,该图形能被简化为一个图形节点。简化后的节点的类型数等于该图形的阶数,该节点的连接链表是该图形的分支节点的链表。

(2) 如果一个图形的维数为 3,阶数为 1,则最接近分支节点的内部节点被保留并转换成图形节点,其余 2 个节点被去除。

(3) 如果一个图形的维数为 3,阶数大于 1 且小于或等于 3,则能与其他 2 个内部节点连成直角的内部节点被保留并转换成图形节点。

(4) 如果一个图形的维数大于 3,选定最接近节点集合几何中心的内部节点为新的图形节点,以代替原来图形的节点集合。

1.2 超级图形

经过冗余消除的基本图形可以准确地表达图形的原始拓扑结构。然而,由于基本图形是建立在像素级上的,对图形的操作一方面受限于图像像素量大小,影响处理速度;另一方面,它没有图形的几何特性表达,难于有效利用图像的几何性质。因此,本文建立了超级图形。

超级图形建立在基本图形的基础上,它具有 2 个特点:①对基本图形的进一步简化,即在超级图形中,消除了基本图形中的类型数为 2 的节点,因为它们大都属于连续曲线中间的点,没有特别的几何意义;②包含更多的图形几何属性,对从文字中检测出干扰线有极大的帮助作用。另外,需要注意到超级图形中没有类型数为 0 和 2 的节点。

超级图形包含两个重要部分:超级节点和超级链。超级节点是类型属性不为 2 的节点,除了含有类似于基本节点的属性外,一个超级节点还有两个额外属性:链链表和角链表。链链表存储了所有连接到该超级节点上的超级链的属性,而超级链提供了原始曲线的重要几何信息,从一个超级节点到另一个超级节点的曲线遍历信息由这 2 个超级节点间的超

级链的属性提供. 角链表提供了连接在超级节点上各超级链的方向信息. 超级链用下列属性来详细描述曲线的几何特征: 端点、长度、直线性、扭点表和扭角表. 这些属性中, 端点是超级链的终止节点. 长度属性用来衡量链的长度, 单位为像素. 对于直线性, 如果曲线实质上是直线, 则它对应的超级链的直线性为 0, 否则等于曲线在多边形化后的扭点数目. 扭点是曲线在多边形化处理后, 邻近两条直线段连接点. 扭点表提供了曲线拐弯位置的坐标和其他相关信息. 扭角表记录了超级链在扭点位置上的方向变化的角度. 超级图形提供了原始曲线的拓扑和几何信息的详尽描述. 图 1 为超级图形概念的示例.

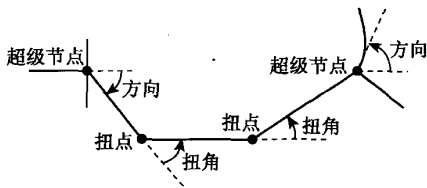


图 1 超级图形示例

Fig. 1 Super graph example

2 文字图像中的干扰线检测

对文字图像中干扰线的检测可以看成是对图像主曲线的检测. 所谓图像主曲线, 是指图像中那些感兴趣的目标的边界或轮廓曲线. 一般而言, 主曲线是具有一定长度的光滑曲线, 允许平缓或尖锐的拐弯. 本文处理的文字中的干扰线基本为直线或者平滑曲线, 因此, 对干扰线的检测即等效于对主曲线的检测.

由前面的讨论可知, 曲线网络可以用超级图形来表示. 在超级图形中的超级链表示了一段不包含分支的曲线, 这段曲线可能是直的或弯的, 长的或短的. 超级节点将超级链连接在一起; 超级节点的类型数范围为 1~8, 代表连接到该节点上的链的数目. 如果是 1, 则该节点是曲线网络的端点, 如果是 n , 则表示有 n 条链连接该节点.

在采用超级图形进行检测主曲线之前, 应该首先确定主曲线检测的判据. 一条主曲线可以是一条单一的超级链或几条超级链平滑地连接在一起. 如果 2 个超级节点间的一条路径是由几条超级链顺次地平滑连接组成, 则这条路径称为平滑路径. 有 2 个条件用以衡量路径的平滑性: ①最短路径特性. 如果 2 个节点间存在几条路径, 则平滑路径必定是最短的路径, 因为只有最短路径才能保证曲线的平滑连接. ②方向偏移. 这一尺度运用于长度小于可接受的万方数据

最短长度的短路径上. 方向偏移定义为

$$D = \sum_{i=1}^N \frac{l_i}{L} |\bar{\theta} - \theta_i|$$

其中: D 为路径的方向偏移; l_i 和 θ_i 分别为第 i 条链的长度和方向; N 为路径上的链的数目; $\bar{\theta}$ 为方向角 θ_i 的平均值; $L = \sum_{i=1}^N l_i$ 为所有路径的长度总和, 每个 l_i 均小于预设的路径长度, L 大于预设的路径长度.

算法中, 采用深度优先搜索算法 (DFS) 及最短路径和最小方向偏移约束对图形网络进行搜索. 假设网络中至少存在一个曲线端点, 主曲线的搜索从这些网络端点开始. 主曲线跟踪的计算步骤为:

(1) 读取一个未被访问过的网络端点节点 (类型数等于 1 的节点) 作为检测入口, 用深度优先搜索算法和最短路径算法延长成一条长度大于预设的门限的路径.

(2) 计算路径的方向偏移, 如果足够小, 则作为主曲线入口, 把入口节点作为根节点.

(3) 用深度优先搜索算法和最短路径算法从根节点开始搜索最短路径直到遇到一个超级节点, 使得这一路径的长度大于主曲线检测的约束, 并记这个节点为路径终点. 在曲线搜索中, 从同一根节点出发可能会在不同的方向找到若干个路径终点. 找出并记下从根节点出发的所有的路径和路径终点.

(4) 计算从根节点到路径终点的各条路径的方向并寻找一条在方向上和前面路径相似度最高的路径.

(5) 记录所选择路径的超级链.

(6) 把所跟踪的路径终点设置为根节点, 跳至步骤 (3), 直至所有超级链都被访问或方向偏差不足要求.

(7) 如果存在还没被访问过的端点, 跳至步骤 (1), 否则结束.

给定一个文字图像超级图, 用主曲线检测算法依次检查每个端点 (类型数为 1). 如果一条平滑路径存在, 则把它作为曲线跟踪的入口. 考虑到字体大小与主曲线长度间的比例是在一定范围内的, 故字体的笔画相对主曲线来说较小, 且很容易通过预设门限等条件排除, 因而, 本方法具有较强的适应性, 对文字、数字、符号中的主曲线检测都是可行的.

3 实验结果

上述操作过程中, 都是在细化后的单像素宽度内进行的检测, 因此文本中的干扰线找出后, 将其与

单像素文本分割相对简单. 实验结果图像中为了清楚起见, 将文字及干扰线进行了一定宽度的放大.

本实验在 Sun sparc 工作站上进行, 其内存为 512 MB, 以 C 语言实现, 实验结果如图 2 所示. 由图可见, 对于文字中的干扰线, 无论是直线还是光滑曲线, 本算法都能准确地将其识别出来 (见图 2(a)~(e)); 此外, 对于有尖锐拐角的干扰线, 本算法也能

成功识别 (见图 2(f)); 最后, 本算法还进行了多条干扰线的同时识别和长文字中大幅度干扰线的去除实验 (见图 2(g)~(j)), 其结果也是令人满意的. 同时, 因为本算法完全采用 C 语言实现, 其实现速度很快, 对于大小为 510×142 像素, 分辨率为 72 像素的图像, 其运算时间为 0.41 s.

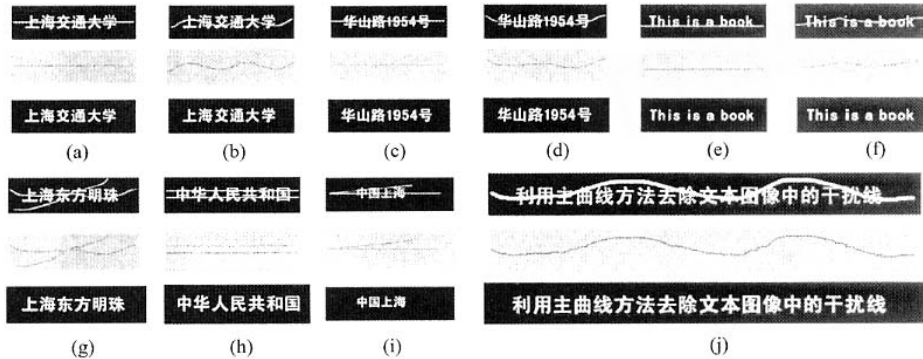


图2 文字图像中的干扰线消除结果图

Fig. 2 The experimental results of eliminating interferential curve of text in images

4 结 语

本文针对文字图像中存在直线或曲线等干扰线的情况, 提出了一种通用的去除干扰线算法. 本算法首先将图像进行预处理, 转换为高级图形, 然后在图形域进行以改进的最短路径算法和方向偏移算法为准则的干扰线的检测, 最后将检测得的干扰线与原图像文字进行分离, 以得到去除干扰线后的文字. 大量的实验结果证明, 本算法对于去除文本图像中的各种常见干扰线非常有效, 并且可同时去除文字图像中的多条干扰线.

参考文献:

- [1] Xu Y, Nagy G. Prototype extraction and adaptive OCR[J]. *PAMI*, 1999, 21(12):1280-1296.
- [2] Wakahara T, Kimura Y, Tomono A. Affine-invariant recognition of gray-scale characters using global affine transformation correlation[J]. *PAMI*, 2001, 23(4): 384-395.
- [3] Reiher E, Hayne C, Lalonde M, et al. A system for

efficient and robust map symbol recognition [A]. *ICPR96 (C90. 18)* [C]. Vienna, Austria; [s. n.], 1996. 783-787.

- [4] Trier O, Taxt T, Jain A. Recognition of digits in hydrographic maps: Binary versus topographic analysis [J]. *PAMI*, 1997, 19(4):399-404.
- [5] Whichello A, Yan H. Fast location of address blocks and postcodes in mail-piece images [J]. *PRL*, 1996, 17(11):1199-1214.
- [6] 杨前邦, 李介谷. 利用标记生长从边缘图像中抽取线段[J]. *上海交通大学学报*, 1999, 33(4):466-468.
YANG Qian-bang, LI Jie-gu. Using labeling and growing to extract line segments from edge images [J]. *Journal of Shanghai Jiaotong University*, 1999, 33(4):466-468.
- [7] Dori D, Velkovitch Y. Segmentation and recognition of dimensioning text from engineering drawings [J]. *CVIU*, 1998, 69(2):196-201.
- [8] Wong K, Chen M. A new robust algorithm for video text extraction [J]. *Pattern Recognition*, 2003, 36(6):1397-1406.