

字符识别中毛刺的去除方法

安 然,张少军,陈 华,喻振华

(北京科技大学 机械工程学院,北京 100083)

摘 要:在字符识别中提取字符特征前经常需要将图像细化,但细化过程会受到照明系统和字符本身质量的影响而产生毛刺。毛刺会使特征提取发生错误,进而影响最终的字符识别结果。针对此问题提出一种基于搜索毛刺生长点的毛刺去除方法,利用此方法可以将毛刺剔除并且保证图像的其它有用信息不被破坏。

关键词:字符识别;特征提取;细化;毛刺

中图分类号:TP391.41

文献标识码:A

文章编号:1673-629X(2007)09-0136-03

Method for Removing Burr to Optical Character Recognition

AN Ran, ZHANG Shao-jun, CHEN Hua, YU Zhen-hua

(College of Mechanical Engineering, Beijing University of Science and Technology, Beijing 100083, China)

Abstract: When recognize a character, it's often needed to thin the image of character before extract the structural feature. It will bring burrs arise from the effect of the system of lighting and the character's quality. The burr can make mistake when extract feature and disturb the result of character's recognition. A method is introduced for removing burr based on searching growth point of burr to resolve this problem. Using this method can remove the burr and ensure the useful part of the image not to be destroyed.

Key words: character recognition; feature extraction; thinning; burr

0 引言

随着工业自动化生产技术的不断进步产生了机器视觉技术。机器视觉是对采集的图像信息进行处理,根据处理信息进行测量或判断以控制下一步动作。这项技术大大提高了产品检测的自动化程度。在机器视觉的应用中有很大大一方面是针对字符的检测识别。如电话卡喷码质量检验、汽车牌照检验^[1]、纸币号码检验^[2]等等。这种检验的核心技术往往是利用成像方法对字符进行识别。字符识别的一般过程如图1所示。

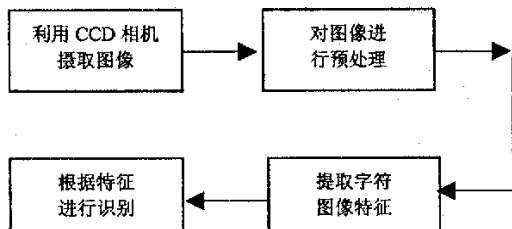


图1 字符识别的一般过程

字符识别的第一步是获取字符图像,在这一步中

要根据具体的项目要求合理选择相机和照明系统。尽可能地获取高质量的图像。第二步是对第一步获取的图像进行预处理^[3]。这一步的目的是将图像中有用信息提取出来,滤去噪声干扰,为下一步的特征提取创造条件。这一步的内容包括很多,要视具体情况处理。一般包括二值化、滤波、字符的裁剪和尺寸归一化、图像的细化、去除毛刺等。第三步是提取字符特征^[4]。字符特征有很多,要挑选能够反映不同字符间差异的特征。第四步是进行字符识别。字符识别的方法也有很多,有采用模板匹配的;有设计分类器的;有设计编码器的;有采用神经网络的。应选取识别率高识别速度快的方法。

1 毛刺的剔除方法

1.1 毛刺对字符识别的影响

对于字符的识别,无论是复杂的手写体字符还是印刷体字符,目前的主要方法有:利用特征分类的方法^[5]、模板匹配的方法^[6]和利用神经网络^[7]来进行识别等等。无论哪种方法一般都需要提取字符的结构特征,而为了提取结构特征往往需要对字符进行细化。对于字符这样较细长的对象可以通过细化得到较丰富的形状信息。目前细化算法有很多种,但无论哪种方

收稿日期:2006-12-13

作者简介:安然(1977-),男,北京人,硕士研究生,主要从事图像检测、图像识别等方面的研究;张少军,教授,主要从事图像测量和冶金设备方面的研究。

法,当受到光照条件、印刷质量和其它干扰噪声的影响时,经常会产生毛刺。这里需要说明的是在现有文章中毛刺的定义一般分两种:一种是图像未被细化前所存在的噪声,另一种是图像细化后产生的多余的分支,文中采用的是后一种概念。毛刺的产生是随机的,一旦经过细化处理后的图像产生了毛刺,很有可能改变原图像内容的结构特征,使提取特征时产生错误,进而造成识别错误。如图2(a)所示是字符“2”经过处理后的图像,(b)是它被细化后的图像。在(b)中的标号“1”、“2”、“3”表示字符的端点,正常的端点应该只有“1”、“3”两个点,。端点“2”是由毛刺产生的。这样,在提取字符的端点特征时,字符“2”的端点数本来应为2个,现在变成了3个,由于特征的提取错误必然会造成识别的错误。所以为了提高识别率,必须将毛刺剔除。



(a) 细化前 (b) 细化后

图2 字符“2”细化前后对比

在剔除毛刺的过程中,还要注意到不能在剔除毛刺的同时破坏字符原本有用的信息。所以,剔除毛刺时,既要将毛刺剔除干净,又要防止有用信息被破坏。下面介绍的方法经实验证明可以达到上述目的。

1.2 剔除毛刺的具体方法

在介绍方法前先做几个定义。由于图像是经过二值化的,只包括黑白两种颜色的像素,所以将黑色像素称为前景图,代表实际字符信息,将白色像素称为背景图,代表不包含字符信息的部分。将毛刺与实际字符像素的公共像素点称为毛刺的生长点。剔除毛刺的主要任务实际就是寻找毛刺生长点。具体方法如下。

在全图范围内搜索端点,其中包括伪端点,即将毛刺误认为端点。记录下搜索的各端点的位置。当每搜索到一个端点后就要从这一像素开始搜索毛刺生长点。搜索的步骤如下。

步骤1:先利用八邻域寻找与此端点像素相连的像素位置。由于判断是否是端点的规则时必须满足如图3的条件,即如果一个像素被认为是端点,以其为中心的八邻域必须是以下几种情况之一。其中字母“P”代表中心像素,数字“1”代表前景像素,数字“0”代表背景像素。

基于以上对端点的定义,则每个端点像素肯定可以在其八邻域内寻找到且只能寻找到一个相邻前景像素点,针对此点判断其是否是毛刺的生长点。判断方法如下。

素点,针对此点判断其是否是毛刺的生长点。判断方法如下。

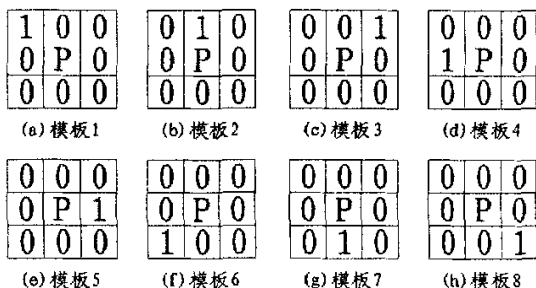


图3 用于寻找端点的八邻域模板

以此像素点为中心取八邻域,除去前一个搜索到的点,然后利用公式(1)计算八邻域的交叉数(交叉数的定义是八邻域像素中黑像素到白像素的变化次数)^[8]。如果交叉数大于1,则说明此时找到的点就是毛刺的生长点。

$$N(8) = \sum_{k \in C} (f(x_k)(1 - f(x_k))) \quad (1)$$

在这里 x_k 的位置如图4所示, $x_1 = x_9$, $C = \{1, 2, 3, 4, 5, 6, 7, 8\}$ 。

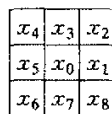


图4 公式(1)中 x_k 的位置说明

步骤2:如果利用上面公式计算的值等于1,说明还不能判断是否找到毛刺生长点。再利用四邻域模板继续搜索判断。如果此时以此像素为中心的四邻域中,除去以前搜索过的点存在两个以上黑色像素则说明此中心像素点就是毛刺生长点。否则四邻域模板应为如图5所示的四种情况之一,其中的字符所代表的含义与上述八邻域相同。由于此时四邻域中除去以前搜索过的点有且仅有一个像素为前景像素,找到此像素。

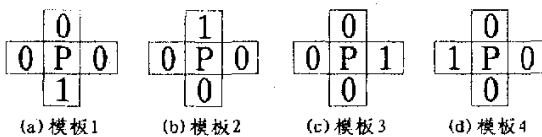


图5 步骤2中用到的四邻域模板

步骤3:将步骤2找到的像素除去以前搜索过的像素,再次利用步骤1中的公式(1)及步骤2判断其是否是毛刺的生长点。

根据经验,毛刺的长度一般不超过四个像素,所以在这里运用步骤1~4最多搜索三个像素,如果在三个像素内都没有找到毛刺的生长点,则说明搜索到的端点是图像中的真正端点,不需要做任何处理。如果在

三个像素内找到了毛刺生长点,则说明搜索到的这个端点不是图像中真正的端点,而是毛刺产生的“伪端点”。将“伪端点”和它与毛刺生长点之间的像素删除,即将毛刺删除。在全图范围内对所有搜索到的端点都进行上述处理,便可将细化图像上生长出来的毛刺都剔除。

2 实验结果与结论

文中的字符图像都经过了中值滤波、阈值分割等预处理,所采用的细化算法是 Zhang 细化算法。Zhang 快速并行细化算法具有速度快和保持细化后的连通性等优点,是一种实用的算法^[9]。图 6 是利用上面方法对几个原来有毛刺的字符图像进行处理的结果。从结果中可以看出,原来图像中的毛刺都被剔除了,而字符真正的端点都被保留,达到了预期的效果。图 6(a)(b)(c)三幅图中左边的是带毛刺的字符,右边是去除毛刺后的字符。

综上所述,文中所采用的毛刺去除方法可以有效去除字符图像细化后所产生的毛刺,并且能够保证其它有用信息不被破坏,字符的整体结构特征得到恢复,效果稳定。为下一步提取字符的特征信息,提高字符的识别率提供了有力保证。

参考文献:

- [1] 李 波,曾致远,周建中,等.车牌识别系统研究与实现[J].计算机技术与发展,2006,16(6):10-14.
- [2] 苑玮琦,张 昱.纸币号码识别中分割方法的研究[J].计

(上接第 135 页)

5 总 结

描述了一个基于 SVM 的加速的人脸检测系统,同时又保证了其相对的检测率。整个系统的提高是通过前两个阶段的分类器,这两个阶段的分类器加速了过滤背景图案的速度及计算时间,对检测潜在的人脸区域起到了很大的帮助作用。与选自 AdaBoost 学习机中的 Haar wavelet 特征的区别是:三层人脸检测系统将特征运用到了整个阶段的分类器中,并充分利用了它的高效展现和快速评估性能。前两个阶段分类器中的瀑布型结构能自适应各种不同的复杂性的输入对象。最后一个阶段中的非线性 SVM 分类器高速并有效地提供了检测的最佳结果。

参考文献:

- [1] Chellappa R, Wilson C L, Sirohey S. Human and machine

20050 20050

(a) 第一组字符

62829 62829

(b) 第二组字符

78725 78725

(c) 第三组字符

图 6 细化字符在毛刺去除前后的效果对比

算机测量与控制,2005,13(8):527-529.

- [3] 尚会超,陈幼平,余文勇,等.印刷字符在线检测的预处理算法及试验研究[J].计算机应用研究,2005(1):243-245.
- [4] 吴 锐,赵 巍,尹 芳,等.特征融合及相似度判据在英文识别中的应用[J].计算机工程与应用,2005(16):55-57.
- [5] 吴 谨,刘 洋.具有抗噪稳健性的数字识别[J].武汉科技大学学报,2004,27(4):406-408.
- [6] 崔 政,李 壮.两种改进的模板匹配识别算法[J].计算机工程与设计,2006,27(6):1083-1085.
- [7] 王建雄,刘应龙.基于人工神经网络的数字识别系统的研究[J].计算机技术与发展,2006,16(5):26-30.
- [8] 谷口庆治.数字图像处理基础篇[M].北京:科学出版社,2002:106-109.
- [9] 章毓晋.图像工程(中册)-图像分析[M].第2版.北京:清华大学出版社,2002:220-223.

recognition of faces: A survey[J]. Proceedings of the IEEE, 1995, 83(5): 705-740.

- [2] 周 杰,卢春雨,张长水,等.人脸自动识别方法综述[J].电子学报,2000,28(4):102-106.
- [3] 卢春雨.人脸自动识别若干问题研究与系统实现[D].北京:清华大学,1998.
- [4] Sung K, Poggio T. Example-based learning for view based human face detection[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1998, 20(1): 39-51.
- [5] Paul V, Michael J. Rapid Object Detection using a Boosted Cascade of Simple Features[C]//In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Kauai, Hawaii, USA:[s. n.], 2001.
- [6] Burges C J C. Tutorial on Support Vector Machines For Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.
- [7] 梁路宏.人脸检测与跟踪研究[D].北京:清华大学,2001.
- [8] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.