

基于数学形态学的手写体汉字细化方法研究

杨 侠,王希常,刘 江

(山东师范大学 信息科学与工程学院,山东 济南 250014)

摘要: 提出了一种基于数学形态学的细化方法,该方法使用结构模板的方式对字符图像进行细化,并针对原有细化方法产生的细化不彻底现象,对原结构模板进行了改进。在常用细化结构模板的基础上,新增了几个结构模板,较好地解决了细化不彻底的现象。实验证明,细化后的图像保持了原图像的连通性并达到了很好的细化效果。

关键词: 手写体汉字特征;字符细化;数学形态学

中图分类号: TP751.1

文献标识码: A

文章编号: 1674-7720(2011)20-0074-03

Research of handwritten Chinese character refinement based on mathematical morphology method

Yang Xia, Wang Xichang, Liu Jiang

(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: This article proposed a refinement method based on the mathematical morphology, which used the structure template to thin the character image. For the phenomenon of the original structure template thinned the character incomplete, the paper improved the original structure template. Based on the common structure template, this paper increased several new structure template, and better solution the phenomenon of refine incomplete. The experiment showed that the refined image used this method maintained the original image connectivity and achieved a good result.

Key words: handwritten Chinese character feature; detailed; mathematical morphology

随着光标字符识别(OCR)技术的不断发展和改进,手写体汉字的识别已成为当前研究的重点。由于汉字结构的复杂性和书写的随意性,手写体汉字的分割和识别一直是研究中的热点和难点,而手写体汉字的细化则是分割和识别的前提和基础。

由于个人在书写时所用的力度和工具不同,导致手写体汉字的笔划不均匀,对汉字结构特征的提取和识别造成了一定的难度。为了消除这个问题,需要在分割和识别之前,对字符图像进行细化处理。所谓汉字细化,主要是指通过一定的处理算法将汉字图像中那些不影响汉字特征和汉字拓扑结构的像素剔除,直至图像汉字的中心骨架形成为止,即细化为单像素宽度的图像^[1]。

在字符细化过程中,要保持字符的结构信息不发生改变,这就要求细化后的字符图像既要保持字符图像笔划的连续性,又要确保字符骨架尽量接近原图像笔划的中心线,无严重的畸变产生。基于数学形态学的图像细

化,能同时探测到图像的外部 and 内部信息,是一种应用较为广泛的图像细化方法。

1 字符细化的研究现状

字符细化有助于突出字符的结构特征,减少冗余的信息量,有利于字符的笔划提取和字符识别。常用的字符细化方法主要有:

(1) 迭代算法^[2-4]。该算法是在早期字符细化时最常使用的一种算法,主要是通过对比笔划边缘像素的迭代剥除来完成的。参考文献[4]中使用的是模板迭代去除算法,主要是通过设计一些去除模板和保留模板,将符合去除模板又不符合保留模板的像素点去掉,这样反复执行迭代过程,直到不发生变化为止。该算法的缺点是在细化过程中只考虑了字符的局部信息,容易造成笔划在拐角处的分叉和短笔划的缩进甚至消失。

(2) 基于笔划类型的细化算法^[1]。该算法与中轴变换算法^[2]具有相类似的特点,可以说是结合了笔划类型和中轴变换的一种方法。该算法在进行字符细化

技术与方法 Technique and Method

时,先判断出笔划的类型,再根据中轴变换算法求取该笔划的中心线作为细化后的骨架,从而避免了细化过程中造成的字型畸变。该方法需要事先提取汉字的笔划宽度,由于手写体汉字的特殊性,实际应用过程中较为复杂。

(3)基于数学形态学的细化算法^[5-6]。近年来,基于数学形态学的细化算法在字符细化过程中得到了广泛的应用。该方法主要是采用结构元素模板的方式对字符结构进行细化,其中,模板的选取对形态学方法的影响较大。参考文献[6]主要使用数学形态学中的腐蚀方法对字符进行腐蚀操作,直至细化为单像素为止。参考文献[5]主要使用形态学方法中的击中与击不中变换,结合印刷体汉字横平竖直的结构特点,对原有的结构模板进行了一定的改进。该方法的优点是能同时探测到字符图像的内部和外部,缺点是在细化过程中容易产生“毛刺”,并出现细化不彻底的现象。

由于字符结构等特性,在进行字符细化过程中容易产生各种问题,其中最常出现的有:(1)细化过程中容易产生“毛刺”,影响字符结构特征的提取;(2)耗时较长,算法复杂度高;(3)细化后汉字笔划的扭曲和歪斜;(4)容易出现细化不彻底的现象。

本文在分析手写体汉字的结构特点和其结构特性的基础上,采用基于数学形态学的方法,对手写体汉字进行细化。并针对细化过程中容易出现的细化不彻底现象,对原有的细化模板进行了一定的改进,使细化结果有了进一步的改善。

2 基于数学形态学的手写体汉字细化

近年来,基于数学形态学的细化算法在图像细化中得到广泛应用。本文采用数学形态学中的击中或击不中变换对数字图像进行细化,该细化方法能够同时探测到图像的内部和外部。

击中或击不中变换(HMT)是数学形态学中的基本运算。 A 被 B 击中或击不中变换可表示为 $A \otimes B$,其中, B 是结构元素对 $B=(B_1, B_2)$,而不是单个元素。击中或击不中变换的定义为: $A \otimes B=(A \odot B_1) \cap (A^c \odot B_2)$,其中, A^c 是 A 的补集, B_2 是 B_1 的补集, $(A \odot B_1)$ 表示 A 被 B_1 腐蚀, $(A^c \odot B_2)$ 表示 A^c 被 B_2 腐蚀。

2.1 图像预处理

为了更好地处理图像的细节部分,减少干扰因素,在进行细化之前,需要对图像进行预处理。首先将一幅手写笔迹图像通过扫描仪输入计算机中,并将其转化为灰度图像;然后根据需要对图像进行去噪、滤波处理,并对该图像进行二值化处理;最后再对图像进行反色变换,即将图像中字符像素部分变为1,背景点部分变为0,这样做的好处是将图像中被处理部分变为1,细化过程即简化为消除图像中多余的“1”像素点。图像的预处理结果如图1所示。



图1 字符图像预处理

2.2 结构元素模板

在基于数学形态学的细化算法中,模板的选取对细化结果影响很大,常用于图像细化的结构元素模板序列如图2所示。

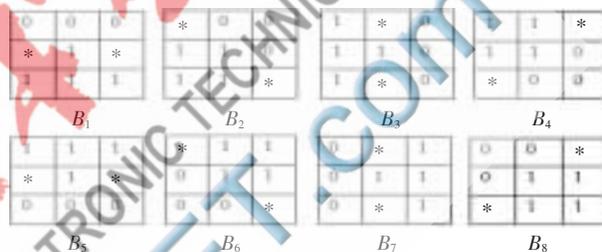


图2 细化过程中的结构元素模板

其中,结构元素模板中“1”表示字符笔划像素点,“0”表示背景像素点,“*”表示既可以是字符笔划像素点,也可以是背景像素点。

通过实验可以发现,仅仅使用上述结构元素模板序列容易造成细化不彻底的现象。本文根据手写体汉字字符图像的结构特征,利用改进的结构元素模板对细化后的字符图像再进行进一步的细化,降低了细化不彻底现象的发生。改进的几个结构元素模板如图3所示。

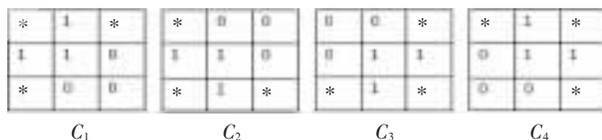


图3 改进的结构元素模板

2.3 手写体汉字细化

使用形态学方法中的击中与击不中变换对手写体汉字进行细化,细化过程可定义为: $A \otimes \{B\}=(\dots((A \otimes B_1) \otimes B_2) \dots) \otimes B_n$,其中,结构元素 $B=\{B_1, B_2, \dots, B_n\}$ 。利用改进的结构元素模板进行细化时,细化过程的结构元素序列可表示为: $M=\{B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8, C_1, C_2, C_3, C_4\}$,即在原有模板的细化基础上对细化后的图像进行进一步的细化处理,该细化过程在一定程度上改善了原有结构模板细化不彻底的现象。

3 实验

本文使用 Matlab 进行实验,首先输入一幅手写体汉

技术与方法 Technique and Method

字图像,并对原图像进行预处理,然后对图像进行数学形态学中的闭运算处理,即对图像先膨胀后腐蚀,目的是填充原始图像笔划中存在的孔洞,使图像笔划变得平滑,有利于下一步的细化操作。手写体汉字的细化主要是通过击中或击不中变换算法,使用结构元素模板对图像进行细化处理。在本实验中,首先采用常用的结构模板对原图像进行细化操作,再使用本文中改进的模板进行进一步的细化处理,实验结果如图4所示。

从以上手写体汉字的细化结果可以看出,基于数学形态学的细化方法很好地保持了字符图像笔划的连续性,很好地改善了原模板中细化不彻底的现象,细化后的骨架比较接近图像笔划的中心线,并保持了字符图像的拓扑性。

同时,本文算法仍有需要改进的地方,细化后的汉字笔划上有“毛刺”现象产生,由于手写体汉字笔划的结构特征等特性,在进行去“毛刺”处理时,易造成某些较短笔划的丢失。

参考文献

- [1] 闫建国,高华.基于笔划类型的字符细化算法研究[J].计算机工程与应用,2001,11(02):83-84.
- [2] LAM L, LEE S W. Thinning methodologies—a comprehensive survey[J]. IEEE, 1992, 14(9): 869-885.
- [3] 张学东,张仁秋.一种快速的手写体汉字细化算法[J].计算机应用与软件,2009,26(11):17-19.
- [4] 柳回春,马树元.手写体数字识别技术研究[J].计算机工程,2003,29(4):24-26.
- [5] 王建平,钱自拓.基于数学形态学的图像汉字笔画细化和提取[J].合肥工业大学学报,2005,28(11):1431-1435.
- [6] 金连文,徐睿.一种基于数学形态学的手写汉字方向特征提取方法[J].计算机工程,2003,29(20):38-39.

(收稿日期:2011-07-19)



图4 实验结果

作者简介:

杨侠,女,1984年生,硕士研究生,主要研究方向:数字图像处理。

王希常,男,1964年生,研究员,硕士生导师,主要研究方向:数字图像处理。

刘江,男,1977年生,硕士,工程师,主要研究方向:数字图像处理。