

## 基于连接段数细化算法的脱机手写体数字识别

陈志坚<sup>1</sup>, 成培<sup>2</sup>, 李峰<sup>3</sup>

(1,2,3 长沙理工大学计算机与通信工程学院, 长沙 410076)

**摘要:** 本文描述了已有的字符细化算法的思想及其缺陷, 在分析细化形变根源的基础上给出了一种新的快速细化算法, 并将该算法应用到脱机手写体数字识别过程中。该算法不会产生毛刺和伪分支点, 细化后字符骨架形变小, 而且速度也比较快, 从而提高了手写体数字的识别速度。

**关键词:** 细化; 连接段数; 骨架; 模板匹配; 字符识别

### 1. 引言

手写数字识别 (Handwritten Numeral Recognition) 在学科上属于模式识别和人工智能的范畴, 它是光学字符识别技术 (Optical Character Recognition, OCR) 的一个分支, 它的研究对象是如何利用计算机自动辨认人手写在纸张上的阿拉伯数字。几乎所有的光学字符识别 (OCR) 都是基于细化算法的, 因此, 细化算法的好坏很大程度上决定了 OCR 系统的好坏。细化的目的是要得到与原来区域形状近似的由简单的曲线组成的图形。在细化过程中需要考虑很多的准则, 其中最重要的是准确性, 同时细化应具备稳定性, 应当可以处理大量的较复杂模式, 在此基础上, 细化的速度还应该越快越好。一个好的细化算法可以减少细化造成的形变, 找到能反映字符真实形状的特征点, 使系统有较高的识别率; 相反, 一个不好的细化算法会产生伪特征点, 给字符的分类带来困难, 甚至导致误识或拒识。本文介绍了几种 OCR 领域常用的细化算法, 分析了其中的缺陷, 并且将一种新的快速细化算法应用于手写体的数字识别过程中。

### 2. 已有的细化算法及其缺陷

图像细化就是把二值图像中具有宽度的

线条状区域变成一条薄线 (即只有一个象素宽度)。图像细化大大压缩了原始图像的数据量, 并保持其形状的基本拓扑结构不变, 从而为文字识别中的特征抽取奠定了基础。细化算法应满足以下条件<sup>[1]</sup>:

- (1) 将条形区域变成一条薄线;
- (2) 薄线应位于原条形区域的中心;
- (3) 薄线应保持原图像的拓扑特性。

现在使用最多的细化算法是边缘侵蚀细化算法, 比如 Hilditch 细化算法、Pavlidis 细化算法等等。字符图像的边缘就是与背景点相邻的前景点的集合, 由于本算法采用的样本图像的背景为 1, 前景为 0, 因此经过去噪处理后的二值化字符图像的边缘就是与 1 象素点相邻的 0 象素点。边缘侵蚀细化算法就是循环检测笔道的边缘, 并且将边缘象素置为背景, 直到所有的笔道都为单象素宽度。很多现有的其他细化算法其实质也都是边缘侵蚀, 只是它寻找边缘的方法是采用轮廓跟踪的方式, 找到一个边缘点后, 搜索与该边缘点 8 邻域相邻的边缘点, 直到重新回到开始的搜索点<sup>[2]</sup>。这些算法最大的缺点就是会产生毛刺和伪分支, 而且还需要对字符图像做多次的边缘检测和去边缘操作, 算法运算量很大, 所以速度较慢。

### 3. 连接段数细化算法

<sup>1</sup> 陈志坚 (1963-), 男, 讲师; 研究方向: 图像处理, 网络和信息安全。

<sup>2</sup> 成培 (1983-), 女, 硕士研究生; 研究方向: 图像处理, 模式识别。

<sup>3</sup> 李峰 (1964-), 男, 教授, 博士, 硕士生导师; 研究方向: 图像处理与识别, 信息安全。

### 3.1 基本概念

为了算法描述的简单，对于二值化字符图像，我们定义以下几个概念：

#### 1) 邻接点

$$\begin{pmatrix} p(i-1, j-1) & p(i, j-1) & p(i+1, j-1) \\ p(i-1, j) & p(i, j) & p(i+1, j) \\ p(i-1, j+1) & p(i, j+1) & p(i+1, j+1) \end{pmatrix}$$

图 1 像素  $p(i, j)$  的邻接点

除了像素的边缘点以外，所有的像素都有 8 个邻接点。它们之间的关系可以用图 1 来表示。

#### 2) 连接段数

文章提出的连接段数是用来描述像素的邻接点的连通状况的，用图 2 来详细说明这个定义。文中使用二值图像，图中\*代表被处理像素，其邻接点的位置如图 2(a)所示，如果属于目标区域，灰度值为 1，否则为 0。

1	2	3	0	1	0	0	1	1	0	1	1
0	*	4	1	*	1	1	*	1	0	*	0
7	6	5	1	0	0	0	0	1	1	1	0
a	b	c	d								

图 2 连接段数

连接段数反映了邻接点的连通区。如图 2(b)所示，该像素有 4 个非零邻接点，其中位于位置 7 和 0 的点是连通的，所以该像素的连接段数为 3；在图 2(c)中，被处理的像素有 5 个非零邻接点其中位置为 2、3、4、5 的点是连通的，所以该像素的连接段数是 2；同理可得图 2(d)的连接段数为 2。

在获取像素的邻接点的个数和位置的过程中，就可以同时求得该像素的连接段数，从而组成了细化的删除规则。

### 3.2 具体细化过程

细化可以理解为一个连续剥离对象的最外层元素直到获得单位宽度的连通线（骨架）的过程，删除规则是算法的关键。细化的基本要求就是保持对象的连通性，同时细化后的骨架应尽量是原图像的中心线，它保持了原对象的几何特征，且其宽度是单像素的。由此而得，像素不能删除的条件是：

(1) 像素是孤立点或者是端点（即只有一个邻接点）；

(2) 如果像素被删除将破坏对象的连通性。

当像素至少有 2 个邻接点的时候，需要判断该点与其邻接点的位置关系，看删除该点是否会影响连通性，从而确定删除该点或者保留。图 3 列出了邻接点分布的一些情形，其中 a-e 的情形是不能删除的，而 f-g 的情形是可以删除的，因为删除该点不影响连通性。

0	1	0	0	0	0	1	1	1	0	0
0	*	0	1	*	1	0	*	0	1	*
0	1	0	0	0	1	1	1	0	1	0
a	b	c	d							
1	1	1	0	1	0	0	1	1	0	0
0	*	0	1	*	0	0	*	1	0	*
1	1	1	0	0	0	0	0	0	0	1
E	f	g	h							

图 3 邻接点的分布图例

根据以上论述，本文提出以下删除规则，它包含了适合删除被处理像素点的所有情况。由已经求出的邻接点的数量（用 node.num 表示），位置（用图 2 的表示方法[0]-[7]表示）和连接段数（用 node.count 表示）可以得到以下删除规则（其中，n[0]-n[7]表示相应位置上的邻接点的灰度值，待处理目标灰度值为 1，背景为 0）。

当 node.num ≥ 2 时，如果 node.count = 1，说明所有的邻接点都是相连的，如图 3 中的 g 和 h，此时该像素可以被删除；当 node.count ≠ 1 时，需要考虑 node.count 值和邻接点的位置。例如当 node.num = 2 时，只有当邻接点如图 4 分布时才可以被删除。依此类推，根据连接段数分析所有的可删除分布。

0	1	0	0	1	0	0	0	0	0	0
1	*	0	0	*	1	0	*	1	1	*
0	0	0	0	0	0	0	1	0	0	1

图 4 可删除邻接点

用二维数组存储给定的二值图像数据，其中目标灰度值为 1，背景值为 0。按照从上到下，从左到右的顺序，运用上述的删除规则来确定处理像素是否被删除，然后按照相反的方向做同样的操作。

如此交替运行直到没有点要被删除为止。

下面给出细化前后的字符的图像,如图 5 所示。

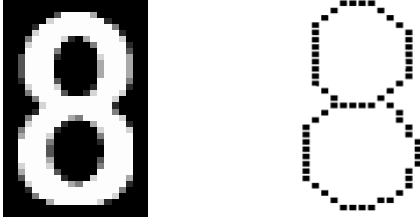


图 5 细化前后的字符图像

#### 4. 模板匹配

##### 4.1 模板匹配的基本原理

模板匹配方法是实现离散输入模式分类的有效途径之一,其实质是度量输入模式与样本之间的某种相似性,取相似性最大者为输入模式所属类别。它根据字符的直观形象抽取特征,用相关匹配原理进行识别,即将输入字符与标准字符在一个分类器中进行匹配<sup>[3]</sup>。以二维图像的处理为例,相关匹配算法如下:设输入字符用输入函数  $f(x, y)$  表示,标准模板用函数  $F(x, y)$  表示,在相关器中比较后输出为  $T(x, y)$ 。由于随机变量用相关量  $x_1, x_2$  表示,此时,相关器输出为:

$$T(x_1 - x_2, y_1 - y_2) = \iint f(x, y) F(x + (x_1 - x_2), y + (y_1 - y_2)) dx dy \quad (1)$$

当  $x_1 = x_2, y_1 = y_2$  时,

$$T(0, 0) = \iint f(x, y) F(x, y) dx dy$$

当  $f(x, y) = F(x, y)$  时,

$$T(0, 0) = \iint f(x, y) f(x, y) dx dy$$

即为输入字符的自相关函数,且有  $T(0, 0) \geq T(x, y)$ 。  $T(x, y)$  在  $T(0, 0)$  处出现主峰,在其它标准字符处出现一些副峰,只要这些副峰不等于主峰,就可用适当的阈值将其鉴别,从而判断并识别出被识字符。这种判别法在一定条件下,其错误概率和拒识概率最小。模板匹配法根据建模时所取特征的不同,有图形匹配、笔划分析、几何特征抽取等几种形式。图形匹配法在建模和匹配比较时,都

是基于字符的图形块本身而进行匹配,并根据其相似程度而得出识别结果,本文所采用的正是这种方法。在用上述图形匹配原理识别字符时,一般都是采用二值化字符:0 表示黑(背景),1 表示白(字符)。对每个字符均建立一个标准模板  $T_i$ ,待识别的图像为  $X$ ,它们的大小均为  $M \times N$ 。将未知的模式逐个与模板匹配,求出其相似度  $S_i$  为

$$S_i = \frac{\sum_{m=1, n=1}^{M, N} (X \times T_i)}{\sum_{m=1, n=1}^{M, N} T_i}$$

其中,“ $X \times T_i$ ”是指矩阵对应像素相乘。事实上,上式表示标准模板与待识别图像上对应点均为“1”的数目与标准模板上“1”点的数目之比。如果  $\max S_i > I$ ,则判定  $X = T_i$ ,否则拒识。这里为拒识阈值(一般由实验得出)。

##### 4.2 改进的模板匹配

模板匹配方法的缺点是抗干扰能力差。在传统的模板匹配方法中,标准模板的制作只是简单的选取一些质量较好的样本模板,然后取其平均值。这样,对于那些稍有变形、位移、旋转的待识模板,用模板匹配方法就很难进行判别。为此,我们设计了一种改进的模板,在原标准模板的基础上,通过对字符特征区域的扩大、加强注意并对那些类间距离差别较大的字符类型制作多个模板,来达到对模板的更有效的匹配结果。首先,对类间距离差别较大的字符类型进行细分,制作多个模板。如图 6 所示,将数字字符“0”细分为左倾、正、右倾三类,将“7”细分为两类,其中第二类用于克服字符切分不精确时带来的误差。



图 6 对类间距离较大的字符细分类别

经过细分后,对模板的特征区域进行扩大和加强注意,具体实现如下:

$$T_1(x,y)=\begin{cases} T(x,y), & \text{if } T(x,y) > \lambda_H; \\ \lambda_H, & \text{if } T(x,y) \leq \lambda_H \text{ and } T(x,y) \geq \lambda_L; \\ 0, & \text{if } T(x,y) < \lambda_L. \end{cases}$$

其中,  $T$  为原模板,  $T_1$  为改进后模板,  $\lambda_H$  和  $\lambda_L$  分别为上下阈值, 其值由实验得出。原模板  $T$  经过公式(5)中的变换后, 模板中的特征区域(如图 1 中的白字区域)被扩大和加强, 由此提高了模板匹配的鲁棒性, 减少了误识的可能。

## 5. 实验与结论

本实验采用  $20 \times 36$  像素的数字的图像, 并且提取了两种统计特征, 一种是网格特征: 把每一副图像分成  $3 \times 3 = 9$  个子图像, 统计每个子图像的黑像素的个数, 形成一个 9 维的矢量; 另外一种交叉点特征, 我们在水平及垂直方向三等分的地方, 作水平或垂直线穿过数字, 看其与数字边缘交叉的次数, 这样又得到 4 个数值, 总共形成一个 13 维的特征矢量<sup>[1]</sup>, 将模板中的特征提取出来保存到文件中去, 遇到一个新字符时, 提取它的特征矢量, 求出其与模板特征矢量的加权距离, 与之距离最小的即为识别结果。

表 1 本算法与传统算法的比较

模板匹配	传统算法	改进后的算法
数字的识别率	96.45%	97.86%
误识率	0.82%	0.65%
拒识率	2.73%	1.49%

由实验数据可以看出, 用改进后的细化算法和模板匹配法进行手写体数字的识别得到了比传统的模板匹配算法更好的结果, 识别率更高, 误识率和拒识率都降低了, 但是本算法在特征提取环节中提取的是数字的空间统计特征, 算法较原始, 所以在特征矢量提取方面还有待改进, 而且本文用到的细化算法对噪声比较敏感, 如果被处理的图像中存在孤立点噪声就无法去除, 不过可以通过前期处理, 如平滑来避免噪声的影响。

### 参考文献:

- [1] 张宏林, 蔡锐编著, 《Visual C++ 数字图像模式识别技术及工程实践》, 北京: 人民邮电出版社, 2003 年 2 月.
- [2] 恭才春, 刘荣兴, 面向字符识别的快速小形变细化算法, Vol. 22, No. 2, Feb., 2004.
- [3] 马俊莉, 莫玉龙, 王明祥, 一种基于改进模板匹配的车牌字符识别方法, Vol. 24, No. 9, Sep., 2003.
- [4] 高彤, 姜华, 吕民, 基于模板匹配的手写体字符识别方法, Vol. 31, No. 1, Feb., 1999.

# Recognition of Off-line Handwritten Numbers Based on Thinning Algorithm with Linked Section

Zhijian Chen<sup>1</sup>, Pei Cheng<sup>2</sup>, Feng Li<sup>3</sup>

(1,2,3 College of Computer and Communicational Engineering, Changsha University of Science and Technology, Changsha 410076, China)

**Abstract:** This paper depicts main ideas and defects of available thinning algorithms, and presents a new fast thinning algorithm with linked section based on analyzing origin of thinning distortion. The algorithm will be used in the recognition progress of off-line handwritten numbers. The algorithm produces no hairs or no spurious branches and makes the skeleton of the thinned character have less distortion. The speed of the algorithm is fast, and it can improve the speed of recognition of the handwritten numbers.

**Key words:** Thinning; Distortion; Skeleton; Pattern Matching; Character Recognition

(责任编辑: Francis, Sabrina, Ivan)