基于像素邻域点信息的藏文图像细化算法研究

刘 芳[1],张云洋[2]
(1. 西藏大学 藏文信息技术研究中心,西藏 拉萨 850000;
2. 西藏大学图书馆,西藏 拉萨 850000)

摘 要:细化是图像处理和模式识别系统中的一个重要过程,在图像分析和图像识别中应用广泛。只有把多像素的线条细化为单像素线条轮廓才能准确地进行字符的切分和文字的特征提取,对后续字符的分析和识别起着关键的作用。根据藏文字符的结构和书写特征,首先对藏文数字图像利用局部自适应方法进行二值化处理,再采用基于基线的滤波处理噪声方法进行去噪处理,以尽量简单直观地还原字符最原始的真实信息。在细化过程中,通过对某个像素点的八个邻域点的连接情况,在对照矩阵中查找对应矩阵项的值判断该点是否能删除,对藏文字符各点逐一进行判断和细化处理,最终得到文字的骨架。该算法在藏文字符数字图像细化实验中效果良好,正确率高,实用性强。

关键词:藏文;数字图像;二值化;去噪;细化;邻域点;对照矩阵
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2018)04-0021-04
doi:10.3969/j.issn.1673-629X.2018.04.005

Research on a Tibetan Image Refinement Algorithm Based on Adjacent Pixel Points' Information

LIU Fang[1], ZHANG Yun-yang[2]
(1. Titetan Information Technology Center of Tibet University, Lasa 850000, China;
2. Tibet University Library, Lasa 850000, China)

Abstract: Refinement is an important part in image processing and pattern recognition system, which has been widely used in image analysis and image recognition. Only by refinement of the multi-pixel lines into a single pixel line contour, the segmentation of characters and the feature extraction of the text can be carried out precisely, which plays a key role for subsequent analysis and recognition. In this paper, according to the structure and writing characteristics of Tibetan characters, the Tibetan digital image is processed in binarization by means of local adaptive method and then denoised by filtering method to deal with noise based on the baseline, in order to restore the original information of the characters as simple and intuitive as possible. In the refining, the refinement algorithm determines whether one pixel point can be deleted from its eight adjacent points' information. It judges Tibetan character's all points one by one, and refines them to produce the text's frame. In Tibetan character recognition experiments this refinement algorithm gets the positive results with satisfactory accuracy and strong practicability.

Key words: Tibetan; digital image; binarization; denoising; refinement; adjacent pixel point; control matrix

0 引 言

图像细化是图像处理的一个重要步骤,在减少图像存储空间,图像分析、特征提取和模式识别中占有重要的地位。在藏文信息处理中,藏文数字图像的细化同样重要,因为只有正确地提取出文字的骨架,才能准确、快速地提取出藏文字符的特征。

图像细化就是在不影响原图像拓扑连接关系下,将宽度大于一个像素的图形线条转变为单像素宽线条的处理过程,也就是提取图像的骨架[1]。图像的骨架就是图像中能精确描述图像结构特征的部分,字符图像的骨架就是能体现出字符中笔划的构成和走向等拓扑结构的部分。

传统的数字图像细化算法有很多,依据是否使用迭代运算可以分为两类[2]:第一类是非迭代算法,即一次产生文字图像骨架,如基于距离变换的方法、游程长度编码细化等。第二类是迭代算法,即重复删除图像

收稿日期:2017-04-19 **修回日期**:2017-08-24 **网络出版时间**:2017-12-05
基金项目:国家自然科学基金(61165010)
作者简介:刘 芳(1981-),女,硕士,讲师,研究方向为藏文信息处理技术。

边缘满足一定条件的像素,最终得到单像素宽带骨架[3]。常用的文字图像轮廓细化算法有基于数学形态学的处理方法[4],基于笔划趋势分析的二值图像细化方法[5],基于边缘细化的角点提取算法[6],等等。

在将藏文文本经过扫描仪等进入计算机后,由于纸张的厚薄、亮度、质量,机器的分辨率等原因都会造成字符的变化,也可能产生断点、污迹、粘连等干扰,这些不利因素都将严重影响字符轮廓的提取。因此,在细化以前,还需要对数字图像去除杂质,消除图像中无关的信息,尽量还原图像中的原始信息,最大限度地将图像中的真实信息更简单直观地显现出来,从而提高后续的字符切分、特征提取和字符识别过程的效率和正确率。

因此,在对藏文字符进行细化以前,需要对藏文的数字图像进行二值化、去噪处理,从而在后面的细化过程中得到更准确的字符轮廓。

1 藏文数字图像的二值化

通过扫描仪、数码相机等 CCD 成像设备输入到计算机中的图像都是多值图像。这就需要对图像进行标准化处理,即把多值图像转换成只有黑、白分布的二值图像,这一过程叫做二值化处理。二值化处理方法可以分成全局二值化方法和局部二值化方法[7]。全局二值化方法对每一幅图计算一个单一的阈值。灰度级深于阈值的像素被标记为背景字符(即白色),灰度级浅于阈值的像素被标记为前景(即黑色)。

局部二值化方法以像素的邻域的信息为基础来计算每个像素的阈值。

二值化首先需要把图像进行灰度处理,灰度处理使用如下灰度变换公式:

结果值=红色分量×0.299+绿色分量×0.587+蓝色分量×0.114+0.5

然后,令红色分量=绿色分量=蓝色分量=结果值即可。

灰度处理之后,便进行二值化处理,由于藏文文字表现形式为线条明显,如果二值化处理不当就会造成关键信息的缺失。在用常规方法对藏文字符图像进行二值化处理的过程中,最典型的是藏文字“ཀ”很容易识别成“ཧ”,从而导致后续特征提取和模式识别中产生错误。为了避免以上的错误,采用局部自适应二值化算法对藏文字符数字图像进行处理。

局部自适应二值化算法,是以像素的邻域信息为基础来计算每一个像素的阈值 k。设给定图像具有 L 级灰度值,对 $1<L<k$ 中的每个 k,将 1,2,…,k 分成两组,计算组 1 的像素数 $\omega_1(k)$,平均灰度 $M_1(k)$,方差 $\sigma_1(k)$;组 2 的像素数 $\omega_2(k)$,平均灰度 $M_2(k)$,方差 $\sigma_2(k)$,则:

组内方差:

$\sigma_W^2 = \omega_1\sigma_2^2$

组间方差:

$\sigma_B^2 = \omega_1\omega_2\ (M_1 - M_2)^2$

对于一幅给定的图像可以证明 $\sigma_B^2 + \sigma_W^2$ = 常数,因此只需求出 σ_B 的最大值,则 σ_W 自然达到最小,用这种算法处理得到效果较好的二值化图片。

2 藏文数字图像的去噪处理

一般来说,原始图像信息在记录、传输、获取等的过程中,都会受到各种噪声的干扰,主要原因有:源文件扫描过程中由于光线的原因,造成图像灰度不均匀,在某些地方产生的阴影;源文件磨损产生的字符笔划断裂;源文件的污迹产生的字符粘连及随机噪声等。

由于藏文的音节分割符号大小与图像中的噪声点几乎相同,甚至有的小于噪声点,如果采用传统的滤波方法很容易将该分割符号滤掉,造成字符切分和细化过程的错误。藏文文字行最大的特点就是:在行中所有的藏文字都以其基线为书写线,形成一条明显的直线,音节分割符正好在基线上,因此可以采用基于基线的滤波处理噪声方法。

首先对图像二值化后,在 Y 轴上进行映射,其中的明显波峰位置便是藏文字行的基线位置,记录基线位置 W_n,n 为基线个数[8]。设:两条基线之间的距离概率等于文字图像中所有单字的最大的文字高度 H,因为藏文字的高宽比例一般最大为 1∶2,且音节分割符号的宽度一般为基字宽度的 1/4,所以可以计算出音节分割符号大小 H_t 的范围:$H/8 \leqslant H_t \leqslant H/4$,其像素个数 F 的范围为:$H_t \times H_t$。

再采用递归的方法,递归黑像素周围八码,其中 P 为黑像素,P_1 至 P_8 中如果有一点为黑像素并且没有标识过已检查,则继续递归该点,并记录到区域 $T(x,y)$ 中,同时标识该黑像素已经检查。最后将区域 $T(x,y)$ 记录的个数与 F 进行比对,如果小于 F,并且区域 $T(x,y)$ 所在位置不在 W_n 上,则标识 P 点为变换点。将所有标识的变换点变为白像素,从而完成对噪声的去除。

3 基于像素邻域点的细化算法

3.1 藏文字的特征

藏文字是一种拼音文字,是基本字符和基本字符通过纵向叠加而成的字符串,构成一个完整藏文词素,基本单位是由藏文中的“音节分割符 tsheg bar”来确定。一个藏文词由一个或多个音节构成。每一个音节包含“基字”和可能跟随的如前加字、上加字、元音符

号、后加字、再后加字[9-13]。一个藏文字的各组成构件如图 1 所示。

图 1　藏文字的各组成构件

首先,藏文字线条明显,且文字中弧线较多,如果预处理中去除的信息过多会造成弧线的断裂或部分误删除[14]。

其次,藏文字的高宽比例会随着基本字符构成数量的不同而不同,也不是标准的方块字,很多笔画也不是规则的横、竖,只有精准的细化以后才能简化特征提取的过程。

在藏文文字的细化过程中,最容易出现的问题是细化以后文字的线条中间出现断节或者是线条中间出现一些小圆圈,最后看见的都不是文字的骨架,严重影响了特征提取和文字的识别[15-16]。

针对藏文文字的特征,提出了一种基于像素邻域点的藏文文字数字图像细化算法。对于二值化后的藏文文字数字图像,对构成文字的每个像素点进行处理,判断这个点是不是藏文字的骨架点,如果是就保留,否则就删除。经过多次实验,根据点的删除原则制定了一个统一的表格,在细化过程中查此表依次判断该点的删除情况。

对藏文文字的细化,最重要的就是判断一个点是否是图像的骨架,并确定删除一个点以后对图像的结果和连接关系是否有影响。算法中主要是根据某一个像素点的八邻域判断其删除情况。

文中算法采用了 256 级灰度图,只用到了调色板中的 0 和 255 两项。通过某点的八个相邻点的情况来判断点的删除情况。

要判断黑色的这个像素点能否被删除,就需要判断它的八个相邻点的情况。在八个相邻点中,与该像素相连的用黑色表示,与该像素不相连的用白色表示。对每一个像素点都用图 1 表示其八邻域的相连情况,再判断该点是否是文字的骨架点。

0	1	2
3		4
5	6	7

图 2　八邻域点信息实例 1

如果某个点的八个相邻点的情况如图 2 所示,这个点是不能删除的。因为该点是一个内部点,如果删除,就提取不到需要提取的文字骨架。

如果某个点的八个相邻点的情况如图 3 所示,那么该点可能是图像噪声或者是笔画粘连,就不是骨架点,是可以删除的。

0	1	2
3		4
5	6	7

图 3　八邻域点信息实例 2

3.2　细化算法中的删除标准

经过大量的实验统计发现,在细化过程中,只要所检测的像素点满足图 4 所示的条件之一,则该点即为文字的骨架不能被删除,否则就要删除。

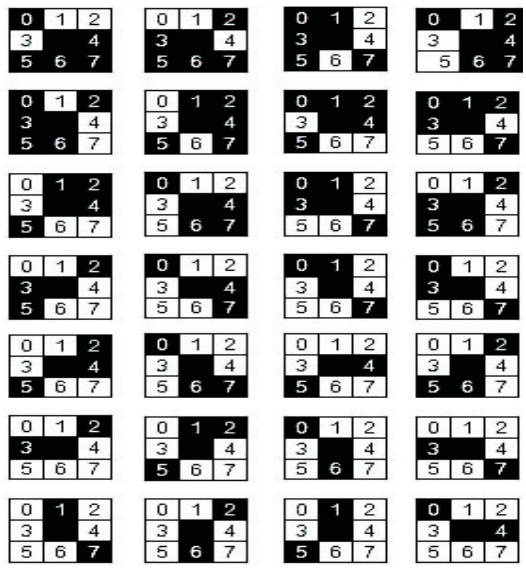

图 4　不能删除的像素点信息图

如图 4 所示,某个像素点的八邻域中,以上 28 种连接情况是该点不能删除的条件。把某个点的八个相邻点黑像素的数目称为该点的权,那些权为二的有 8 种情况;权为三的有 8 种情况;权为四的有 4 种情况;权为五的有 8 种情况。

3.3　细化处理过程

确定了像素点的删除标准后,采用数学矩阵来使用这个标准对每个点进行处理。

根据 8 位二进制所表示的范围制作了一个矩阵,里面表示 0 ~ 255 共 256 项,每一项的值为 0 或 1。在程序中,把八个相邻点中的白像素用 1 表示,黑像素用 0 表示。把八个相邻点从最低位开始分别乘以 2^0、2^1、2^2、2^3、2^4、2^5、2^6、2^7,然后到矩阵中查找对应矩阵项的值。如果查得为 1,则表示该点可以删除,否则该点就是骨架,不能删除。

像素点的八个相邻点的各种情况对照的矩阵如图

5 所示。

```
矩阵={
    0,0,1,1,0,0,1,1, 1,1,0,1,1,0,0,1,
    0,1,0,0,1,0,1,1, 0,0,0,0,0,0,0,1,
    1,0,1,0,0,0,1,0, 1,1,0,1,1,0,0,1,
    0,0,0,0,1,0,1,1, 0,0,0,0,0,0,0,1,
    1,1,0,0,1,0,0,0, 0,0,0,0,0,0,0,0,
    1,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,
    1,0,0,0,1,0,0,0, 1,1,0,1,1,1,0,1,
    0,0,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,
    0,0,1,1,0,0,0,0, 1,1,0,1,0,0,0,1,
    1,1,0,0,1,0,1,1, 0,0,0,0,0,0,0,1,
    0,0,0,0,0,0,0,0, 0,0,0,1,0,0,0,1,
    0,0,0,0,0,0,1,1, 0,0,0,0,0,0,0,0,
    1,1,0,0,0,0,0,0, 0,0,0,0,0,0,0,0,
    1,1,0,0,1,1,1,1, 0,0,0,0,0,0,0,0,
    1,0,0,0,0,0,0,0, 1,1,0,1,1,0,0,0,
    1,1,0,0,1,0,1,0, 1,1,0,0,1,0,0,0
};
```

图 5 细化算法中的对照矩阵

如果某像素的八邻域信息如图 6 所示,该像素的八个相邻点按以上规则进行计算为:$1*2^1+1*2^2+1*2^3=14$。再对照以上的矩阵,找到第 14 个值为 0,则表示该点是不能删除的。

0	1	2
3		4
5	6	7

图 6 八邻域点信息实例 3

在细化过程中,通过以上方法对藏文字的逐个点进行判断是否能删除,最后得到的就是文字的骨架。这些骨架保留了藏文字本身的结构特征,不影响其轮廓。减少了很多不需要的像素点后,既提高了特征提取的效率,又能正确地对藏文字进行模式识别。

4 结束语

藏文图像细化是藏文信息处理领域的一个很重要也很困难的方面,寻求一种图像骨架提取正确率高、图像失真率小的细化算法对藏文文字识别等结果有着至关重要的影响。针对藏文数字图像,提出了一种基于像素邻域点信息的细化算法。依次对图像中的每一个像素点根据其八个邻域点的连接情况判断是保留还是删除,最后得到图像的骨架,提取图像的信息。该算法的原理简单,在PC机上易实现,有效性和实用性都非常好。

参考文献:

[1] 李 甦,谭永龙. 基于生成树的图像完全细化算法[J]. 计算机工程与设计,2006,27(21):4006-4007.

[2] WANG W,DUAN M,YANG Y. The study of a new image thinning algorithm[C]//3rd international conference on electrical and electronics engineering. [s. l.]:[s. n.],2014.

[3] 程永上. 工程扫描图的矢量化和图档管理系统设计[D]. 南京:河海大学,2004.

[4] 韦 冰,陈宇拓. 基于数学形态学的图像轮廓细化方法[J]. 科技信息:学术版,2006(9):13-15.

[5] 刘桂雄,申柏华,冯云庆. 基于笔划趋势分析的二值图像细化方法[J]. 光学精密工程,2003,11(5):527-530.

[6] 邹琼兵,周东翔,蔡宣平. 基于边缘细化的角点提取算法[J]. 计算机应用与软件,2006,23(3):110-112.

[7] 刘 悦,刘明业,尚振宏. 快速响应矩阵码的多级阈值化方法[J]. 计算机应用研究,2006,23(8):177-179.

[8] 欧 珠,赵栋材. 基于水滴渗透算法的木刻经书藏文字切分研究[C]//2010 年全国模式识别学术会议(CCPR2010). 重庆:中国图象图形学学会,2010.

[9] 边巴旺堆. 基于 ISO/IEC10646 藏文编码字符集标准的藏文排序算法设计与实现[D]. 拉萨:西藏大学,2009.

[10] 春 燕. 藏文编码识别与转换算法的研究与实现[D]. 成都:西南交通大学,2010.

[11] MAI Z,XIANG W. Tibetan recognition system based on the improved extraction algorithm of complexity index feature [C]//5th international symposium on knowledge acquisition and modeling. [s. l.]:[s. n.],2015.

[12] 白玛玉珍. 藏文文字特征提取方法的研究[J]. 电脑知识与技术,2013,9(28):6362-6364.

[13] WAN F C,HE X Z,YU H Z,et al. Tibetan semantic information parsing for Tibetan - Chinese machine translation [C]//Proceedings of 2014 international conference on artificial intelligence and industrial application. [s. l.]:[s. n.],2014.

[14] YAN Xiaodong,ZHAO Xiaobing. Research on Tibetan text orientation identification[J]. Journal of Computers,2014,9(7):1598-1605.

[15] ZHOU D,HE W H,WU T T. The research on Tibetan text classification based on n-gram model[J]. Applied Mechanics and Materials,2014,543-547:1896-1900.

[16] KOJIMA M,KAWAZOE Y. Progress in automatic recognition of Tibetan Buddhist literatures by using object oriented design[J]. Joho Chishiki Gakkaishi,2008,18(5):481-482.