文章编号:1674 - 2869(2011)01 - 0084 - 04

OCR 中一种基于最小一乘法的连通域去噪方法

刘 渝[1,2],张彦铎[2*],鲁统伟[1,2]

(1. 武汉工程大学计算机科学与工程学院,湖北 武汉 430074;
2. 武汉工程大学智能机器人湖北省重点实验室,湖北 武汉 430205)

摘 要:提出了一种在 OCR 过程中利用最小一乘法分析连通域大小从而实现去噪功能的方法. 在图像中存在少量颜色与提取信息相似且大小具有明显差异的噪声的前提下,采用最小一乘法取代传统的最小二乘法对所有连通域的大小进行分析,从而获得需要的连通域阈值,实现去噪功能. 数据结果显示,在规定前提下该方法比经典的最小二乘法所确定的范围更为准确;实验结果显示,该方法能够达到去噪效果.

关键词:OCR;最小一乘法;连通域;去噪

中图分类号:TP391 **文献标识码**:A **doi**:10.3969/j.issn.1674-2869.2011.01.021

0 引 言

OCR (Optical Character Recognition 光学字符识别)技术,是指电子设备(例如扫描仪或数码相机)检查纸上打印的字符,通过检测暗、亮的模式确定其形状,然后用字符识别方法将形状翻译成计算机文字的过程. 在识别过程中,需要通过去噪获得只有要提取字符的图像. 根据实际图像的特点、频谱分布规律和噪声统计特征,对于去噪已经提出了很多研究方法,比如均值模板法、领域平滑法、连通域分析、滤波等等.

最小一乘法准则是 1760 年波斯科维奇在子午线长度研究中提出的. 由于计算上比较复杂,研究长期处于停滞状态,1805 年法国数学家勒让德在天文学和测地学研究中提出了最小二乘法,由于其计算上的方便和高斯等数学家的推崇,回归分析中多采用最小二乘法进行计算. 但是在降低异常值影响方面,最小一乘法更为优秀[1-2].

本例中,根据提供的对象,对具有相似特征的图片,对比最小二乘法提出一种基于最小一乘法的连通域去噪方法.

1 基于最小一乘法的连通域去噪方法

1.1 方法和思路

第一步,找到图像中的所有连通域,并获取它们各自所含像素点的个数;第二步,根据最小一乘法,计算出阈值中间值;第三步,根据阈值中间值确定阈值范围;第四步,根据阈值范围的限定对连通域进行相应的处理.

所提出的方法的创新之处在于以下几点:

a. 运用深度优先算法结合匹配矩阵查找连通域,在匹配矩阵中用不同的数字表识不同的连通域,提高后续工作的效率.

b. 运用最小一乘法对获取的连通域信息进行处理,从而找到阈值范围的中间值. 在 OCR 的背景下,根据字符特征和找到的中间值,确定阈值范围.

1.2 图像特征

本例中要处理的四幅图片尺寸均为 512 * 384,位深度为 24. 其特征为:

(1)待获取区域面积占总图像比例较小.

(2)待获取信息轮廓清晰.

(3)噪声与待获取信息颜色相似.

源图像如图图 1 所示.

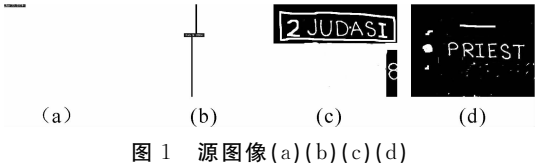


图 1 源图像(a)(b)(c)(d)

Fig. 1 Source images (a) (b) (c) (d)

图 1(a)和(b)中待获取信息放大 20 倍后图像如图 2、图 3 所示.

1.3 准备工作

第一步,利用 YIQ 公式将 RGB 图像转化为灰度图像.

收稿日期:2010 - 10 - 31

作者简介:刘 渝(1986 -),男,湖北宜昌人,硕士研究生. 研究方向:智能系统理论和应用.

指导老师:张彦铎,男,教授,博士,硕士研究生指导老师. 研究方向:智能系统理论和应用. * 通信联系人

图 2　图 1(a)待测目标放大图

Fig. 2　Amplify goal in Fig. 1 (a)

图 3　图 1(b)待测目标放大图

Fig. 3　Amplify goal in Fig. 1 (b)

第二步，利用最大阈值分割算法将灰度图像转化为二值图像.

第三步，利用深度优先算法对图像进行搜索，查找连通域，并对最终的结果进行不同的数字标识. 匹配矩阵标识效果如图 4 所示.

$$\begin{Bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 255 & 255 & 255 & 0 & 0 & 255 & 255 & 255 & 255 \\ 0 & 255 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 255 \\ 0 & 255 & 255 & 255 & 0 & 0 & 0 & 0 & 255 & 0 \\ 0 & 0 & 0 & 255 & 0 & 0 & 0 & 255 & 0 & 0 \\ 0 & 0 & 0 & 255 & 0 & 0 & 0 & 255 & 0 & 0 \\ 0 & 0 & 255 & 0 & 0 & 0 & 255 & 0 & 0 & 0 \\ 0 & 0 & 255 & 0 & 0 & 0 & 255 & 0 & 0 & 0 \\ 0 & 255 & 0 & 0 & 0 & 255 & 0 & 0 & 0 & 0 \\ 255 & 0 & 0 & 0 & 0 & 255 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

$$\begin{Bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 2 & 2 & 2 & 2 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \end{Bmatrix}$$

图 4　匹配矩阵标识效果图

Fig. 4　Result for match matrix show

第四步，根据先验知识，在连通域大小小于 10 的情况下，很难构成一个完整的字符，所以将连通域大小小于 10 的连通域变为背景，不列入后续的统计.

1.4　最小二乘法和最小一乘法

如果有 n 个数，$a_1, a_2, a_3, \cdots, a_n$，现在需要得到一个 x 使得它和这 n 个数的偏差 $a_1-x, a_2-x, \cdots, a_n-x$ 在总体上尽可能的小.

最小二乘法便是在这一思想上提出的，其表达式为 $D_1 = f_{\min}(x) = \sum_{i=1}^{n}(x-a_i)^2$；最小一乘法对于这一思想的表达则是 $D_2 = f_{\min}(x) = \sum_{i=1}^{n} |x-a_i|$. 根据本例中的情况，在一维状态下对比两种方法.

a. 从形式上看，D_1 表达的是一个圆域，解是唯一的；D_2 表达的是一个方域，解可能是不唯一的.

b. 从解的性质上看，求解 D_1 中的 x，令 $\frac{\mathrm{d}D_1}{\mathrm{d}x} = 2\sum_{i=1}^{n}(x-a_i) = 0$，则 $x = \frac{a_1+a_2+\cdots+a_n}{n}$ 时，D_1 达到最小值，所求解为 n 个数的算术平均数；求解 D_2 中的 x，令 $a_{(1)} \leqslant a_{(2)} \leqslant a_{(3)} \leqslant \cdots a_{(n)}$ 为 $a_1, a_2, a_3, \cdots, a_n$ 按由小到大的排列，即 $a_1, a_2 a_3, \cdots, a_n$ 的顺序统计量，①当 n 为奇数时，$x = a_{(\frac{n+1}{2})}$ ②当 n 为偶数时，$x = a_{(\frac{n}{2})}$ 或 $x = a_{(\frac{n}{2}+1)}$，所求解为 n 个数的中位数[1].

c. 从回归分析上看，最小二乘法在计算上更为简单，不需要分情况讨论，但是其受异常值的影响较大；最小一乘法缺乏显式表达，有时会出现多解的现象，但是其受异常值的影响较小[3].

本例中针对最小一乘法中多解的问题，可以通过求中间两个值的平均数得到解决[4]. 由于模板中小写字母 i 和大写字母 M 之间的差距最大，M 所包含的像素总和是 i 的 4 倍，所以在获得阈值中间值 x 后，确定的阈值范围为 $(\frac{x}{4}, 4x)$.

2　数据结果分析

将图 1 中(a)、(b)、(c)、(d)经过准备工作处理后得到四幅图的连通域个数，如表 1 所示.

表 1　图 1(a)、(b)、(c)、(d)中连通域个数

Table 1　Number of connected domain in Fig. 1 (a), (b), (c), (d)

图	图 1(a)	图 1(b)	图 1(c)	图 1(d)
连通域(个)	10	10	11	25

图 1 中(a)、(b)、(c)、(d)经过准备工作处理后得到的连通域大小数据如表 2、表 3、表 4 和表 5 所示.

表 2　图 1(a)中连通域数据

Table 2　Data of connected domain in Fig. 1 (a)

17	25	35	29	29	23	22	22	23	195528

表 3　图 1(b)中连通域数据

Table 3　Data of connected domain in Fig. 1 (b)

77218	22	25	28	28	115847	24	32	29	25

表 4　图 1(c)中连通域数据

Table 4　Data of connected domain in Fig. 1 (c)

830	104751	64	968	754	749	872	814	1577	591	9634

表 5　图 1(d)中连通域数据

Table 5　Data of connected domain in Fig. 1 (d)

1216	352	701	956	702	602	765	1132	541	137
58	34	27	35	222	12	25	21	30	22
18	17	11	46	15					

表 2 中的 11 个数据利用最小二乘法公式求得 $x=19\ 575.3$，阈值范围 δ_1 为[4 894,78 301]；利用最小一乘法思想求得 $x=\frac{23+25}{2}=24$，阈值范围 δ_2 为[10,95]. 根据实际情况最佳阈值范围 δ_3 应该是[17,35]，噪声阈值范围 δ_4 为[195 528,195 528].

表 3 中的数据共有 10 个，利用最小二乘法公式求得 $x=19\ 327.8$，阈值范围 η_1 为[4 832,77 311]；利用最小一乘法思想求得 $x=\frac{28+28}{2}=28$，阈值范围 η_2 为[10,111]. 根据实际情况最佳阈值范围 η_3 为[22,32]，噪声阈值范围 η_4 为[77 218,115 847].

表 4 中的数据共有 11 个，利用最小二乘法公式求得 $x=11\ 054.909$，阈值范围 φ_1 为[2 764,44 219]；利用最小一乘法思想求得 $x=830$，阈值范围 φ_2 为[208,3 319]. 根据实际情况最佳阈值范围 φ_3 为[591,1 577]，噪声阈值范围 φ_4 为[10,64]∪[9 634,104 751].

表 5 中的数据共有 25 个，利用最小二乘法公式求得 $x=307.88$，阈值范围 ψ_1 为[77,1 231]；利用最小一乘法思想求得 $x=46$，阈值范围 ψ_2 为[12,183]. 根据实际情况最佳阈值范围 ψ_3 为[541,956]，噪声阈值范围 ψ_4 为[11,352]∪[1 132,1 216].

通过求解表 2 和表 3 中的数据可以得出，$\delta_3\not\subset\delta_1$，$\eta_3\not\subset\eta_1$，说明利用最小二乘法没有能够获得正确的阈值范围；与此同时 $\delta_3\subset\delta_2$ 且 $\delta_4\not\subset\delta_2$，$\eta_3\subset\eta_2$ 且 $\eta_4\not\subset\eta_2$，说明利用最小一乘法不但获取了正确的阈值区间，而且没有包含噪声的阈值空间. 通过分析可以看出，当噪声不多，且需要获取信息的连通域大小彼此十分相近，相对于有着相同颜色的噪声所在的连通域大小具有明显差异时，将噪声所在连通域大小当做异常值进行处理，由于在一维状态下使用最小一乘法和最小二乘法的开销相差不大(一个进行排序，一个求平均数)，但是运用最小一乘法能够更好的减小异常值带来的影响，由此不难得出与 1.4 节中 **c** 吻合的结论，对于异常值过大或过小的情况运用最小一乘法应该能够得到更好的效果.

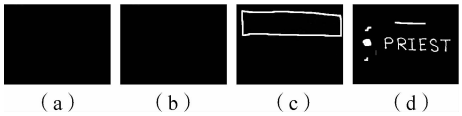
通过求解表 4 中的数据可以得出，φ_1 与 φ_3 没有交集却与 φ_4 有交集，说明利用最小二乘法不但没有能够获得正确的阈值范围，而且选出了噪声；与此同时 $\varphi_2\subset\varphi_3$ 且 $\varphi_2\not\subset\varphi_4$，说明利用最小一乘法不但获取了正确的阈值区间，而且没有包含噪声的阈值空间. 通过分析可以看出，在与表 2 和表 3 具有相同特征的情况下，缩小目标与噪声所在连通域大小的差距，最小一乘法仍然可以比最小二乘法更精确地命中目标.

通过求解表 5 中的数据可以得出，$\psi_1\supset\psi_3$ 且 ψ_1 与 ψ_4 有交集，说明利用最小二乘法获得了正确的阈值范围，但也选出了噪声；与此同时 $\psi_2\not\subset\psi_3$ 且 $\psi_2\subset\psi_4$，说明利用最小一乘法不但没有能够获得正确的阈值范围，而且选出了噪声. 通过分析可以看出，相比于前三种情况，本例中噪声较多，且与目标所在连通域大小差值不明显，利用最小一乘法错误的将噪声选成了阈值中间值，最终导致了区域的选择错误. 而此处利用最小二乘法则获得了更准确的结果.

通过以上数据分析可以总结出，当噪声不多，且噪声所在连通域大小与目标所在连通域大小具有明显差异时，利用最小一乘法能够取得比最小二乘法更准确的结果，这是因为异常值偏离实际中心过远导致的，最小二乘法对于异常值的影响要比最小一乘法敏感，据此说明具有类似特征的图片在使用连通域去噪时，利用最小一乘法进行处理要更为合理一些. 但当图像中噪声较多或大小区分不明显时，最小一乘法相比最小二乘法不具备优势[5].

3 实验结果

用最小二乘法对图 1(a)、(b)、(c)、(d)进行处理后效果如图 5(a)、(b)、(c)、(d)所示.

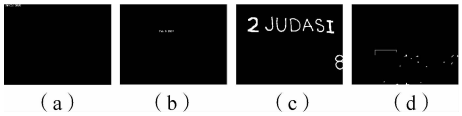


(a) (b) (c) (d)

图 5 最小二乘法处理效果图

Fig. 5 Result for dealing with least squares method

用最小一乘法对图 1(a)、(b)、(c)、(d)进行处理后效果如图 6(a)、(b)、(c)、(d)所示.



(a) (b) (c) (d)

图 6 最小一乘法处理效果图

Fig. 6 Result for dealing with least absolute deviation

图 6(a)和图 6(b)中的待获取信息放大 10 倍后如图 7、图 8 所示.

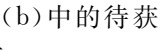


图 7 图 6(a)目标结果放大图

Fig. 7 Amplify goal in Fig. 6 (a)

图 8　图 6(b)目标结果放大图

Fig. 8　Amplify goal in Fig. 6 (b)

图 5 和图 6 的效果符合数据分析的预期结果. 从对比中可以看出：对于噪声不多，且噪声所在连通域大小与目标所在连通域大小具有明显差异的图像，利用最小一乘法能够取得比最小二乘法更准确的结果；对于噪声较多或大小区分不明显的图像，最小二乘法相比最小一乘法能够获得更准确的结果.

4　结　语

通过实验证明，最小一乘法对于解决异常值影响比最小二乘法更为优秀. 与传统的最小二乘法相比，最小一乘法更适合运用在目标轮廓清晰，噪声偏少且所在连通域大小与目标所在连通域大小相差很大情况下的连通域去噪中. 针对本例中出现的噪声与目标连通域大小相同的情况，在识别时通过匹配度门限的设置，就可以将通过连通域大小无法判别的噪声去除；针对噪声过多的情况，还有待以后的研究进一步完善[6].

参考文献：

[1] 王文平. 从算术平均数谈最小一乘法与最小二乘法的区别[J]. 武汉船舶职业技术学院学报，2009(6)：40－41.

[2] 李仲来. 最小一乘法介绍[J]. 数学通报，1992(2)：42－45.

[3] 夏勇，戴汝为，肖柏桦，等. 基于 OCR 与词形状编码的英文扫描文档检索[J]. 模式识别与人工智能，2009，22(3)：488－493.

[4] 徐国庆，张彦铎，王海晖. 基于多分辨分解的乐音水印算法实现[J]. 武汉工程大学学报，2008，30(2)：91－93.

[5] 张煜东，颜俊，王水花，等. 非参数估计方法[J]. 武汉工程大学学报，2010，32(7)：99－106.

[6] 胡学军，滕达，胡林文. 基于 MATLAB 的时滞对象控制算法仿真分析[J]. 武汉工程大学学报，2010，32(5)：92－95.

Method of connected domain denoise based on least absolute deviation in OCR

LIU Yu[1,2]，***ZHANG Yan-duo***[2]，***LU Tong-wei***[1,2]

(1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. Hubei Province key Laburatory of Intelligent Robot, Wuhan Instinte of Technology, Wuhan 430074, China)

Abstract: A denoise approach using analyzing size of connected domain by least absolute deviation is proposed in process of OCR. There is a little color which is similar to the needed information but different from it in size. Instead of the least square method, the least absolute deviation is proposed to analyze the size of the connected domain, get the threshold of the connected domain and denoise at last. The data results demonstrate that this method compared with the classic least squares method is more accurate under the precondition. The experimental results demonstrate that this method can achieve the desired effect.

Key words: OCR; least absolute deviation; connected domain; denoise

本文编辑：陈小平