基于最小二乘法拟合的汉字图像笔画点研究*

刘荣生，傅惠南，库才高
（广东工业大学 机电工程学院，广东 广州 510006）

摘 要：结合C程序，将预处理后的单个汉字图像与原图像进行逐个像素对比以判断读写，描写出原汉字字形。对经过预处理的单个汉字图像进行分析，提出了运用最小二乘法对二值化笔画点进行分组拟合的方法，从分布散乱的像素点中拟合出直线或曲线，画出汉字笔画，并计算相关系数、相关指数、残差及其平方和等参数，评估相关性、回归特性等拟合效果。最后，采用计算坐标平均的方法平整左右上下线，将其矫正成左右边竖直、上下边水平的口字形。

关键词：图像分析；最小二乘法；拟合；像素对比；汉字笔画
中图分类号：TP391.41 文献标识码：A 文章编号：1674-7720(2011)20-0044-03

Study based on least-squares fitting of the character image point of stroke

Liu Rongsheng, Fu Huinan, Ku Caigao
(College of Mechanical and Electrical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: With C programming, the paper proposes a single Character image after the pre-image and the original character pixel-by-pixel to determine reading and writing, and to describe the original Chinese Characters. Almost all single character pre-processing image are analysed. Least square method is proposed to binary grouping strokes point fitting method. From the distribution of scattered pixels in a straight line or curve fitting, it draws the character strokes and calculate the correlation coefficient, correlation index, residuals and their squares, and other parameters, assesses the correlation and regression fitting effect characteristics.

Key words: image analysis; least squares; fitting; pixel contrast; chinese Character strokes

由于数字图像的复杂性，至今仍没有一种通用的处理检测算法。在处理时，对被处理的图像有相当的依赖性，不同的算法都有其优点，同时也存在各自的不足[1]。

将原图像与预处理后的图像进行像素对比读写，从而描绘出与原图像相仿的汉字字形。该方法以C程序来实现，简单而实用。

通过预处理操作，单个汉字图像的笔画会变成一些看似有规律分布的像素点，其中，不少的像素点已经被处理掉，笔画变得断断续续、参差不齐，不好判断其原字形。应用最小二乘法进行拟合能将这些点按照某种规律连续起来，可以很大程度地还原笔画，为进一步的识别打下基础[2]。

曲线拟合中最基本和最常用的是直线拟合[3]。设 x 和 y 之间的函数关系为：

$$y(x)=a_0+a_1x \tag{1}$$

设已知数据点 (x_i,y_i)，$i=1,2,\cdots,m$，分布大致为一条直线。作拟合直线：

$$y(x)=a_0+a_1x \tag{2}$$

该直线不是通过所有的数据点 (x_i,y_i)，而是使偏差平方和为最小。偏差平方和为：

$$F(a_0,a_1)=\sum_{i=1}^{m}(a_0+a_1x_i-y_i)^2 \tag{3}$$

其中，每组数据与拟合曲线的偏差为：

$$y(x_i)-y_i=a_0+a_1x_i-y_i \quad (i=1,2,\cdots,m) \tag{4}$$

用最小二乘法原理估计参数，使 $F(a_0,a_1)$ 有极小值，a_0 和 a_1 应满足下列条件：

$$\begin{cases}\dfrac{\partial F(a_0,a_1)}{\partial a_0}=2\sum\limits_{i=1}^{m}(a_0+a_1x_i-y_i)=0\\[2ex]\dfrac{\partial F(a_0,a_1)}{\partial a_1}=2\sum\limits_{i=1}^{m}(a_0+a_1x_i-y_i)x_i=0\end{cases} \tag{5}$$

整理后得到正规方程组：

$$\begin{cases}a_0m+a_1\sum\limits_{i=1}^{m}x_i=\sum\limits_{i=1}^{m}y_i\\[2ex]a_1\sum\limits_{i=1}^{m}x_i^2+a_0\sum\limits_{i=1}^{m}x_i=\sum\limits_{i=1}^{m}x_iy_i\end{cases} \tag{6}$$

* 基金项目：国家自然科学基金项目(50675037)

解正规方程组便可求得直线参数 a_0 和 a_1 的最佳估计值:

$$\hat{a}_0=\frac{(\sum x_i^2)(\sum y_i)-(\sum x_i)(\sum x_i y_i)}{N(\sum x_i^2)-(\sum x_i)^2}$$

$$\hat{a}_1=\frac{N(\sum x_i y_i)-(\sum x_i)(\sum y_i)}{N(\sum x_i^2)-(\sum x_i)^2}$$

对于有拟合的显著性检验、偏差系数、相关系数等计算,将在下文的图像数据实验中具体分析给出[4]。

1 像素对比

预处理是绝大多数图像处理过程必需的步骤,有很多处理算法和技巧,但是在其之后所要进行的针对特定目标的识别处理则相当关键,像素对比读写法是一种新的简单算法,下面对算法原理进行介绍。

原图(图 1(a))经过灰度化、二值化、边缘检测、腐蚀运算、开运算处理后,得到字形轮廓图像,如图 1(b)所示,将其与原图像进行逐个元素对比读写,即当图像中像素值为 255 (白) 时, 则对应的原图像像素点保留不变,否则,变为 255。经过每一行的扫描对比后得到图 1(c),这样不仅避免了图 1(b)笔画过粗造成的误差,而且将原字形提取出来,背景变为白色,字形更加明显。

(a) 原图像 (b) 开运算 (c) 对比法

图 1 图像预处理

2 分组拟合

2.1 拟合评定参数计算

直线拟合的结果可由式(2)算出,当把观测数据点 (x_i,y_i) 作直线拟合时,用相关系数 r 来衡量一组测量数据 x_i、y_i 的线性相关程度[5]。

$$r=\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_i (x_i-\bar{x})^2\cdot\sum_i (x_i-\bar{y})^2\right]^{1/2}} \quad (7)$$

r 值介于 -1~$+1$ 之间,即 $-1\leqslant r\leqslant 1$。当 $r>0$ 时,为正相关;当 $r<0$ 时,为负相关。当 $|r|=1$ 时,全部数据点 (x_i,y_i) 都落在拟合直线上;若 $r=0$,则 x 与 y 之间完全不相关;r 值越接近 ±1,则它们之间的线性关系越密切。

相关指数 R^2 表示一元多项式回归方程估测的可靠程度的高低,$R^2=1-(\sum(y-y\text{ 估测值})^2\div\sum(y-y\text{ 平均值})^2)$。残差平方和 RSS 可由式(8)计算得到:

$$RSS=\sum[y_i-(a+bx_i)]^2 \quad (8)$$

相关系数检验部分数据如表 1 所示,例如,当 $|r|\geqslant r_{n-2}^{0.05}$ 时,认为 $|r|$ 在 0.05 水平上相关性显著;当 $|r|\geqslant r_{n-2}^{0.01}$ 时,认为 $|r|$ 在 0.01 水平上相关性高度显著。

2.2 直线附近数据点的拟合

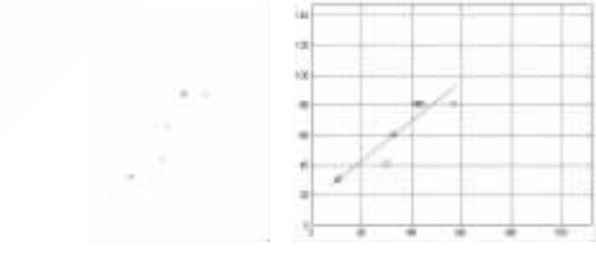
图 2(a)为一幅分布着黑色数据点的二值化图像,以图像左下角为坐标原点,数据点的坐标值如表 2 所示。黑色像素点近似分布在一条直线附近,因此对其进行最小二乘法直线拟合,拟合结果如图 2(b)所示。

表 1 相关系数检验表

$n-2$	$r_{n-2}^{0.05}$	$r_{n-2}^{0.01}$
1	0.996 9	0.999 9
2	0.950 0	0.990 0
3	0.878 3	0.958 7
4	0.811 4	0.917 2
5	0.754 5	0.874 5
6	0.706 7	0.834 3
7	0.666 4	0.797 7
8	0.631 9	0.764 6
9	0.602 1	0.734 8
10	0.576 0	0.707 9
11	0.552 9	0.683 5
12	0.532 4	0.661 4
13	0.513 9	0.641 1
14	0.497 3	0.622 6
15	0.482 1	0.605 5

表 2 图 2(a)实验数据点

序号	x	y	残差
1	10	30	0.621 4
2	11	31	0.295 3
3	30	41	−14.091
4	33	61	1.119 6
5	42	81	9.184 1
6	43	81	7.857 9
7	44	81	6.531 7
8	57	81	−10.708



(a) 直线点图 (b) 拟合结果

图 2 直线点图及拟合结果

计算结果与拟合效果评定如下:

斜率 $b=1.326\,167\,238\,195\,73$;截距 $a=16.116\,855\,710\,894\,2$;回归方程为 $y=1.326\,167\,238\,195\,73x+16.116\,855\,710\,894\,2$;相关系数:$r=0.929\,217\,368\,156\,058$,正相关很强;相关指数 $R^2=0.863\,444\,917\,282\,872$,回归效果很好;残差平方和为 527.222 104 985 4。

2.3 “口”字形笔画点

采用最小二乘法进行拟合的方法进行口字分组拟合直线,以提取“苦”中的“口”字形。鉴于像素点数量过

大不便于进行拟合实验，本文先进行二值化处理减少笔画点，再进行直线拟合。图 3 为选取各组不同阈值进行二值化得到的结果，可以看出，当阈值变小时，笔画点的数量也将减少。

从图 3 可以看出，当阈值取 60 时，笔画的数量合适，便于进行拟合分析，因此选取图 3(d)图进行坐标点赋值和拟合实验。

(a) 阈值为 128 (b) 阈值为 100 (c) 阈值为 80 (d) 阈值为 60

图 3 不同阈值的二值化结果

3 实验分析

3.1 实验数据

结合图像的存储结构进行 C 程序设计[6]，获得“口”字形笔画点的坐标共 43 个，如表 3 所示。

表 3 口字形笔画点实验数据

序号	1	2	3	4	5	6	7	8	9
x	69	70	71	73	74	74	46	47	45
y	0	0	0	1	1	2	5	5	7
序号	10	11	12	13	14	15	16	17	18
x	45	45	77	77	77	77	77	44	44
y	8	9	10	12	13	14	15	23	24
序号	19	20	21	22	23	24	25	26	27
x	78	44	44	44	78	43	43	48	49
y	24	25	26	27	28	29	30	34	35
序号	28	29	30	31	32	33	34	35	36
x	79	52	53	79	79	79	60	79	79
y	36	37	39	41	42	43	44	44	45
序号	37	38	39	40	41	42	43		
x	79	65	70	78	78	75	77		
y	46	47	50	50	51	52	52		

3.2 分段拟合

由表 2 获得的数据，对比放大图。为了获得口字形轮廓，根据数据点位置和坐标值，分 4 段进行直线拟合，如图 4 所示。

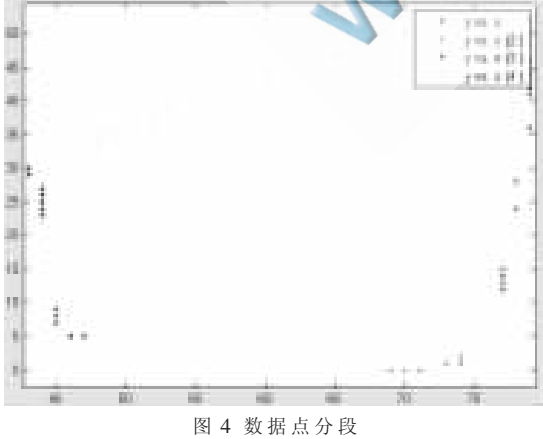


图 4 数据点分段

3.3 数据点计算分析与拟合图

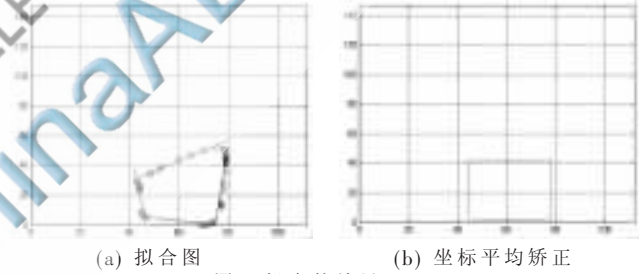
(1)右线选取序号为 5、13、14、15、16、19、23、28、31、32、33、35、36、37、40、41 等数据点进行直线拟合，回归方程为 $y=10.044\,083\,526\,682\,1x-751.248\,259\,860\,789$；相关系数 $r=0.832\,200\,397\,156\,514$，正相关很强；相关指数 $R^2=0.692\,557\,501\,027\,458$，回归效果较好；残差平方和为 1 206.385 150 812 07。

(2)底线选取序号为 1、2、3、4、5、6、7、8、9 等数据点进行直线拟合，回归方程为 $y=-0.185\,845\,047\,401\,111x+14.082\,870\,219\,025\,8$；相关系数 $r=-0.915\,704\,730\,983\,46$，负相关很强；相关指数 $R^2=0.838\,515\,154\,345\,491$，回归效果较好；残差平方和为 9.043 151 356 652 5。

(3)左线选取序号为 7、8、9、10、11、17、18、20、21、22、24、25 等数据点进行直线拟合，回归方程为 $y=-7.933\,333\,333\,33334x+371.2$；相关系数 $r=-0.902\,266\,317\,138\,268$，负相关很强；相关指数 $R^2=0.814\,084\,507\,042\,254$，回归效果较好；残差平方和为 215.6。

(4)上线选取序号 26、27、29、30、34、38、39、42 等数据点进行直线拟合，回归方程为 $y=0.681\,944\,444\,444\,444x+2.015\,277\,777\,777\,78$；相关系数 $r=0.993\,105\,425\,680\,536$，正相关很强；相关指数 $R^2=0.986\,258\,386\,516\,119$，回归效果很好；残差平方和为 4.665 277 777 777 77。

拟合结果如图 5(a)所示。



(a) 拟合图 (b) 坐标平均矫正

图 5 拟合的结果

3.4 口字形坐标平均矫正

由于是手写的汉字，拟合出的口字形左、右、上、下四边都是倾斜的，本文采用坐标平均的方法将其矫正成左右边竖直、上下边水平的口字形。由于右线选用的数据点分布在一条竖线附近，因此计算这些数据点的 x 坐标平均值作为右竖线的笔画坐标，$X_{右}=\sum_{i=1}^{16}x_i/16=77.93$。同理，分别对下线、左线、上线的数据点分布及坐标值用坐标平均的方法计算，$Y_{底}=\sum_{i=1}^{9}y_i/9=2.33$，$X_{左}=\sum_{i=1}^{12}x_i/12=44.50$，$Y_{上}=\sum_{i=1}^{8}y_i/8=42.25$。

将 $X_{右}$、$Y_{底}$、$X_{左}$、$Y_{上}$ 作为笔画的四边，得到矫正后的口字形，同时 4 个对角点位置由所得 4 个平均坐标组合得到，确定笔画的边界。坐标平均矫正的结果如图 5(b)所示。

本文对最小二乘法拟合原理计算公式进行了阐述，提出的像素点对比方法达到了提出汉字整体轮廓的目的。同时也提出将最小二乘法直线拟合运用到单个汉字笔画点字形提取当中，得到了与原字形相符的“口”字形笔画，拟合效果好，达到了预期的目的，同时为汉字的自动识别提取研究打下基础。

另外，对以下几个方面作进一步说明：(1)对于弯曲曲线的字形笔画，可以尝试用最小二乘法进行曲线拟合，同时，其他高等数学拟合方法也可以用来对笔画点进行分析；(2)汉字笔画点自动识别提取，实现对笔画点自动拟合；(3)笔画点二值化处理的阈值选择与笔画点数量的确定需要进行更客观的规律分析，达到阈值的优化选择，笔画点数量范围更大。

参考文献

[1] 朱辉,杨扬,颉斌,等.SVM 在小字符集手写体汉字识别中的应用研究[J].微计算机信息,2004(8-1):21-23.

[2] 樊钧,王润生.从图像中提取文字[J].国防科技大学学报,2002(01):59-62.

[3] 党兴菊,吴文良.最小二乘法拟合直线公式的推导[J].重庆科技学院学报(自然科学版),2010,12(4):185-187.

[4] 薛鹏涛,雷金山,肖立.土工直剪试验的最小二乘法拟合[J].中外公路,2007,27(5):41-44.

[5] 丁克良,沈云中,欧吉坤.整体最小二乘法直线拟合[J].辽宁工程技术大学学报(自然科学版),2010,29(1):44-47.

[6] 马建波.C 语言图像处理程序集[M].北京:海洋出版社,1992.

(收稿日期:2011-06-30)

作者简介：

刘荣生，男，1986 年生，硕士研究生，主要研究方向：超精密加工制造与检测。

傅惠南，男，1956 年生，教授，研究生导师，主要研究方向：微纳超精密技术。

库才高，男，1987 年生，硕士研究生，主要研究方向：金刚石研磨与动力学仿真技术。