关于限制性手写体汉字预处理探索

傅 彦
(电子科技大学计算机系 成都 610054)

【摘要】 就限制性手写体汉字的预处理问题进行了一些讨论,对通过摄像机摄入的限制性手写体汉字提出了一种预处理方法,对汉字采用了位置正规化、大小正规化、平滑处理、细化处理的相结合的方法,通过预处理的汉字,使汉字识别率大大的提高。

关 键 词 模式; 正规化; 近旁; 连通性; 连结数; 可删点
中图分类号 TP391.4

由于汉字的多样性和手写体汉字的不规范性,书写在一标准方框中的汉字的位置和大小仍有不同,它们之间存在着很大的差别,加之输入设备的影响,产生多余的信息和噪声。另外,这种汉字在其大小尺寸上与标准字典中的标准汉字的尺寸大小也可能存在很大的差异。为此,在进行汉字识别前必须对其进行适当的加工处理,即对汉字进行预处理。预处理阶段主要涉及如下一些内容:位置正规化、大小正规化、平滑处理、细化处理等。

1 位置正规化

输入汉字在机器内是以点阵为其表示形式,该点阵可以用一函数 f 表示,此 f 在点 (i,j) 的值为 $f(i,j)$,汉字点阵 f 是一离散函数;另用大写 F 表示对应模式,由于汉字点阵是一二值点阵,为此规定:当 $f(i,j)=1$ 时表示黑点,当 $f(i,j)=0$ 时表示白点。其中,设汉字点阵字模由 $M\times N$ 像点矩阵构成,为分析方便,扩充此阵为 $(M+2)\times(N+2)$ 的矩阵,同时规定

$$f(k,j)=0 \qquad (k=0,M+1;\ j=0,1,2,\cdots,N+1)$$

$$f(i,l)=0 \qquad (i=0,1,2\cdots M+1;\ l=0,N+1)$$

对于书写在标准方框位置不规则的汉字,移至方框中央,首先求汉字的重心位置 (i_0,j_0)

$$i_0=\frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{N}if(i,j) \qquad j_0=\frac{1}{n}\sum_{i=1}^{M}\sum_{j=1}^{N}jf(i,j)$$

式中 $n=\sum_{i=1}^{M}\sum_{j=1}^{N}f(i,j)$,此时网格中心为 $([\frac{M}{2}],[\frac{N}{2}])$,如 $|i_0-[\frac{M}{2}]|\leqslant 3$,$|j_0-[\frac{N}{2}]|\leqslant 3$,则约定不平移。否则将其文字的重心坐标 (i_0,j_0) 平移到 $([\frac{M}{2}],[\frac{N}{2}])$,而其他黑点 (i,j) 也进行相应平移

$$i\to i+|[\frac{M}{2}]-i_0| \qquad j\to j+|[\frac{N}{2}]-j_0|$$

此时若位置 (i,j) 处使 $f(i,j)=1$,则平移后,$f(i+|[\frac{M}{2}]-i_0|,j+|[\frac{N}{2}]-j_0|)=1$,由此而得到的新的二值点阵,其文字是位于点阵的中央。

1996年 12月 24日收稿,1997年 1月 14日修改定稿
* 女 33岁 硕士

2 大小的正规化

汉字的大小千差万别,必须将其压缩或扩大成 24× 24的标准二值图像(设标准汉字的尺寸大小为 24× 24网格点阵),同时测出文字图形的外接边框

$$i_B = \{i \mid (f(i,j) = 1) \wedge ((f(i+1,j+1) = 1) \vee (f(i+1,j) = 1) \vee (f(i+1,j-1) = 1))\}$$
$$i_E = \{i \mid (f(i,j) = 1) \wedge ((f(i-1,j+1) = 1) \vee (f(i-1,j) = 1) \vee (f(i-1,j-1) = 1))\}$$
$$j_B = \{j \mid (f(i,j) = 1) \wedge ((f(i-1,j+1) = 1) \vee (f(i,j+1) = 1) \vee (f(i+1,j+1) = 1))\}$$
$$j_E = \{j \mid (f(i,j) = 1) \wedge ((f(i-1,j-1) = 1) \vee (f(i,j-1) = 1) \vee (f(i+1,j-1) = 1))\}$$

其中 $0 < i < M+1$; $0 < j < N+1$; i_B, i_E, j_B, j_E 分别表示 i, j 方向左右、上下边框。其中原文字模式用 $F[0: M+1, 0: N+1]$表示,则切出包围 F 中的文字模式的最小方形子模式为 $F_1[i_B: i_E; j_B: j_E]$,对此子模式进行大小的正规化。利用曲面内插法按下式进行[1]

$$F_1[i_B: i_E; j_B: j_E] \xrightarrow{\text{扩大(缩小)为}} F_2[0: 0.25; 0: 0.25]$$

大小不规范性的原文字模式,经上面求出的 i_B, i_E, j_E, j_E 的纵横比值有时可能出现明显的大或小,直接处理时可能产生不必要的失真,必须给予修改。这里设其纵横比是 C_{ij},则 $C_{ij} = (i_E - i_B)/(j_E - j_B)$,则有以下三种情况:

1) 当 $C_{ij} < 0.5$时,文字成竖状,为此进行加宽: $i_B - ((j_E - j_B) - (i_E - i_B))/2 \Rightarrow i_B$; $i_E + ((j_E - j_B) - (i_E - i_B))/2 \Rightarrow i_E$; j_B, j_E 不变。 2) 当 $C_{ij} > 2$时,文字成横条状,为此进行加长: $j_B - ((i_E - i_B) - (j_E - j_B))/2 \Rightarrow j_B$, $j_E + ((i_E - i_B) - (j_E - j_B))/2 \Rightarrow j_E$, i_B, i_E 不变。 3) 当 $0.5 \leqslant C_{ij} \leqslant 2$时,不作任何修正。

经上述修正后,纵横比值 C_{ij} 趋于 1。若外接文字框的尺寸 $(i_E - i_B)$或 $(j_E - j_B)$大于某给定的阈值时,说明该输入图形为一汉字,必须进行正规化处理。设 (i,j)表示任意一个点,其中: $i = 0, 1, 2 \cdots 25$, $j = 0, 1, 2 \cdots 25$,则新模式 F_2 在点 (i,j)的取值为: $f_2(0,j) = f_2(25,j) = f_2(i,0) = f_2(i,25) = 0 (i = 0, 1, 2 \cdots 25, j = 0, 1, 2 \cdots 25)$; $f_2(i,j) = C_1 x(i) y(j) + C_2 x(i) + C_3 y(j) + C_4 (i = 1, 2, \cdots 24, j = 1, 2, \cdots 24)$。 $i_B \leqslant x(i) = i_B + [(i_E - i_B)(i-1)/(24-1)] \leqslant i_E$; $j_B \leqslant y(j) = j_B + [(j_E - j_B)(j-1)/(24-1)] \leqslant j_E$, $C_1 = Z_1 - Z_2 - Z_3 + Z_4$, $C_2 = l(j)(Z_2 - Z_4) + (l(j)+1)(Z_3 - Z_1)$, $C_3 = k(i)(Z_3 - Z_4) + (k(i)+1)(Z_2 - Z_1)$, $C_4 = Z_1(l(j)+1)(k(i)+1) - Z_2 l(j)(k(i)+1) - Z_3(l(j)+1)k(i) + Z_4 l(j)k(i)$, $k(i) = [x(i)]$, $l(j) = [y(j)]$,规定: $k(i) \leqslant x(i)$, $l(j) \leqslant y(j)$, $Z_1 = f(k(i), l(j))$, $Z_2 = f(k(i), l(j)+1)$, $Z_3 = f(k(i)+1, l(j))$, $Z_4 = f(k(i)+1, l(j)+1)$. 将 C_1, C_2, C_3, C_4 分别代入 $f_2(i,j)$中易知: $0 \leqslant f_2(i,j) \leqslant 1 (i = 1, 2, \cdots 24, j = 1, 2, \cdots 24)$。

为处理问题方便,仍将 $f_2(i,j)$转换成二值函数,为此进行阈值为 f 的二值处理。经上述扩大或缩小以后,就将其汉字的二值化点阵转化成一个标准大小和位置的二值化点阵。

3 汉字的平滑处理

如果书写时的笔质量不好,文字背景有时不小心溅上了油墨,加上光电转换装置本身的噪声等因素的影响,使经光电转换后文字笔划凹凸不平,背景中出现污点或在文字笔划中出现空白凹陷或毛刺等,这同样会给以后的工作带来一些不必要的误差,致使降低识别积累率。因此,必须预先对文字进行平滑处理,填上凹陷,除去毛刺,除去孤立点,使笔划平整光滑,以利细化。

首先考虑 3× 3窗口,如图 1所示。图中: $x_0 = f(i,j)$, $x_1 = f(i+1,j)$, $x_2 = f(i+1,j-1)$, $x_3 = f(i,j-1)$, $x_4 = f(i-1,j-1)$, $x_5 = f(i-1,j)$, $x_6 = f(i-1,j+1)$, $x_7 = f(i,j+1)$, $x_8 = f(i+1,j+1)$。当 $k > 8$时,令 $x_k = x_{k-8}$,当 $k < 1$时,令 $x_k = x_{k+8}$,要判断一个点 x_0 是否是毛刺或凹陷,可通过如

图 1的 3× 3窗口里的图像元素,使用明确的逻辑规则来判断。

x_4	x_3	x_2
x_5	x_0	x_1
x_6	x_7	x_8

图 1　3× 3窗口示意图

1) 如 x_0是一个黑点,则需判断 x_0是否属于毛刺,为此规定: 如一黑点出现连续环绕它的五个点都是白点,而其余三个点满足一定的条件,则判 x_0是毛刺。

(1) 如 x_0= 1,其余 x_i= 0(i= 1, 2… 8),则 x_0为孤立点。 (2) 如 x_0= x_6= x_7= x_8= 1,其余 x_1= x_2= x_3= x_4= x_5= 0,则 x_0为毛刺。 (3) 如 x_0= x_6= x_7= 1,其余 x_1= x_2= x_3= x_4= x_5= x_8= 0,则 x_0为毛刺。(4) 如 x_0= x_6= x_8= 1其余 x_1= x_2= x_3= x_4= x_5= x_7= 0,则 x_0可能是连接点。 (5) 如 x_0= x_7= 1,其余 x_1= x_2= x_3= x_4= x_5= x_6= x_8= 0,则 x_0可能是端点。

将上述旋转 90°的整倍数,则包含一切情形: 对于 (1)~ (3)应删去,对于 (4)~ (5)应保留。

上述删除点的条件可由下式表示: 如 x_0= 1,则 (1)$\sum_{k=1}^{8} x_k$= 0; (2)$\sum_{l=k}^{k+4} x_l$= 0并且$\sum_{l=k+5}^{k+7} x_l$= 3(k= 1, 3, 5, 7); (3) $\sum_{l=k+7}^{k+4} x_l$= 0并且$\sum_{l=k+5}^{k+6} x_l$= 2(k= 1, 3, 5, 7)之一成立,则该黑点应该删去,否则保留。

2) 如 x_0是一个白点,即 x_0= 0,则需判断 x_0是否属于凹陷,为此规定: 如果环绕某个白点 x_0的五个点都是黑点或者 x_1, x_3, x_5, x_7都是黑点,则判 x_0是一个小凹陷。即如 x_0= 0时,若 (1)$\sum_{l=k}^{k+4} x_l$= 5 (k= 1, 3, 5, 7); (2)$\sum_{l=1}^{4} x_{2l-1}$= 4之一成立,则此时该白点应该填上。

4 汉字的细化处理

尽管对手写汉字加了一定的限制,但毕竟不可能绝对标准,难免会出现很多不规范的地方。 如在书写时文字用力不均,笔来水不均,以及光电转换的影响,造成了汉字笔划的粗细不均匀,这就产生了很多不必要的信息。 因此,对粗细不均的汉字笔划统一进行标准型的细化处理,减少识别的工作处理量,以便于分离和抽取汉字的特征。

细化和平滑处理有某些相似之处,细化一般采用对汉字进行层层剥离的技术。 根据笔划的粗细,一个字一般要经过若干次的剥离才能达到细化的目的。当然对细化后的汉字一定不能破坏原文字图像的形状,并保证细化后的文字图像的连通性。

仍采用平滑时所示的 3× 3窗口,为保证图像的连通性,用一个刻划连通性的连通数 N_{C4}(或 N_{C8})来反映,为此首先给出一些基本定义[2]。

定义 1　称 $\{x_k \mid k \in S\}$为 x_0的 8近旁, $\{x_k \mid k \in S_1\}$为 x_0的 4近旁,其中: S= {1, 2,… 8}, S_1= {1, 3, 5, 7}。

定义 2　对于具有相同值的两个元素 a_1, a_2,如存在着与此具有同样值的元素系列 y_0(= a_1), y_1, y_2… y_n(= a_2),对于全部的 i($1 \leqslant i \leqslant n$), y_i是存在于 y_{i-1}的 4(8)近旁时,称元素 a_1和 a_2在连结意义下是可连结的。

定义 3　设 x_0是一个黑点,则 x_0的连结数 N_c可由下式定义

$$N_{C4} = \sum_{k \in S_1} (x_k - x_k x_{k+1} x_{k+2}) \qquad N_{C8} = \sum_{k \in S_1} (\bar{x}_k - \bar{x}_k \bar{x}_{k+1} \bar{x}_{k+2})$$

其中　S_1= {1, 3, 5, 7}; $\bar{x}_k$= 1- x_k。

定义 4　在文字图像中,只要两个点间是 4(8)连通的,则称该两点是连通的。

定义 5　如果一个点的删除不影响文字的连通性,称该点为 4(8)—— 可删点。

定理 1 一个点 x_0 是一个 4(8)—— 可删除点当且仅当 $N_{C4}=1$(或 $N_{C8}=1$)

性质 1 根据连结数 $N_{C4}(N_{C8})$的值,边界点可分类如下:

$$N_{C4}=\begin{cases}1 & \text{端点}\\ 2 & \text{连结点(内点)}\\ 3 & \text{分岐点}\\ 4 & \text{交叉点}\end{cases}\qquad N_{C8}=\begin{cases}1 & \text{端点或三分岐点}\\ 2 & \text{内点}\\ 0 & \text{交叉点或孤立点}\end{cases}$$

从以上的性质可知,当 $N_{C4}=1$或 $N_{C8}=1$时,并不一定是可删点,此时,可能是端点或分岐点。为此定理 1可改述为定理 2

定理 2 1) 一个点 x_0 是一个 4(8)—— 可删除点当且仅当 $N_{C4}=1$(或 $N_{C8}=1$),并且 x_0 不是端点或三分岐点; 2) 一个点 x_0 是一个 4(8)—— 可删除点当且仅当 $N_{C4}=1$(或 $N_{C8}=1$),并且只存在唯一的一个 $k\in S$ 使得 $x_k=1$,其余的 $x_l=0(l\neq k, l=1, 2,\cdots 8)$这样的情形不成立。

任用 N_{C8}来判断时,除端点外(同 N_{C4}一样),还存在三分岐点,此时,除 $x_k=x_{k+2}=x_{k+4}=1$外($k=1, 3, 5, 7$)其余的 $x_l=0$

为方便起见,仍采用 3× 3窗口进行讨论,同时 x_i 与 x_0 是 4连通的,则此时有一个点 x_0 是 4—— 删除点当且仅当 $N_{C4}=1$,并且 x_0 不是端点。

为了使汉定图像在细化后结果尽量是原文字笔划中心线,程序设计采用多方向扫描进行反复剥离,细化中,为了不使删除影响下一步细化,要进行并行处理,一般在该点不被利用时才删除它。

规定扫描方向的顺序为: 细化方向: (1) 上→下; (2) 左→右; (3) 下→上; (4) 右→左。 对应的扫描点移动方向: (1) 左→右; (2) 下→上; (3) 右→左; (4) 上→下。反复沿上述的扫描方向顺序进行细化处理,直到笔划为 1 bit为止。

通过对汉字的位置正规化,大小正规化、平滑处理、细化处理等预处理,可大大提高汉字识别的正确率,为此,汉字的预处理在汉字识别中有着十分重要的作用。

参 考 文 献

1 Duda R O, Huet P E. PaHezn Classification and Scene Analysis. New york, 1989

2 Fujimucza O. Computer Input-Output of Chinese Characteas. Wisconsin, 1992

Probing of Preprocession Based Limited-handwritten Chinese Character

Fu Yan

(Dept. of Computeas, UEST of China Chengdu 610054)

Abstract The limited-handwritten chinese character is discussed in this paper. The preprocessing method is proposed. The limited-handwritten chinese character is produced by pickup camcra. The proposing algorithms for chinese is a method combining regular position, regular big-small, level and smooth processing, slender processing. The recognition rate is raised greatly for chinese character.

Key words pattern; regulation; nearby; connectedness; connective numbers; deletion points

编辑 徐安玉