文章编号:1001 - 9081(2006)03 - 0714 - 03

基于部件复用的分级汉字字库的构想与实现

冯万仁,金连文
(华南理工大学 电子与信息学院,广东 广州 510641)
(wrfeng@ scut. edu. cn)

摘 要:为减少汉字图像字库的存储量,提出了一种基于仿射变换的分级汉字字库构造的新方法。该方法设计了汉字经常使用的基本部件库,通过仿射变换重复使用这些部件来自动生成任意汉字。提出了使用仿射变换来实现部件与汉字之间的转换的方法。试验表明,该方法可行,在减小字库存储量上具有显著的优点。

关键词:分级汉字字库;仿射变换;部件;笔画
中图分类号: TP391.12; TP391.41 **文献标识码**:A

Hierarchical Chinese character database based on radical reuse

FENG Wan-ren, JIN Lian-wen
(*School of Electronic and Information Engineering, South China University of Technology, Guangzhou Guangdong* 510641, *China*)

Abstract: In order to reduce the storage of Chinese character bitmap database, a new hierarchical Chinese character database design method based on affine transform was proposed. The basic radicals library which were frequently used was designed. Each Chinese character could be generated from the basic radicals library. Affine transform was brought forward to generate characters from radicals. Experiments show that the proposed method works very well. The storage of bitmap Chinese database can be decreased greatly by using this new method.

Key words: hierarchical Chinese character database; affine transform; radical; stroke

0 引言

汉字作为历史最悠久的一种文字,拥有世界上最多的使用人群。但汉字在信息时代也暴露出词汇量多、存储量大的弱点。这些弱点给汉字图像字库在信息时代各种产品中的使用增加了困难。

分级字库就是一种减小汉字图像存储量的办法。有人提出使用分级的模板来生成手写体汉字,将一个汉字分割成几块,再根据块间的比例将部件拼接起来[1]。有人提出根据汉字部件之间的结构将各种部件拼接起来,调整部件的比例结构从而达到较好的效果[2,3]。这些方法都使用了部件的概念,但生成汉字时使用的拼接方法会使生成的汉字比较呆板。在通过部件生成汉字的时候,不仅需要将部件平移、缩放,还要形变。本文提出使用仿射变换就可以很好地解决这个问题。

本文提出了这样一种分级字库的构想。我们知道,一个汉字由几个部件和笔画构成,这些部件包括偏旁部首、字元还有基本笔画等。许多的部件和笔画在不同的字中经常出现,所以,字库中只存储常用的部件和笔画,而每一个汉字都通过这些部件和笔画来生成,这样就可以达到减少存储量的目的。

本文所介绍的分级字库的第一部分是总结出来的常用部件和基本笔画,第二部分是每个汉字使用的部件在部件库中的索引和信息。这样,汉字图像就可以通过索引和信息在部件库中提取的部件和笔画来构成,而不再需要保存每一个汉字的图像。这可以大幅度减小字库的存储量。但是简单的部件拼接无法得到较好的字形生成效果,所以本文利用仿射变换来实现部件和笔画生成汉字。

1 分级汉字字库的原理

汉字有许多是形声字,它们的形旁就代表它们的特点,例如:“林”、“枝”、“杯”、“柜”、“枇”、“枣”、“果”,这些汉字含有“木”字旁,就代表它们与树木有关联。“江”、“河”、“湖”、“海”,都含有“水”旁,这些就都与水有关。除此之外还有许多含有相同字元的汉字,例如:“投”、“殴”、“殴”、“股”、“段”、“殷”。虽然它们所含有的相同部分不一定代表什么含义,但这些在字库里面都是可以共同使用的东西。根据汉字的这个特点,我们可以将汉字中重复使用的部件和一些基本笔画集合起来,成为一个基本库,这个库包含的部首和字元可以涵盖整个国标码 GB18030-2000 中的约 27 000 个字,这样全部的汉字都可以由这个库生成。

虽然每个汉字都由字库中的部件生成,但许多汉字相同的部件却不在同一个位置,大小也不一样,甚至形状也不同。而对于同一个部件,在字库中的位置大小形状都是固定的。要重复利用部件和笔画,利用简单的部件拼接是行不通的。本文用仿射变换来使得一个部件可以表示出不同的位置、大小和形状,从而达到部件复用的目的。

1.1 汉字的数学描述

一个汉字 CC 由若干个部件 C_i, $i = 1, 2, \cdots, T$,构成,一个部件 C_i 由若干笔画 S_i, $i = 1, 2, \cdots, n$ 构成,每个笔画 S_i 由若干个黑像素 $\{(x_i, y_i), i = 1, 2, \cdots, m\}$ 构成,即:

$$CC = \{C_1, C_2, \cdots, C_T\} \quad (1)$$

$$C_i = \{S_1, S_2, \cdots, S_n\} \quad (2)$$

$$S_i = \{(x_1, y_1)^T, (x_2, y_2)^T, \cdots, (x_m, y_m)^T\} \quad (3)$$

收稿日期:2005 - 09 - 09 **修订日期**:2005 - 11 - 24
基金项目:国家自然科学基金资助项目(60275005);广东省科技计划项目(2003C50101, 04105938)
作者简介:冯万仁(1981 -),男,广东茂名人,硕士研究生,主要研究方向:图像处理、模式识别; 金连文(1968 -),男,贵州都匀人,教授,博士生导师,主要研究方向:图像处理、模式识别.

1.2　基于仿射变换的部件复用

(a) 部件原图　　(b) 仿射变换后的效果

图 1　仿射变换的例子“城”,“费”,“处”,“意”

仿射变换[4]是在平面上的一种线性变换,可以将图像平移、缩放、旋转和倒映。所以仿射变换可以达到部件复用需要的功能。式(4)是仿射变换的一般形式:

$$\begin{aligned} X &= a_1x + b_1y + c_1 \\ Y &= a_2x + b_2y + c_2 \end{aligned} \tag{4}$$

在对部件进行仿射变换生成汉字时,部件上每一个点(x, y)就会通过式(4),生成汉字上的一个点(X, Y),因此我们需要确定$[a_1 \quad b_1 \quad c_1][a_2 \quad b_2 \quad c_2]$这 6 个参数。

首先分别将部件和现有汉字的图像细化,在部件和汉字的骨架上各选取 3 个点分别是(x_1, y_1),(x_2, y_2),(x_3, y_3)和(X_1, Y_1),(X_2, Y_2),(X_3, Y_3),将这些点的坐标代入下面的方程组:

$$\begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$
$$\begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix} \cdot \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \tag{5}$$

解式(5)的方程组可以得到:

$$\begin{aligned} a_1 &= \frac{(X_1 - X_2)(y_1 - y_3) - (X_1 - X_3)(y_1 - y_2)}{(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)} \\ b_1 &= \frac{(X_1 - X_2) - a_1(x_1 - x_2)}{y_1 - y_2} \\ c_1 &= X_1 - a_1x_1 - b_1y_1 \\ a_2 &= \frac{(Y_1 - Y_2)(y_1 - y_3) - (Y_1 - Y_3)(y_1 - y_2)}{(x_1 - x_2)(y_1 - y_3) - (x_1 - x_3)(y_1 - y_2)} \\ b_2 &= \frac{(Y_1 - Y_2) - a_2(x_1 - x_2)}{y_1 - y_2} \\ c_2 &= Y_1 - a_2x_1 - b_2y_1 \end{aligned} \tag{6}$$

最后,根据式(4)和计算所得的 6 个参数,目标图像上的每一对(X,Y)都可以在部件图像上找到一对(x,y)与之对应。根据部件图像上(x,y)坐标处像素点灰度就决定了生成的目标图像(X,Y)处的灰度。

这样的方法,就使得同一个部件可以改变位置、大小和形状,变换出一个与不同汉字相应部位都可以相似的部件。从而实现了使用几百个基本的部件可以生成所有汉字的目的。如图 1 所示。

1.3　仿射变换的选点方法

由图 1 和仿射变换的式(4)~(6)可以看出,仿射变换的关键在汉字和部件上选取的 3 个对应点。这 3 个点在变换后是与原汉字图像位置不变的,从而它们完全决定了变换后的汉字部件的位置、大小和形状。选点不当会造成生成的汉字与需求差别非常大,使得字库的使用不理想。如图 2 所示。

(a) 部件原图　　(b) 选点不当的结果　　(c) 选点正确的结果

图 2　选点不当的仿射变换

针对这个情况,我们做了 500 个仿射变换模拟汉字的试验。我们的仿射变换是先在图像的骨架上选取 3 点来计算参数的,主要考虑的特征点有端点、分叉点等。根据实验,总结出一些选点规律如下:

1)“竖”、“勾”、“撇”等长笔画:可选一端点的两个分叉点和另一端点,两分叉点决定笔画的宽度,另一端点决定笔画的长度。如图 3(a)。

2)部件旋转时,可选两头的端点和中间一点。如图 3(b)。

3)方形的部件:一般选四个角中的三点。需要上部对齐则选上部左右两角和下部一点,需要下部对齐则相反。如图 3(c)。

4)三角形的部件:一般选最上点和最下两点。如图 3(d)。但如果部件没有对正,则可适当在部件或汉字上向左或右偏移下面的某一点,直至对齐。

5)在左边的偏旁:可选最上一点、最下一点和最左一点,在右边的偏旁则反之。如图 3(e)。

(a)“撇”在“税”中的变换　(b)“勾”在“仿”中的变换　(c)“口”在“和”中的变换　(d)“上”在“让”中的变换　(e)单人旁在“仪”中的变换

图 3　选点方法

2　基本部件库构造需要注意的几个问题

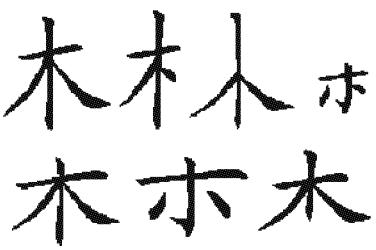


图 4　部件库中的 7 种“木”

虽然有很多汉字使用相同的部件,但实际上这些部件是有区别的,从而使得它们不可以公用同一个部件。例如:林、李、果,这三个字都含有“木”这个部件,而且“林”字包含两个“木”,但我们可以观察到,作为偏旁部首的“木”和作为字元的“木”是不同的,而李、果中的“木”就更加不一样。基于这个原因,在我们设计的部件基本库中,需要包含了 7 种不同的“木”,如图 4 所示。

在部件库中,既可作为偏旁,也可作为字元的部件一般都具有这种情况,例如“大”、“米”、“牛”等。

部件中还有一个特点就是,有些部件可以相互涵盖,例如:“了”和“子”,其中“子”可以用“了”和“一”来生成,但因为“子”的使用面很广,许多字都要用到,都用“了”和“一”就会给下面的

工作带来很多麻烦。而在“辽”等字中也确实具有“了”这个字元,所以基本部件库就同时包含“了”和“子”。这虽然增加了一些存储量,但也给汉字的生成提供了更多的选择。

3 实验

部件库中每个图像用 Bezier 轮廓曲线[5]保存,现在已经总结出 330 个常用基本部件,存储量为 242KB。

在这些部件的基础上,用一级汉字字库前 500 个汉字做了试验。在存储汉字时,每个部件名(即索引信息)有 3 个字节,每个参数是包括小数点在内的 10 位浮点数,所以汉字中每个部件的总储存量为 63 字节。分级字库的存储量与 Bezier 曲线字库的比较如表 1 所示。

表 1 分级字库与 Bezier 曲线字库存储量的比较

汉字个数	汉字 + 部件(分级字库)		Bezier 曲线存储每个汉字	
	总存储量	平均每个汉字存储量	总存储量	平均每个汉字存储量
500	335 KB	680 B	395 KB	809 B
3755	968 KB	264 B	2968 KB	809 B

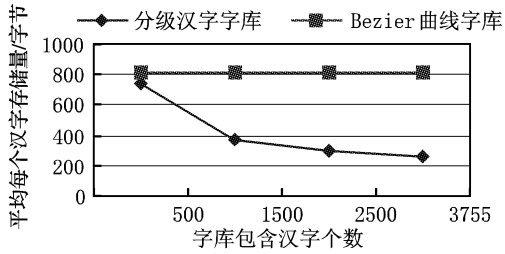
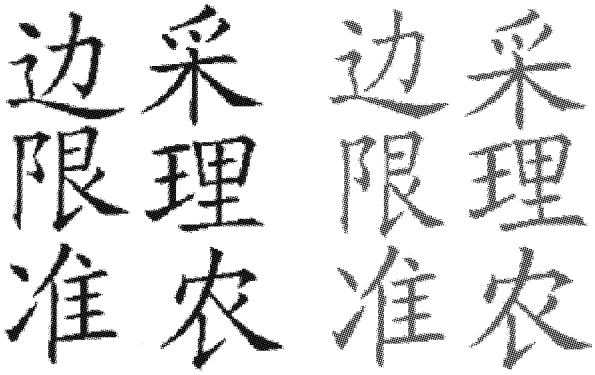


图 5 平均存储量随汉字容量变化图



(a) “楷体_GB2312”的汉字 (b) 生成的汉字

图 6 部件生成的汉字效果图

由表 1 可知,与使用 Bezier 曲线所保存的汉字相比,在存储 500 个汉字的时候,分级汉字字库每个汉字的存储量少了 16%。而随着生成的汉字增加到一级字库的 3755 个,平均每个汉字的存储量显著下降了 67.3%,因此,本文方法在大字库时存储量优势十分明显。

部分生成的汉字效果如图6所示,图6(a)为通过存储的部件索引和参数信息调用部件生成的汉字,图 6(b)为原“楷体_GB2312”的汉字。由图可见,构造出的汉字与真实汉字图像形状非常一致。

4 结语

本文提出了基于仿射变换重复利用部件来生成每一个汉字的新方法,总结了常用的部件,大大减小了汉字字库的存储量。实验结果表明,通过部件生成的汉字效果不错,与原来的楷体相比,在笔画和结构上都非常相近。存储量与 Bezier 曲线字库相比有明显的下降。

由于存储量的大大减少,分级汉字字库的方法有着广泛的应用,如:动态汉字字库[6]装入 PDA、嵌入式系统或单片机中,基于分级汉字字库就可以大大减少存储量需要的成本。

由于手工选点做仿射变换将会有很大的工作量。我们构想了自动对汉字和部件分别选点来进行仿射变换的一套算法。过程步骤大致如下:

1) 分别对汉字和部件提取骨架;

2) 在骨架上提取特殊点(包括:端点、分叉点和拐点),设汉字上提取的特殊点数量为 m,部件上为 n 个;

3) 分别在 m,n 个点中随机取出 3 点,即有 $P_m^3 \times P_n^3$ 种组合;

4) 对每一个组合做仿射变换,即得出 $P_m^3 \times P_n^3$ 组的参数;

5) 每一组参数根据部件仿射模拟出一个结果,分别与汉字原图像进行比较求欧氏距离[7],得出结果最小,也就是效果最好的结果所对应的一组参数即为所求。

虽然这种方法在运算时间等方面还有不少改进的地方,但通过这样的方法,可以在建立分级汉字字库的时候大大减少人工的工作量,使得分级字库的汉字容量能够更容易地扩展,在实用性方面有很大的提高。

参考文献:

[1] LEE H - J, HSU H - C. A hierarchical model - guided generation of Chinese characters[R]. National Chiao Tung University, Hsinchu, Taiwan, 1994.

[2] LAI P-K, YEUNG D-Y, PONG M-C. A Heuristic Search Approach to Chinese Glyph Generation Using Hierarchical Character Composition[J]. Computer Processing of Oriental Languages, 1996, 10(3).

[3] WANG J-H, OZAWA S. Automated Generation of Chinese Character Structure Data Based on Extracting the Strokes[Z]. Department of Electrical Engineering, Keio University, 1993.

[4] 魏海涛. 计算机图形学[M]. 北京: 电子工业出版社, 2003. 51 - 68.

[5] 马小虎, 潘志庚. 高质量 Beizer 曲线描述轮廓库自动生成算法[J]. 自动化学报, 1994, 20(1): 121 - 125.

[6] 陈东明, 金连文. 基于骨架自动跟踪的动态汉字字库的设计与实现[A], 彭群生. 中国计算机图形学进展 2004——第五届中国计算机图形学大会论文集[C]. 西安: 西北工业大学出版社, 2004. 457 - 460.

[7] DUDA RO, HART PE, STORK DG. Pattern Classification[M]. Second Edition. 北京: 机械工业出版社, 2003.

(上接第 713 页)

参考文献:

[1] LIANG G-H, TJAHJADI T, YANG Y-H. Roof Edge Detection Using Regularized Cubic B-Spline Fitting[J]. Pattern Recognition, 1997, 30(5): 719 - 728.

[2] ZIOU D. Line Detection Using An Optimal IIR Filter[J]. Pattern Recognition, 1991, 24(6): 465 - 478.

[3] 张小莉, 王敏, 黄心汉. 一种有效的基于机遇 Freeman 链码的拐角检测法[J]. 电子测量与仪器学报, 1999, 13(2): 14 - 17.

[4] 李翊华, 胡匡祜, 苏万芳. 一种提取微血管边缘曲线角点策略[J]. 中国生物医学工程学报, 1998, 17(4): 289 - 194.

[5] LIU Y, HUANG TS. Determining Straight Line Correspondences From Intensity Images[J]. Pattern Recognition, 1991, 24(6): 489 - 504.

[6] LEE J-W, KWEON I-S. Extraction of Line Feature in A Noisy Image[J]. Pattern Recognition, 1997, 30(10): 1651 - 1660.

[7] 温熙森, 胡芶庆, 邱静. 模式识别与状态监控[M]. 长沙: 国防科技大学出版社, 1997.

[8] 沈清, 汤霖. 模式识别导论[M]. 长沙: 国防科技大学出版社, 1991.