基于统计的笔段网格字形笔画曲线美化方法

张麦库,林 民,黄含泉
(内蒙古师范大学 计算机与信息工程学院,内蒙古 呼和浩特 010022)

摘 要:已设计的汉字笔段网格字形笔画曲线美化算法对笔画曲线美化处理比较符合人对汉字的认知习惯,需要用户通过相应笔画的调节点对曲线笔画做局部调节从而达到满意的效果,对每个网格字形进行美化处理都要经过上述的处理,当处理大量字形时比较繁琐费时;针对该问题文中抽取有代表性的有限数量的网格字形作为训练集,在此方法的基础上,对训练集中的字形做美化调节处理并记录不同曲线笔画类型中调节点的位置信息,然后用统计方法算出不同曲线笔画类型中相应调节点的位置特征信息,用这些调节点位置信息结合上述方法对网格字形进行自动美化处理,对处理大批量网格字形效率高、自动处理能力强且效果好。

关键词:调节点;训练集;统计方法
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2012)10-0083-04

Stroke Curve Beautification Algorithm of Stroke-segment-mesh Glyph Based on Statistics

ZHANG Mai-ku,LIN Min,HUANG Han-quan
(Computer & Information Engineering College,Inner Mongolia Normal University,Hohhot 010022,China)

Abstract:The beautification of Chinese character stroke-segment-mesh glyph stroke curve in processing stroke curve beautification is more in line with the cognitive habits of the the Chinese character,in order to achieve satisfactory glyph results,user needs to adjust the adjust point of the corresponding strokes do local adjust on strokes,for each grid glyph beautify to go through the above processing,more cumbersome and time-consuming when dealing with a large number of glyphs. For the problem,extract a representative of a limited number of grid-glyph as a training set,on the basis of the method,do beautify adjust processing to the training set and record the location information of adjust point in different types of curve strokes,and then use statistical methods to calculate the adjust point location characteristics of the different of curve strokes,using the adjust point location information combined with the above method auomatic beautification of grid-glyph,it is high efficiency and high automatic processing ability and better result in processing large quantities of grid-glyph.

Key words:adjust point;training set;statistical method

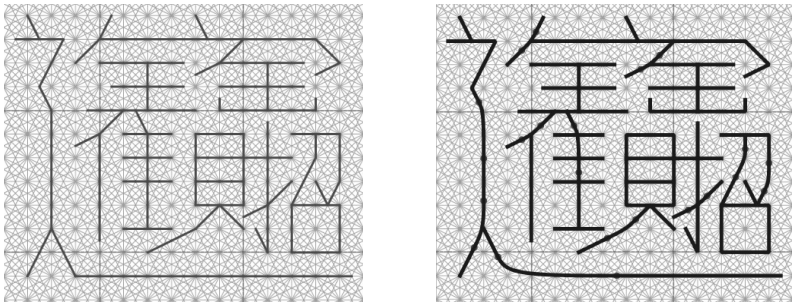
0 引 言

一种笔段网格汉字字形描述方法[1]是从字形的角度描述汉字,汲取了表意文字描述序列[2]、字符描述语言[3],采用颗粒度适当、规范化、无歧义的笔段基元描述各种可能的字形(包括正字、错字、古籍异体字以及拼合字)骨架异同,该方法简化了字形的描述方法,从而高效支持字形的整体与局部自动比对计算。由于该方法只能用直线段描述字形,从而使描述后的字形丧失了字形的曲线特征,在需要曲线描述的笔画中两折线段处出现了多余的尖角,影响了字形的显示效果。基于该问题设计了汉字笔段网格字形笔画曲线美化算法工具,通过人工交互能较好地解决这个方面的问题。

该方法对网格模型中的字形提取各笔画,结合人对汉字的认知自动判断字形中需要曲线化描述的笔画,然后在需曲线化描述的两直线段在折点两边的折线段上各增加相应的调节点并做相应的标记信息,同时提取笔画中起点、调节点、折点、尾端点等关键点相关信息,结合已设计的汉字笔段网格字形笔画曲线美化算法对字形描绘,增加的调节点用户可以自由移动从而调节笔画的曲线效果,从而使网格字形显示效果更美观。图 1 为拼合字“招财进宝”笔段网格字形经过该方法的处理过程及显示效果。

在图 1(c)过程中对曲线笔画的调节比较繁琐费时, 调节后的曲线笔画类型当在别的网格字形中出现

收稿日期:2012-02-15;**修回日期**:2012-05-21
基金项目:国家自然科学基金资助项目(60863007)
作者简介:张麦库(1987-),男,山东菏泽人,硕士研究生,主要从事自然语言处理、汉字信息处理;林 民,博士,教授,研究方向为自然语言处理、人工智能。



a 拼合字“招财进宝“的笔段网格字形描述　b 拼合字“招财进宝”初次美化处理后的效果

c 拼合字“招财进宝”交互调节后的效果　d 拼合字“招财进宝”最终美化效果

图 1　网格字形笔画曲线美化处理过程及显示效果

时仍需要做上述相同的复杂处理。根据傅永和对 11834 个汉字的笔画的统计表明,频率最高的汉字笔画是“横”(包括“提”),占 30.30%;其次是“竖”,占 19.38%;再次是“折”,占 17.95%;再次是“撇”,占 15.74%;最少的是“捺”,它和“点”合在一起占 16.64%[4]。结合笔段网格模型特征对网格字形中的笔画分类,在图 1(c)对网格字形曲线笔画调节的过程中,记录属于不同曲线笔画类型的各笔画调节点的位置特征信息,结合统计方法计算出在不同曲线笔画类型中调节点位置特征信息,然后用这些信息结合汉字笔段网格字形笔画曲线美化算法处理网格字形,能较好地解决大批量的网格字形笔画曲线美化处理。

1　笔段网格字形曲线笔画分类及命名规则

根据汉字字形形式化描述方法及应用研究[5]对网格笔画分为基本笔画、简单笔画、复合笔画,简单笔画是由基本笔画组合而成,简单笔画分为简单笔画 Heng、Shu、Pie、Na,编码分别为 0、1、2、3;简单笔画 Pie 又细分为平撇、斜撇、立撇三种子类型,编码分别为 0、1、2;简单笔画 Na 又细分为平捺、斜捺、竖捺三种子类型,编码分别为 0、1、2,见表 1。复合笔画由简单笔画相互组合而成,各笔画的具体定义见文献[5]。

同曲线笔画类型定义:尾首顺次相连组成笔画的各直线段所属简单笔画类型及子类型都相同的不同形状的笔画都属于同种曲线笔画类型;同类型的笔画由于组成笔画的直线段长度长短不同显示效果也不同,为了详细地区分各种笔画采用如下笔画标识规则:曲线笔画类型名由顺次组成笔画的各直线段所属简单笔画类型编码及子类型编码组合而成,标识为 C_iS_i(分别表示为笔画中第 i 个直线段所属的简单笔画的编码及子编码);同曲线笔画类型的笔画名称由直线段所属简单笔画类型的编码及其在水平、竖直方向上长度占半个基准小矩形长度的个数组合而成,标识为 $C_iL_xL_y$(分别为第 i 个直线段所属笔画的类型编码及在水平、竖直方向占半个基准小矩形的个数);如笔画类型为 3130 对应的笔画名称及图例,见表 2。

表 1　简单笔画类型定义及图例

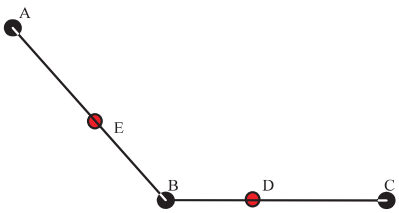
笔画类型	笔画含义	类型编码		简单笔画类型图例		
Heng	横	0				
Shu	竖	1				
Pie	平撇	2	0			
	斜撇		1			
	立撇		2			
Na	平捺	3	0			
	斜捺		1			
	立捺		2			

表 2　笔画类型 3130 对应的笔画名称及图例

344,363	322,342	344,321	366,321

2　同种曲线笔画类型调节点获取方法

在网格字形中需要曲线化描述的笔画,其起点、折点、尾端点已确定,笔画曲线的形状主要受调节点位置的控制,如图 2(a)调节点 E、D 所在的位置经曲线化处理后的效果如图 2(b)所示。根据人对汉字字形的认知,对网格字中需要曲线化处理的笔画对其调节点位置做相应的调节,使笔画曲线达到更美观,同时记录曲线笔画类型、笔画名标识、调节点位置值及对应出现的次数、每一笔画出现的次数,分类统计出同种曲线笔画类型各调节点位置信息,然后用这些信息作为网格字形中需曲线化处理的同类型笔画中调节点的位置信息,然后用三次 B 样条曲线[6]对其描绘。



a 需曲线化的两折线段及关键点

b 曲线化处理后的笔画效果

图 2　需曲线化处理的两折线段调节点的位置对笔画曲线效果的影响

2.1　网格字形训练集提取

采用笔段网格字形描述方法[7,8]开发的工具已完成了 GBK 字符集中全部 20902 个汉字,GB2313 字符集中 6763 个简体字;根据文献[5]对网格字形笔画的分类,从中选取需要曲线化处理的笔画类型,再从 6763 个简体字中抽取 400 个网格字,其中包含了选取的笔画类型,作为已开发的汉字笔段网格字形笔画曲线美化算法美化工具的训练字形。

2.2　曲线笔画类型调节点位置设计

在网格字形中需要曲线化处理的笔画其起点、折点、尾端点已确定,笔画的曲线形状主要受可以动态移动调节点位置的影响,如图 2(a)起点 A、折点 B、尾端点 C 位置已确定,图 2(b)曲线形状主要受调节点 E、C 的位置影响;在属于 Pie、Na 类型的直线段,调节点位置取值范围在以直线段为对角线的矩形框内任意位置值,如图 2(a)所示,调节点 E 在以直线段 AB 为对角线的矩形范围内;在属于 Heng 类型的直线段调节点只能在其水平方向上调节移动;在属于 Shu 类型的直线段调节点只能在其竖直方向调节移动。为简化计算,调节点位置取值用调节点到折点的距离与其对应直线段长度的比值表示,取值分别为 0、0.05、0.1、0.15…0.95、1,共 21 个取值,每个取值相差 0.05,当取值为 0、1 时表示调节点与端点重合,为简化存储处理对调节点位置进行如下规整处理:$V = 2 \times V1 \times 10$($V1$ 为实数值,V 为处理后的值),经处理后调节点位置取值分别为 0、1、2、3…19、20。

调节点位置值具体表示用调节点与折点距离在水平方向、竖直方向长度分别与其对应的直线段在水平、竖直方向长度的比值表示,分别表示调节点距折点在水平、竖直方向的值,如图 2(a)中,调节点 E 距折点 B 的位置为 EB 在水平方向、竖直方向长度分别与 AB 在水平方向、竖直方向长度比值;对 400 个网格训练字形进行人工笔画曲线美化调节处理,记录曲线笔画类型、笔画名称、各笔画出现的数、调节点取值及对应出现的次数,如对笔画类型为 3130 各笔画对应调节点位置等信息如表 3 所示(此笔画类型共两个调节点,调节点比例取值[0,20],第一行为两调节点各自的取值范围,第一列为不同笔画标识,笔画标识行对应的取值为调节点在不同位置时出现的次数):

表 3　曲线笔画类型为 3130 各笔画对应调节点位置及出现的次数记录信息

	10	11	12	13	14	15	16	17	18	0	…	13	14	15	16	17	18	19
388384			1												1			
344384					3										1	1	1	
322342						1									1			
3443105						1						1						
366363							1										1	
322384				1			1	1					1		1		1	
366342				1												1		
344342						2		2							2	2		
366384			1												1			
31010321		2	1			1							1	1	1	1	1	
344363		2	1					1	1					2	1	1	1	
366321	2	3	2	1									1	3	1	1	1	2
344321			1										1					

经过对 400 个网格字形的笔画曲线美化调节处理,获取 442 个需要曲线化描绘不同笔画,分为 72 种笔画类型,通过对同种曲线笔画类型的各笔画调节点位置信息分析,得出不同笔画标识的各调节点位置都比较集中在相应的某个值附近,基于此采用统计方法的思想分析同种笔画类型不同笔画各调节点位置的共性,最后算出同种笔画类型各调节点的位置信息,用这些信息作为网格字形中同笔画类型的不同笔画对应各调节点的位置信息,这样描绘后的网格字形与人工笔画曲线美化调节后的网格字形有极大的相似性。

2.3 获取同种曲线笔画类型调节点位置

由于人对汉字字形审美有稍微的差别,在对笔段网格字形曲线美化时,不同的人对同一个字形笔画曲线美化调节处理获取的同一笔画调节点位置信息不完全相同,即使同一个人对同一个网格字形在不同的时间内对其曲线美化调节获取的同一笔画调节点位置也不完全相同,但调节点大概在某个值附近内排布,如表 3 所示,为此从属于同种曲线笔画类型的众多笔画中,找出笔画中各调节点位置的一个合理值,笔画中相应调节点位置值都在该值附近均匀分布,把该值作为同曲线笔画类型相应调节点位置值。

具体方法如下:

(1)对网格字形训练集进行训练学习,并对需要曲线化描述的笔画记录其曲线笔画类型、笔画名称、笔画中各调节点位置取值及各调节点在取不同值时出现的次数、同笔画名称出现的总次数。

(2)对曲线笔画类型分类,属于同种曲线笔画类型的笔画放在同一组内;对笔画中各调节点按照其出现次数的大小对其相应出现的位置从大到小排序,根据经验对笔画中各调节点取前四个值的信息。

(3)统计出属于同种曲线笔画类型的笔画出现的总次数记为 C,第 i 个笔画出现的次数记为 C_i,笔画中调节点个数记为 N,第 i 个笔画中第 j 个调节点取第 $k(k=1\cdots4)$ 个位置值记为 P_{ijk},其出现次数记为 C_{ijk},用如下方式求出同种曲线笔画类型各调节点位置值:

For ($j=1;j<=N;j++$)

{

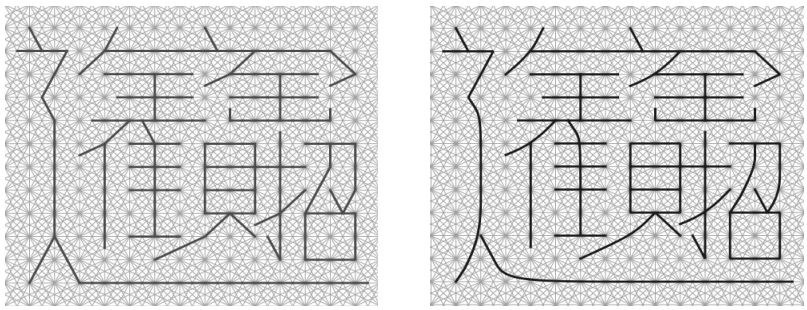
$$P_j=\sum_{i=1}^{C}\left(\frac{C_i}{C}\bullet\sum_{k=1}^{4}\left(\frac{C_{ijk}}{C_i}\bullet P_{ijk}\right)\right);$$

}

通过同样的方法求出不同种曲线笔画类型各调节点位置信息。

3 网格字形绘制效果对比试验

用这些已知曲线笔画类型调节点位置信息对网格字形中需要曲线化描绘的同种笔画类型的笔画调节点做替换,然后用已设计的汉字笔段网格字形笔画曲线美化算法[9]对网格字形进行描绘。用此方法对拼合字“招财进宝”的网格字形处理前后效果如图 3 所示。



a 拼合字“招财进宝“的笔段网格字形描述 b 拼合字“招财进宝”美化处理后的效果

图 3 拼合字“招财进宝”的网格字形处理过程及显示效果

按照汉字信息字典[10,11]中结构类型的独体字、上下结构、左右结构、包容结构、嵌套结构对笔段网格字库中的简体、繁体各取 100 个已选取训练集外的网格字形,用此方法对其曲线美化处理,都达到满意的效果;对留学生错字字库的 200 个汉字字形也做了测试,绝大部分字形效果较好,少部分字形中需要人工调节,主要原因是错字字形中笔画随意,导致找不到与训练集中相同的曲线笔画类型,所以用此方法无法处理。

4 结束语

从网格字形库中合理抽取一定数量的字形作为训练集,在已设计好的汉字笔段网格字形笔画曲线美化算法工具上按照人对汉字字形的认知对字形训练集进行笔画曲线美化调节处理,同时记录曲线笔画类型、笔画名标识、调节点位置及对应值出现的次数、每个笔画出现的次数。从属于同笔画类型的众多笔画中统计算出曲线笔画类型调节点的特征位置信息,然后再用这些信息对网格字形中属于同类型的笔画各调节点做替换处理,最后再对网格字形进行美化处理绘制。该方法对处理大批量网格字形笔画曲线美化处理效率高、自动处理能力强且效果好,若对笔画曲线有更细致的要求用户仍可以对笔画曲线做局部的调整。

用该方法对网格字形笔画曲线美化处理过程中,需曲线化处理的笔画所属的曲线笔画类型能在训练集库中找到的笔画能较好地处理,找不到的曲线笔画类型还不能很好地处理,处理不好的笔画主要是一些错字字形中的奇怪的组合笔画,由于其随意性大造成笔画形状奇特。

(下转第 90 页)

3.5 与最简规则比较分析

表 2 的最简规则(表 3 左)和其决策树(见 3.4)展开规则(表 3 右)相比较,两者都只涉及{a,d}属性;展开规则个数多 1 个。若排除树规则有重值的因素,决策树展开规则与最简规则基本接近,可见决策树是较优的决策树。

表 3 最简规则与树展开规则比较

d. 1->e. 3	d. 1->e. 3
d. 2∧a. 2->e. 2	d. 2∧a. 1->e. 1
d. 2∧a. 3->e. 2	d. 2∧a. 2->e. 2
	d. 2∧a. 3->e. 2
d. 3∧a. 2->e. 2	d. 3∧a. 1->e. 1
a. 1->e. 1	d. 3∧a. 2->e. 2

4 结束语

(1)粗糙集是以 $\{An\}$ 与 D 划分关系为出发点,而本算法是以 $Ai \in \{An\}$ D 划分关系为出发点,首要解决的是在 $\{An\}$ 中确定分裂属性而并非确定核属性。

(2)将 Ai 按属性值划分为若干粒,提出以粒为基本单位确定分裂属性且直接建树,不涉及等价类和属性约简等复杂运算的中间过程。

(3)算法具有可扩展性,当属性划分力接近时,可采用扩级方法进一步确定分裂属性。

参考文献:

[1] 杨 静. 决策树算法的研究与应用[J]. 计算机技术与发展,2010,20(2):114-120.

[2] 丁春荣. 基于粗糙集的决策树构造算法[J]. 计算机工程,2010,36(11):75-77.

[3] Wang X J, Reyes J L, Chua Nam-Hai, et al. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNAtargets[J]. Genome Biol. ,2004(5):1-15.

[4] 覃伟荣. 基于粗糙集理论的条件属性动态约简算法[J]. 计算机技术与发展,2008,18(8):23-25.

[5] 李永华. 一种基于 rough 集的属性约简的改进算法[J]. 计算机应用,2008,28(8):200-202.

[6] 李 鸿. 知识粗糙性与知识粒度的关系研究[J]. 计算机技术与发展,2007,17(8):117-119.

[7] Kim Sung-Kyu, Nam Jin-Wu, Rhee Je-Keun, et al. miTarget: microRNA target gene prediction using a support vector machine[J]. Bioinformatics,2006(7):411-422.

[8] 葛 浩. 一种改进的基于二进制可分辨矩阵属性约简算法[J]. 计算机技术与发展,2008,18(8):12-15.

[9] 王国胤,张清华,胡 军. 粒计算研究综述[J]. 智能系统学报,2007,2(6):8-26.

[10] Chou Chi-Hung, Yang Tsung-Hsien, Tsao Shih-Chiang, et al. Standard operating procedures for embedded linux systems [J]. Linux Journal,2007(160):10-14.

[11] 周 军. 基于粒度商的决策树构造算法[J]. 计算机工程与设计,2009,30(16):3826-3829.

[12] 高 静. 一种改进的基于正区域的决策树算法[J]. 计算机科学,2008,35(5):138-142.

[13] 陈 婷. 基于粒度的粗集-决策树雷达信号识别模型[J]. 微电子学与计算机,2008,25(12):13-16.

(上接第 86 页)

下一步主要工作是对任何笔画都能较好地处理,主要思想:

(1)不能处理的复杂笔画都是由简单曲线笔画类型或组合曲线笔画类型拼合而成,因此对于不能处理的复杂曲线笔画类型把它分解成两个或更多的训练集库中简单曲线笔画类型或组合曲线笔画类型来分段对其处理。

(2)对一些奇怪的曲线笔画类型单独做成美化曲线笔画类型库,当遇到不能处理的笔画优先去美化曲线笔画类型库中查找处理。

参考文献:

[1] 林 民,宋 柔. 一种笔段网格汉字字形描述方法[J]. 计算机研究与发展,2010,47(2):318-327.

[2] Ideographic Description[EB/OL]. 2003. http://www.unicode.org/versions/Unicode4.0.0/ch11.pdf.

[3] Cook R. A Specification for CDL(Character Description Language):an extract of[D]. USA:UC Berkeley, Dept. of Linguistics,2003.

[4] 傅用和. 中文信息处理[M]. 广州:广东教育出版社,1999.

[5] 林 民,宋 柔. 汉字字形形式化描述方法及应用研究[D]. 北京:北京工业大学,2009.

[6] 罗笑南,王岩梅. 计算机图形学[M]. 广州:中山大学出版社,2003:188-196.

[7] 林 民,宋 柔. 汉字的笔段网格字形描述及字形比对计算[J]. 计算机辅助设计与图形学学报,2009,21(9):1298-1307.

[8] 林 民,宋 柔. 一种面向构形计算的汉字字形形似化描述方法[J]. 中文信息学报,2008,22(3):115-123.

[9] Zhang Maiku, Lin Min, Huang Hanquan. Beautification of Chinese Character Stroke-Segment-Mesh Glyph Stroke Curve [J]. Advances in MSEC,2012,2:101-110.

[10] 上海交通大学汉字编码组. 汉字信息字典[M]. 北京:科学出版社,1988.

[11] 国家语言文字工作委员会. GF3001-1997 信息处理用 GB13000.1 字符集汉字部件规范[S]. 北京:语文出版社,1997.

基于统计的笔段网格字形笔画曲线美化方法

作者： 张麦库， 林民， 黄含泉
作者单位： 内蒙古师范大学 计算机与信息工程学院,内蒙古 呼和浩特 010022
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(10)