

多重复字字库的制作与使用

申请号：[201610403423.7](#)

申请日：2016-06-12

申请(专利权)人 [朱信](#)

地址 100037 北京市西城区阜外大街南礼士路甲一号西八楼201号

发明(设计)人 [朱信](#)

主分类号 [G06F17/22\(2006.01\)I](#)

分类号 [G06F17/22\(2006.01\)I](#)

公开(公告)号 106021204A

公开(公告)日 2016-10-12

专利代理机构

代理人



(12)发明专利申请

(10)申请公布号 CN 106021204 A

(43)申请公布日 2016.10.12

(21)申请号 201610403423.7

(22)申请日 2016.06.12

(71)申请人 朱信

地址 100037 北京市西城区阜外大街南礼
士路甲一号西八楼201号

(72)发明人 朱信

(51)Int.Cl.

G06F 17/22(2006.01)

权利要求书1页 说明书3页

(54)发明名称

多重复字字库的制作与使用

(57)摘要

扫描古籍,将其字做成字库,再用该字库排印该古籍,是之谓“古籍字库级翻印”,可兼具影印的真实、免校和排印内容可编辑等优点于一身。但通用型字库的编码系统不能容纳同一字符因书法行气变化产生的多个样式(重复字),而保真的翻印古籍恰恰要求这个。本技术通过对这种字库自编码,以容纳多重复字;再根据该古籍的文本与自编码字库的编码顺序做一耦合文档,使自编码字库与古籍文本关联,以便编辑。采用GB18030-2005《中文编码字符集》,可对161万字以内的古籍达成字库级翻印。

1.在通用字库ttf(即TrueType,包括OpenType(otf),下同)中,通过对字符码位的自定义编码,即为加入ttf的字符图像在unicode合理范围内,将同一字符的多个不同艺术样式(非异体字)存储在同一ttf中,制成“多重复字字库”。

2.制作和使用多重复字字库的关键是要对所表达(古籍)的字符图像和txt文本按照“含义顺序对应一致”原则:1,顺序读取一个txt文本中的字;2,取得对应的字符图像;3,重新给定码位并将图像置入ttf字库;4,对该txt文档进行“耦合改造”:令各重复字与新制ttf中与之对应编码字相一致;改造后的文本,对其适用对应的“多重复字字库”后,原含义字符图像不仅原序显示,且表现出在不同上下文同一字的不同变化造型,从而达到对中文古籍刻本的字库级再现;如用通用字库显示,则含义不明。

3.通过采用超大字符集GB18030-2005《信息技术中文编码字符集》,单个ttf字库可容纳1611668个字以内的古籍刻本原著且被通用文字编辑程序支持;如能对重复量超多的常用字进行样式分类整理,可容纳更长文本。

4.采用多重复字字库,可“一书一字库”;多重复字字库可容纳多个重复字,容纳多字体也不成问题,因此可不必再为同一古籍中不同字体另做字库。

5.还可将二值墨线画稿制作为字库的一个“字符”,这对古籍刻本中有插图著作的翻印也有方便存储、编辑、使用、提高印刷质量、降低印刷成本等好处。

多重复字字库的制作与使用

技术领域

[0001] 本发明涉及在同一ttf(truetype简称ttf,包括otf,下同)字库中存储与调用多个重复字(同一字符图像,但各自艺术造型不同)的字库制作与使用方法,适用于中文书法性质古籍刻本的字库级翻印。

背景技术

[0002] 用古籍扫描制作ttf字库,然后用字库翻印古籍的好处是:1,字库制作是对古籍进行了字符级的矢量保真分割;2,字库排印古籍,可用清晰字对讹、错、损等进行较高质量的修补,还方便增改现代句读、版式。传世古籍汗牛充栋,出版、印刷、保存质量多有不佳者,如皆影印传承,质量差强人意。尽管用古籍自身字符制成该古籍字库再翻印,似乎解决了这些问题,但同字符同模样地重复出现,违背古籍刻本的书法原貌,反不如影印传神。刻本质是书法。书法不是活字。书法因行气需要,同一字符针对不同上下文其长、短、大、小、肥、瘦、粗、细、内宫位置、动态样式、字间距等都呼应有变化。现通用ttf必须在N多不同艺术造型中“一字一码”地选用一个,并以此单一字型适应各种上下文,再好,也与古籍原著书法风格不符。活字性质ttf与书法根本区别:ttf中同一字在任何上下文只能同一模样,无变化,而书法却要求根据上下文行气,变化字符造型应对。因此,活字从原理上不能再现书法。古人之所以不爱用活字出书,其“非书法性”是根本原因。现电子时代,用字库排印出版虽然质量、效益最佳,但一般的通用字库,即使用古籍自身字符制成该古籍的ttf翻印该古籍,也从根上破坏了中国传统书法的艺术效果。要想解决ttf的这个问题,必须将其改造为能存储和使用多个重复字。

[0003] 当下要想为一种字体安排多个重复字,似乎只能用类似“汉语拼音字库”一种解决方案:同一字体,copy成5个(音调)分库,需用哪个选哪个。若用此法表达书法,可能比一字一码稍好,但数量必然有限(很难想象为表现书法艺术而做N多分库),使用上也有诸多不便,制作时还会出现各字库间不易对行气、整体协调困难等问题。字库级翻印古籍要求ttf包含多个重复字并存储于同一字库。本技术通过对制作古籍ttf时对重复字的重新编码与对应文本中对应字符的“耦合改造”,变“一字一码”为“一型一码”,从而在兼容现有字符编辑软件工具的前提下,解决了上述矛盾。

[0004] 多重复字字库制作与对应文本“耦合改造”技术实现过程:1,准备好该古籍原文文本;2,根据该文本字序(包括重复字)全部重新编码,给出新码位;3,在新给码位上,按字序置放从该古籍图像上扫描生成的ttf字符图像;4,同时修改该文本文档,令其中该字为对应的新码位字。例如“白云长长长长长长消”,重新编码、将原著的该字符图像置入ttf以及对txt文本进行耦合改造后,该文本如用通普字库显示,为一串数量与原文字数相等但含义不可理解的字符。但在对该文档选用新编码字库后,不仅字符还原显现,且由于是原位置字符,虽是字库,亦呈现原著状态,达到古籍的字库级恢复。简单说,就是根据古籍顺序,重新安置所有字(包括重复字)。此法翻印古籍,原著中字符如不清楚或损泐,可根据上下文及书法行气状态,择其它合适者替代。

[0005] 或许有人会担心这种字库不能表达长文本。unicode给汉字的编码位置,加上扩展A,只有27484个码位,照此,最长适用文本只能三万字,而古籍超过三万字的很多。可喜的是,2005年国家已颁行了超大字符集:GB18030-2005《信息技术中文编码字符集》,其四字节字字符编码容量为1611668个码位。《三国》、《史记》等都只几十万字。再造善本或古籍整理目录中标注上千万字者,非一人所能为。长版刻本为多人协同,书法、刻字风格亦有协同。多重重复字字库表现长版刻本,可取其一部分(譬如前几十万字)生成多重重复字字库,适用到后边页面。超长文本中常用字必定超级重复。如能对超大量的重复字分类整理,合理缩减,不仅可容纳更长内容,制作与使用也会更便捷。

[0006] 现在实际使用四字节字的只有中文。四字节字的编码方案是用双字节汉字中最后两个区扩编而成。按常理,重复字编码,在原双字节码后直接扩编最方便,但四字节字有特殊性,这种方法unicode不支持,排版工具也不支持。这正是采用GB18030-2005《信息技术中文编码字符集》编码的原因。使用国标18030-2005提供的这1611668个码位制成的“超大字符集字库” ttf,为现有排版工具承认并兼容。

[0007] 由于自定义编码,多重重复字字库还能做到“一书一字库”。古籍刻本中常有一本书中多个字体者,很多都是楷书、行书、隶书并存。由于多重重复字字库可容纳多个重复字,容纳多字体更不成问题。用多重重复字字库后,可不再需要为同一古籍刻本中不同字体另做第二个字库,使用起来更方便,也更容易协调。多重重复字字库还可将黑白墨线稿制为字库的一个“字符”,这对古籍刻本中插图的翻印,具有方便存储、编辑、使用、提高印刷质量、降低印刷成本等诸多好处。此外,由于对字库自定义编码和对文本耦合改造的目的是为古籍善本的字库级恢复,而非信息传播,因此多重重复字字库还可有效预防盗版。

[0008] 多重重复字字库是采用现代计算机技术,改造原本产生于西方拼音文字环境并只适应“横排右行”的ttf字库,令其适应中国传统书法文化的同时,还要适应现代桌面出版印刷系统。但要想达成用字库对中文古籍的准确表现,仅用多重重复字字库技术还不够。因为即便达成了多重重复字的存储与表现,现在所有文字编辑工具对“竖排左行”的支持,仅限于用对“横排右行”字库转竖排完成。这种安排本质是将适应横排右行的等高字符,一个落一个地“落”成一列,以此“兼容”竖排。其在ttf中表现为em方阵的纵向叠加。拼音字母高低一致,宽窄不一;可中文书法汉字却恰恰相反:宽窄一致,高低不一。由于ttf中em方阵纵向固定,无法准确表现中文书法字符高低不一的样子,所有高低方向越界的字符图像都会被上限(WinAscent)和下限(WinDescent)截掉。所以,要想在编辑工具中达成对传统中文书法古籍“竖排左行”汉字高低变化的精确表达,多重重复字字库技术须与《CN201410745134-中文竖排古籍、书法专用字库制作方法》结合使用,才能对书法古籍做到字库级完美翻印。

发明内容

[0009] 在通用字库ttf(即TrueType,包括OpenType(otf),下同)中,通过对字符码位的自定义编码,即,为加入ttf的字符图像在unicode合理范围内,将同一字符的多个不同艺术样式(非异体字)存储在同一ttf中,制成“多重重复字字库”。

[0010] 制作和使用多重重复字字库的关键是要对所表达(古籍)的字符图像和txt文本按照“含义顺序对应一致”原则:1,顺序读取一个txt文本中的字;2,取得对应的字符图像;3,重新给定码位并将图像置入ttf字库;4,对该txt文档进行“耦合改造”:令各重复字与新制ttf

中与之对应编码字相一致;改造后的文本,对其适用对应的“多重复字字库”后,原含义字符图像不仅原序显示,且表现出在不同上下文同一字的不同变化造型,从而达到对中文古籍刻本的字库级再现;如用通用字库显示,则含义不明。

[0011] 通过采用超大字符集GB18030-2005《信息技术中文编码字符集》,单个ttf字库可容纳1611668个字以内的古籍刻本原著且被通用文字编辑程序支持;如能对重复量超多的常用字进行样式分类整理,可容纳更长文本。

[0012] 采用多重复字字库,可“一书一字库”;多重复字字库可容纳多个重复字,容纳多字体也不成问题,因此可不必再为同一古籍中不同字体另做字库。

[0013] 还可将二值墨线画稿制作为字库的一个“字符”,这对古籍刻本中有插图著作的翻印也有方便存储、编辑、使用、提高印刷质量、降低印刷成本等好处。