北京大学学报(自然科学版) 第 54 卷 第 1 期 2018 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 54, No. 1 (Jan. 2018)
doi: 10.13209/j.0479-8023.2017.082

基于部件拼接的高质量中文字库自动生成系统

刘成东 连宙辉† 唐英敏 肖建国

北京大学计算机科学技术研究所, 北京 100871; † 通信作者, E-mail: lianzhouhui@pku.edu.cn

摘要 针对中文字库制作开销过大的问题, 提出一种基于部件拼接的高质量中文字库制作方法。参考汉字相关规则与信息, 选取供用户书写的少量汉字集合, 将书写的汉字分割至部件级别。根据汉字部件成字关系, 拼接产生剩余汉字, 最终生成完整 GB2312 标准(共包含 6763 个中文字符)的高质量中文字库。实验表明, 所提出的字库制作系统实现了快速生成个性化中文字库的功能, 在保证生成字库质量的前提下, 显著降低了个性化中文字库的制作时间。

关键词 个性化中文字库; 汉字输入集; 部件分割; 缩放拼接

中图分类号 TP399

Automatical System to Generate High-Quality Chinese Font Libraries Based on Component Assembling

LIU Chengdong, LIAN Zhouhui†, TANG Yingmin, XIAO Jianguo

Institute of Computer Science and Technology, Peking University, Beijing 100871;
† Corresponding author, E-mail: lianzhouhui@pku.edu.cn

Abstract Current Chinese font library generation systems bear a major drawback that the user is required to write all characters contained in the font library, which is rather boring and time consuming. This paper proposes a system to automatically generate Chinese font libraries of high quality based on component assembling. An input set of a few characters for users is selected to write according to the instructive information of Chinese characters. Then components of each written characters are extracted. Several selected components are combined to construct each unwritten character. Finally the complete Chinese personal font library is obtained, which contains 6763 Chinese characters according to the GB2312 standard. Experimental results show that the proposed system can generate personal Chinese font libraries with dramatically shorter time and still keep excellent quality.

Key words personal Chinese font libraries; character input set; component extraction; resizing and assembling

作为世界上最古老的文字系统之一, 汉字在历史发展与文化转播过程中发挥着至关重要的作用。从绘图文字出现至今, 汉字陆续经历了甲骨文、篆书、隶书、楷书、行书及草书等字体。随着印刷技术的发展, 各种印刷字体也应运而生, 作为文化知识的载体, 文字扮演的角色也随之增强。汉字激光照排系统的产生, 解决了汉字在信息化时代的存储以及印刷的历史性难题, 汉字字体也被引入计算机领域。

计算机性能的不断提升促进了字库制作产业的蓬勃发展, 以北大方正电子有限公司和汉仪科印信息技术有限公司为代表的字库公司设计了大量中文字库, 并推广到日常工作生活中。近年来, 网络社交平台在社会生活中扮演着重要的角色, 拥有专属的个性化字库, 并将其应用到网络社交领域, 符合当前提倡的张扬个性的历史潮流。因此, 快速制作高质量的中文字库需要利用字形分析、字形缩放等算法, 在字形学术研究中具有深远意义。

国家自然科学基金(61472015, 61672056, 61672043)、北京市自然科学基金(4152022)和国家语委重点项目(ZDI135-9)资助
收稿日期: 2016–12–07; 修回日期: 2017–04–26; 网络出版日期: 2017–05–04

1 相关工作

利用传统方法制作标准中文字库费时费力，从书写到设计调整，需要花费数月甚至数年的时间。个性化字库的制作成本相对较低，诸多名人与字库公司合作，生成专属个性化字库。北大方正电子有限公司与艺人徐静蕾合作生成了“方正静蕾简体”字库①，该字库由徐静蕾亲自书写 GB2312 标准对应的所有汉字及字母符号，通过方正字库处理技术生成完整字库。汉仪科印信息技术有限公司与作家郭敬明合作，通过书写部分汉字，配合字库自动生成技术，制作了“汉仪郭敬明体”字库②。

明星字体一经推广，便获得广泛关注与好评。受明星效应带动，网络用户对个性化字体的需求与日俱增。但字库公司的相关业务主要面向明星等具有商业价值的客户，对普通个体来说，制作独一无二的字库难以获得技术支持。本研究组开发了 FlexiFont 个性化字库制作系统[1]，该系统与“方正静蕾简体”制作方式类似，用户书写 GB2312 标准对应的完整汉字字符集(以下简称“完整字符集”)，即生成量身定制的个性化字库。然而该系统存在局限性，完整字符集包含 6763 个汉字，需要数周甚至数月的书写周期，导致字库制作效率低下。若能由用户书写少量汉字，然后通过计算机技术对其进行处理，生成完整的字库，便可大幅度缩短字库制作周期，推动全民个性字库的发展进程。

Xu 等[2]提出一种基于笔画表示的手写体汉字生成方法，根据书写汉字生成其他文字。但是，由于算法的局限性，该方法需要大量人工干预，不能实现自动化。由其衍生的相关方法[3]仍然存在类似问题，无法推广到实际应用中。

目前，基于部件拼接复用的汉字字库构建方法[4–5]是快速制作中文字库的主要手段，将相同部件复用在不同的文字上，可以实现字库快速制作以及字库压缩等目的。然而，该类方法面向标准的中文矢量交叉字库，在个性化中文字库制作过程中，普通用户可提供最方便的数据为文字图片，因此面向矢量交叉字库的方法不适用于普通用户个性化字库制作的全面推广。在部件缩放过程中，该方法不能维持笔画宽度不变，在缩放拼接后，会导致文字笔画一致性产生失真，因此无法达到高质量字库快速制作的工业化需求。其他研究者提出面向英文字库[6]以及日文字库[7]的制作方法，也无法直接扩展为中文字库制作。

本研究开发一种基于部件拼接的高质量中文字库自动生成系统，用户只需书写 775 个汉字，便可获得完整字符集对应的中文字库。

2 字库生成系统

本文利用部件分割拼接的方式，根据用户书写的少量汉字产生完整字符集，最终生成高质量的中文个性化字库(以下简称字库)。具体步骤如下：1) 确定输入字符集供用户书写，收集用户书写数据；2) 根据参考字信息，将用户书写的汉字分割至部件级别；3) 根据汉字部件成字关系，利用分割获得的

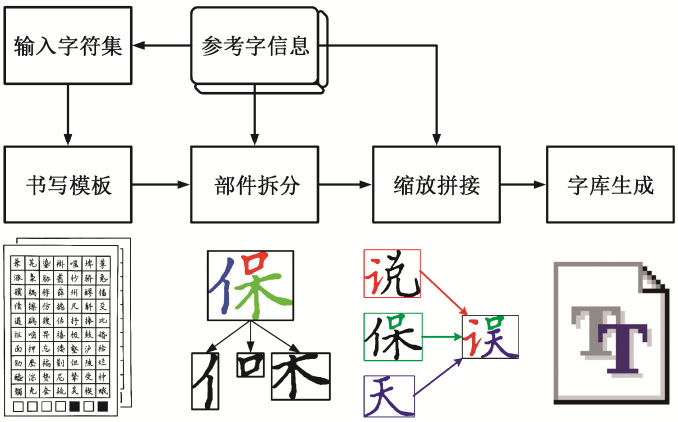


图 1 系统流程图
Fig. 1 System work flow

① https://baike.baidu.com/item/%E6%96%B9%E6%AD%A3%E9%9D%99%E8%95%BE%E7%AE%80%E4%BD%93/3808038?fr=aladdin
② https://baike.baidu.com/item/%E6%B1%89%E4%BB%AA%E9%83%AD%E6%95%AC%E6%98%8E%E4%BD%93/17506765?fr=aladdin

部件进行缩放拼接，得到完整字符集的汉字图片；4) 利用字库制作工具，将字符图片处理生成 TrueType 文件。

图 1 是字库生成系统的流程图。

2.1　输入集选取与数据采集

汉字是一种会意文字，每个汉字由若干部件组合而成。本文参考 2009 年发行的《现代常用字部件及名称规范》[8]，并根据文字书写及表意规则，将完整字符集包含的汉字划分为 19182 个部件。同时，根据汉字成字规则以及书写习惯，人工将所有部件归组分类。设部件类集合为 RadicalSet，输入集为 InputSet，输入集的选取满足：

$$\forall s(s \in \text{RadicalSet}) \rightarrow \exists c(c \in \text{InputSet} \wedge r \in c \wedge r \in s),$$

即输入集需要覆盖所有部件类集合。s 为部件类，c 为输入集包含的文字，r 为文字 c 包含的部件。

根据教育部 1986 年统计数据，日常使用频率最高的 3500 个汉字累积使用频率为 99.48%[9]。基于新浪微博社交平台数据统计的最新汉字累积使用频率如图 2 所示。完整字符集包含大量生僻字，由于生僻字既不常用，也不易书写，因此输入集的选取中尽量避开生僻汉字。这样既能保证采集美观度较高的优质书写数据，又能提高用户亲自书写汉字在实际使用中出现的概率。

复杂汉字包含较多笔画及部件，书写时间开销大且美观度较差。根据书法家经验，部件数量为 2 或 3 的汉字书写美观度最佳。另外，计算机技术处理复杂汉字时难度更大，因此为部件数量较少的汉字赋予更高的选取“优先级”。本模块还增加预选功能，根据实际需求预选特定汉字。

首先通过预选功能将生活中最常用的 170 个汉字加入输入集 InputSet，这些汉字的累积使用频率接近 50%。然后迭代处理部件类，直到所有部件类被输入集覆盖。标记未被 InputSet 覆盖的部件类，利用贪心算法，循环处理每一类部件。对于常用字包含的部件类，优先选取使用频率较高的常用字。部分部件类未被常用字包含，因此从非常用字中选取“优先级”较高的汉字。按照以上处理流程，本系统选取规模为 775 个汉字的输入集(共 1012 个部件类)，仅包含 57 个非常用字，且所有汉字的累积使用频率为 59.2%。输入集选取流程如图 3 所示。

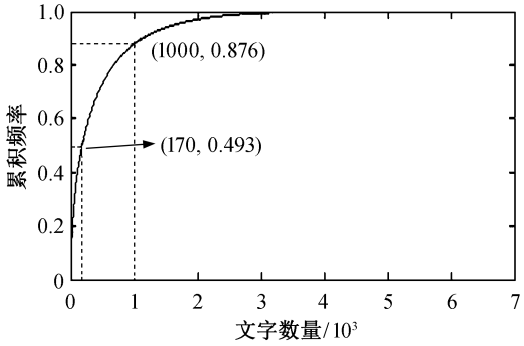


图 2　汉字累积使用频率

Fig. 2　Cumulative frequency of Chinese characters

利用本研究组开发的在线数据采集平台 FlexiFont，将输入集书写模板提供给用户，用户下载模板并完成书写工作，然后将模板拍照上传。

2.2　输入集部件分割

汉字部件分割方法已得到学术界关注。Ma 等[10−11]提出若干种部件分割方法，但无法解决复杂汉字部件分割的问题。我们利用本研究组提出的基于数据驱动的部件分割方法[12]，参考楷体的相关数据，对书写的独立汉字图片进行部件分割，并获得良好的分割结果。首先提取书写目标字的骨架，然

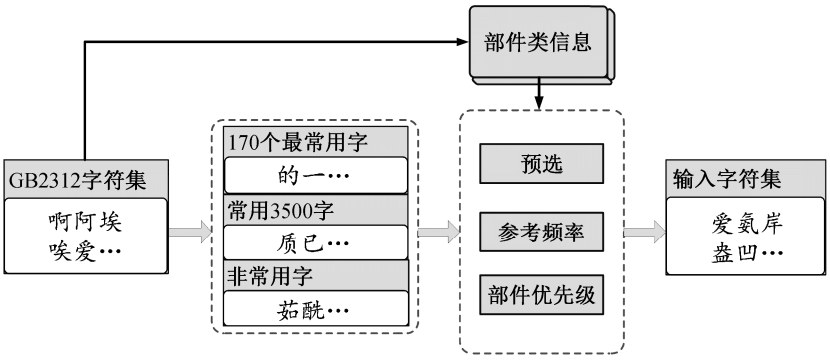


图 3　输入集选取流程

Fig. 3　Work flow of input set selection

后将目标字骨架点与参考楷体字的骨架点进行匹配。根据匹配结果对目标字进行标注分割，确定每个部件对应的图像信息，所有分割标注的部件组成后续拼接工作的基础部件集合。本研究组还开发了手动分割汉字的软件，对部件分割效果差的汉字进行简单人工干预，以提升拼接字库视觉效果。

2.3 缩放拼接产生所有汉字

将输入集作为训练集，剩余汉字作为测试集。根据部件分类关系以及成字规则，为测试集的每个汉字选取合适的基础部件。根据参考字测试集部件与选取的基础部件的尺寸，确定部件的缩放比例 Scale，并计算缩放后部件的位置。

笔画宽度是汉字的重要结构信息，尤其是硬笔字体，整体笔画宽度基本上一致。对于测试集中的汉字，由于其基础部件可能来自不同的汉字，不同部件的缩放比例也存在差异，导致缩放后拼接生成的汉字笔画宽度不一致，严重失真而影响拼字效果。传统的插值缩放方法[13]无法维持汉字的笔画宽度，自然图像的缩放方法[14–21]无法处理文字图像。本文提出一种维持笔画宽度的缩放算法，将部件抽象为轮廓与骨架，确定轮廓点与骨架点的控制关系；然后利用仿射变换对骨架点进行平移，根据轮廓点与骨架点的相对位置关系，确定轮廓点的目标位置；最后将轮廓点闭合填充，得到缩放结果。

首先进行轮廓与骨架提取，本文采用 Canny 算子进行轮廓检测，根据轮廓点的连通关系，将其保存为有序的点集 P_{c}。利用细化方法[22]进行骨架提取，去除骨架中的分叉点以及端部的冗余骨架信息，设骨架点集为 P_{s}。

然后确定轮廓点与骨架点的控制关系，为每个轮廓点分配一个骨架控制点。设 $P_{\mathrm{u}}=P_{\mathrm{c}}\cup P_{\mathrm{s}}$，利用 Delaunay方法[23]对 P_{u} 进行三角剖分，同时剔除不在前景区域的剖分结果，设有效剖分结果的三角形集合为 T。选取既包含轮廓点又包含骨架点的三角形

$$t\in T\wedge\exists v_i\in P_{\mathrm{c}}\wedge\exists v_j\in P_{\mathrm{s}},$$

确定轮廓点与骨架点之间的控制关系，其中 v_i 和 v_j 是三角形 t 的顶点。处于同一三角形中的轮廓点被骨架点控制。若轮廓点被多个骨架点控制，则将控制该轮廓点所有骨架点的质心作为控制点。

由于Delaunay三角剖分方法的特殊性，部分三角形只包含轮廓点或骨架点，致使轮廓点无对应控制点，造成细节信息丢失，因此需要进行特殊区域处理。遍历轮廓点集 P_{c}，如果轮廓段S上所有点都无对应控制点，且 S 包含轮廓点数量大于阈值 L (L 为平均笔画宽度的1/2)，则将轮廓段 S 前一邻居轮廓点 p_i 的控制点 $\overline{p}_k$ 设置为 S 上所有轮廓点的控制点，以保持笔画端部及拐点处的细节信息；若S包含轮廓点数量小于 L，则舍弃 S。确定控制点集合 P_{control} 与有效轮廓点集合 P_{contour}，对于 $p_i\in P_{\mathrm{control}}$ 且 $\overline{p}_k\in P_{\mathrm{contour}}$ 的点对 $<p_i,\overline{p}_k>$ 表示p_i和 $\overline{p}_k$ 之间存在控制关系，计算点对的相对位置关系:

$$D_j=p_i-\overline{p}_k。$$

根据缩放比例 Scale 对 P_{control} 中的控制点进行仿射变换，得到目标控制点集合 P'_{control}。遍历 P_{contour} 并对轮廓点进行仿射变换，得到目标轮廓点集合 P'_{contour}。对于原始轮廓点 $\overline{p}_k$ 及其目标轮廓点 $\overline{p}'_k$，存在

$$\overline{p}'_k=p'_i-D_k,\ <p_i,\overline{p}_k>,$$

p'_i 是控制点 p_i 仿射变换后的位置，计算方式为

$$p'_i=p_i\times\mathrm{Scale}。$$

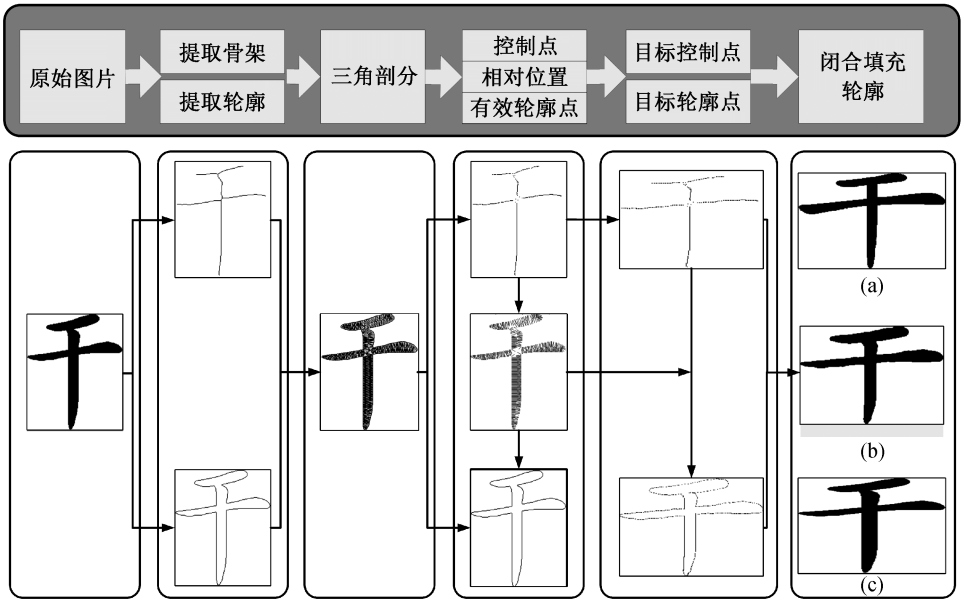
利用直线将 P'_{contour} 中相邻轮廓点连接，得到闭合轮廓 C，并以 P'_{control} 为种子对 C 进行填充。最后对填充后的结果进行闭运算以减少噪声，得到最终的缩放结果。该算法维持了部件在不同缩放比例下的笔画宽度。缩放算法流程如图 4 所示。根据参考数据的部件位置，将缩放后的部件放置在合适的位置，获得目标字图片。

2.4 字库生成

利用文献[1]的字库生成技术，将完整字符集对应的文字图片进行矢量化，并生成 TrueType 文件。经代码重构后，可以在数分钟内产生字库文件。

3 实验测试与结果分析

本文设计文字图灵测试[24]来验证系统生成完整字库的仿真性。从输入集以及拼接产生的字符集中各随机选取 100 个汉字作为测试数据，将其随机展示在同一页面。从输入集中另选 50 个汉字作为展示样例，邀请用户判断测试数据中每个汉字是否为用户亲自书写。若测试用户的判断准确度为 50%，则证明计算机生成的汉字字形与书写者的字形相似度极高。



(a) 提供缩放比例的目标部件; (b) 采用本文方法缩放获得的部件; (c) 采用传统插值方法缩放获得的部件

图 4　维持笔画宽度缩放算法流程

Fig. 4　Stroke-width preserving resizing method

利用本研究组开发的数据采集平台，选取 3 名书写者 775 个汉字的输入集，通过本文的字库制作系统生成完整字库。邀请 20 名没有字形计算技术背景的用户参与图灵测试，结果表明用户的识别率为 55%，证明本文提出的方法取得良好的仿真效果。

已有方法无法达到快速和自动化的要求，不适合普通用户的个性字库制作。本文方法只需用户书写少量汉字，加以微量人工干预，便在数小时内生成实用性强、仿真度高的个性化字库。本系统主要时间开销如表 1 所示。

不同书写者书写模块以及部件人工干预模块的时间开销存在差异，平均生成一套完整字库需要的时间约为 7.5 小时，除去 1.5 小时的人工干预(需要进行人工部件分割的平均文字数量为 120 个左右)，其他步骤均完全自动化，可达到产业化、快速制作个性化字库的目标。

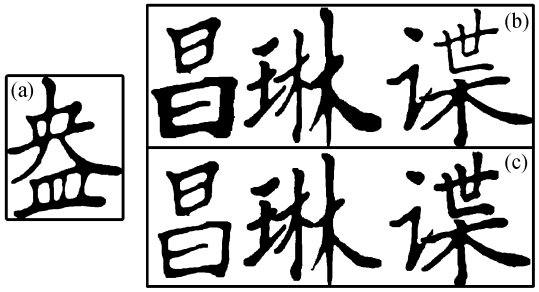
表 1　系统时间开销

Table 1　Time cost of auto-generation system

处理模块	时间/h
用户书写	≈1
部件自动分割	4
部件分割人工干预	≈1.5
产生完整字库	< 1

维持笔画宽度的部件缩放算法是拼字系统的重要组成步骤。图 5 为不同缩放方法拼字结果的比较，可以看出，本文提出的字形缩放方法既可以维持生成汉字笔画宽度的一致性，又能保证生成字形与用户书写风格相同。

本文提出的部件缩放算法亦可用于汉字缩放，解决非等比例缩放过程中传统插值方法导致汉字的笔画宽度不一致的问题。图 6 是对硬笔楷书及楷体两种字体的文字横向拉伸 1.5 倍的结果，可以看出，与插值方法相比，本文方法很好地维持了汉字笔画宽度，保证了汉字书写的整体风格。



(a) “盎”字由用户书写; (b) 利用传统插值缩放算法的拼字结果; (c) 本文提出的维持笔画宽度的部件缩放算法的拼字结果

图 5　不同缩放方法拼字结果比较

Fig. 5　Generated characters based on different resizing methods

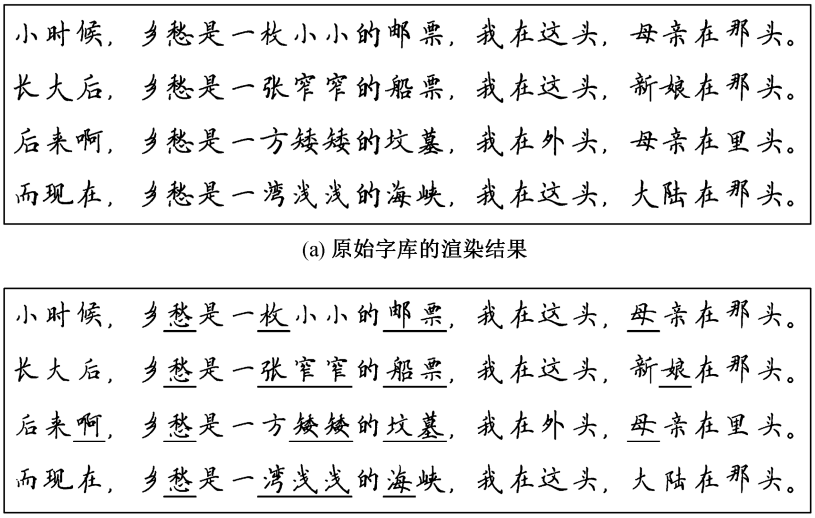
图 6 不同缩放算法结果

Fig. 6 Resizing results of our method and interpolation

基于部件拼接复用的汉字字库构建方法[4–5]处理对象为标准矢量字库文件，而本文方法的处理对象为手写体文字图片，因此无法进行直接实验对比。两种方法的基本思路均为部件复用，然而在实际操作过程中，本文方法对用户亲和度更好。输入集的文字累积使用频率为59.2%，日常工作生活中，使用的汉字约有60%为用户亲自书写，而生成的文字与用户书写风格类似，保证了高仿真的特性。图 7 为利用本系统对“方正佩安手写体”的实验结果与原始字库的比较。系统生成的字库可与原始字库相媲美，达到快速制作的行业标准，可推动字库制作的工业化进程。图 8 为利用本系统生成的 3 套字库渲染文字的效果，其中英文字母与标点符号为用户亲自书写，左侧英文缩写为相应字库名称。

4 结语

本文提出基于部件拼接的高质量中文字库自动生成算法。参考《现代常用字部件及名称规范》、汉字使用频率以及汉字书写习惯等信息，选取固定输入集供用户书写。然后通过部件分割技术，将书写的汉字分割至部件级别。同时提出能够维持笔画宽度的部件缩放算法，根据楷体参考数据以及汉字成字关系，对分割的部件进行缩放拼接，自动生成其他汉字。最后利用字库生成技术，获得完整的矢量化手写体中文字库。图灵测试及实验结果表明，该系统具有制作周期短，字库实用性强，仿真度高等优点，能够快速自动生成与用户亲自书写结果相



(a) 原始字库的渲染结果

(b) 生成字库的渲染结果

下划线标注文字由系统生成，未标注文字包含在输入集中

图 7 个性化字库与原始字库渲染结果比较

Fig. 7 Rendered results of auto-generated Chinese font library and original PAT font library

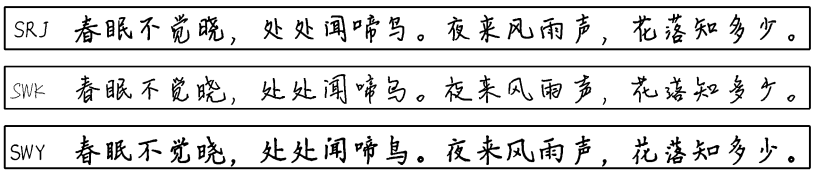


图 8 系统生成个性化字库展示

Fig. 8 Results of auto-generated Chinese font libraries by proposed system

媲美的中文字库。

本方法存在的不足是，分割算法未能实现完全自动化，需进行少量人工干预；此外，由于骨架提取算法的局限性，为缩放拼接过程中部件缩放比例过大时，会在笔画交点及端部产生失真。未来将考虑优化部件分割算法，同时寻找适合汉字的骨架提取方法，实现高质量字库的完全自动生成。

参考文献

[1] Pan W, Lian Z, Sun R, et al. FlexiFont: a flexible system to generate personal font libraries // Proceedings of the 2014 ACM symposium on Document Engineering. Fort Collins, 2014: 17–20

[2] Xu S, Lau F C M, Cheung W K, et al. Automatic generation of artistic Chinese calligraphy. IEEE Intelligent Systems, 2005, 20(3): 32–39

[3] Kwok K W, Wong S M, Lo K W, et al. Genetic algorithm-based brush stroke generation for replication of Chinese calligraphic character // IEEE Congress on Evolutionary Computation. Vancouver, 2006: 1057–1064

[4] 冯万仁，金连文．基于部件复用的分级汉字字库的构想与实现．计算机应用, 2006, 26(3): 714–716

[5] 唐英敏，张艳霞，吕肖庆．基于汉字构形的 TrueType 字库压缩方法．微电子学与计算机，2007，24(6): 52–55

[6] Phan H Q, Fu H, Chan A B. FlexyFont: learning transferring rules for flexible typeface synthesis. Computer Graphics Forum, 2015, 34(7): 245–256

[7] Dolinsky J, Takagi H. Synthesizing handwritten characters using naturalness learning // IEEE International Conference on Computational Cybernetics. Gammarth, 2007: 101–106

[8] 中华人民共和国教育部．GF0014-2009 现代常用字部件及部件名称规范[S]．北京：语文出版社, 2009

[9] 国家语言文字工作委员会．现代汉语常用字表．北京：语文出版社, 1988

[10] Ma L L, Liu C L. On-line handwritten Chinese character recognition based on nested segmentation of radicals // Chinese Conference on Pattern Recognition. Nanjing, 2009: 1–5

[11] Ma L L, Liu C L. A new radical-based approach to online handwritten Chinese character recognition // 19th International Conference on Pattern Recognition. Tampa, 2008: 1–4

[12] Lian Z, Xiao J. Automatic shape morphing for Chinese characters // SIGGRAPH Asia 2012 Technical Briefs. Singapore, 2012: No. 2

[13] Keys R. Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(6): 1153–1160

[14] Chen L Q, Xie X, Fan X, et al. A visual attention model for adapting images on small displays. Multimedia Systems, 2003, 9(4): 353–364

[15] Santella A, Agrawala M, DeCarlo D, et al. Gaze-based interaction for semi-automatic photo cropping // Proceedings of the SIGCHI conference on Human Factors in computing systems. Quebec, 2006: 771–780

[16] Setlur V, Takagi S, Raskar R, et al. Automatic image retargeting // Proceedings of the 4th international conference on Mobile and ubiquitous multimedia. Christchurch, 2005: 59–68

[17] Avidan S, Shamir A. Seam carving for content-aware image resizing. ACM Transactions on graphics (TOG), 2007, 26(3): No.10

[18] Achanta R, Süsstrunk S. Saliency detection for content-aware image resizing // 16th IEEE International Conference on Image Processing (ICIP). Cairo, 2009: 1005–1008

[19] Wang Y S, Tai C L, Sorkine O, et al. Optimized scale-and-stretch for image resizing. ACM Transactions on Graphics (TOG), 2008, 27(5): No. 118

[20] Zhang G X, Cheng M M, Hu S M, et al. A shape-preserving approach to image resizing. Computer Graphics Forum, 2009, 28(7): 1897–1906

[21] Chang C H, Chuang Y Y. A line-structure-preserving approach to image resizing // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, 2012: 1075–1082

[22] Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 1984, 27(3): 236–239

[23] Shewchuk J R. Triangle: engineering a 2D quality mesh generator and Delaunay triangulator // Applied Computational Geometry Towards Geometric Engineering. Berlin, 1996: 203–222

[24] Baird H S, Coates A L, Fateman R J. PessimalPrint: a reverse Turing test. International Journal on Document Analysis and Recognition, 2003, 5(2/3): 158–163