汉字字形结构式压缩方法的研究和实现

高锐智 华成英
(清华大学自动化系 北京100084)

Chineses Font Structural Compression Method and Implementation

GAO Rui-Zhi HUA Cheng-Ying
(Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract Because outputing chinese font in computer is limited in complicated building process and mass storage, we propose a structural-based compression scheme. The approach bases on the stroke-assembled shape character of chinese font, extracts stroke outlines and clusters these outlines to create stroke template. This paper also introduces font structural compression software developed by us, explains the process of our structural-based compression scheme.

Keywords Technology of chinese font, Structural font, Outline font, Stroke extraction, Stroke outline clustering

1 引言

汉字字型技术作为计算机汉字信息处理技术的重要组成部分,经历了点阵汉字、矢量汉字发展到现在广泛应用的曲线轮廓汉字,其研究和应用取得了长足的进步[2]。汉字字形有两个特点:汉字是基本笔划和字根在二维空间的组合,其字形结构复杂,笔划数目变化大,而且不同字体的形状特点也变化很大;汉字的字符集十分庞大,其中一二级汉字字符集(GB2312)就包括6763个汉字[2],刚刚颁布施行的GB18030标准的编码字符更多达2.6万个。现有的汉字字型技术大都是在引进和吸收西文字型技术的基础上发展起来的,很少考虑到汉字字形的特殊性。因此,现有的汉字字型技术的字形制作生成工作十分复杂,且字形占用的存储空间很大,这样的缺陷限制了汉字字形的应用,特别是在PDA等存储空间有限的设备上的应用。这两个问题也因此成为汉字字型技术研究的重点和难点。本文简要介绍现有汉字字形压缩技术,着重研究正在兴起的汉字字形结构式压缩方法,并且针对结构式压缩方法在字形生成中存在的问题,结合我们开发的汉字结构式字形计算机辅助设计系统介绍基于笔划抽取和聚类的轮廓字形压缩方法及其实现过程。最后,通过对实验结果的分析给出结论。

2 汉字字形压缩

点阵式字形[2]是计算机字形最初的表示形式。点阵式字形技术是将字形转换为m×n(通常m=n)点阵的图片存储和输出。它的优点是输出处理简单快速,并且可以通过制作不同大小点阵的字形库在各种输出分辨率下得到高质量的字形。但是,汉字点阵式字形占用的存储空间大,且点阵式字形库只能输出固定大小的点阵式字形,要获得不同大小的字形,就需要保存多个点阵式字形库,这又大大增加了字形的存储空间。点阵式字形的制作工艺复杂,需要有经验的设计人员对字形逐个进行字模设计,再利用计算机辅助设计的方法,通过人机交互将字模转存为点阵式字形库。

点阵式字型技术的局限,促使字型技术研究人员深入研究字形的压缩表示形式,由此产生了多种字形压缩技术。字形轮廓压缩法和结构式字形压缩法是两种最具代表性的汉字字形压缩方法,我们在表1中概括了两种方法的特点。

表1 汉字字形压缩方法的比较

汉字字形压缩方法	压缩率	失真率	字形复原速率	压缩字形自动变换能力	压缩字形自动生成
字形轮廓压缩	中	极低	中	较好	可以
结构式压缩	高	低	低	好	很难

2.1 字形轮廓压缩法

图形压缩法是字形压缩的基本方法,它是针对字形的二维图形信息进行压缩的方法,适用于西文和汉字字形的压缩存储。字形轮廓压缩法[2]是技术最成熟、应用最广泛的一种图形压缩法,它利用字形轮廓与点阵式字形一一对应的关系,通过获取字形边缘线并对其进行编码表示实现字形的压缩存储。轮廓字形的生成过程主要包括:轮廓抽取、轮廓线描述和建立字形轮廓库三个步骤。轮廓抽取是从点阵字形获取字形的边缘线;轮廓线描述过程用数学表达方法描述抽取得到的字形边缘线。这两步是字形轮廓压缩法最重要也是最难的部分,现在已经有多种算法可以实现从点阵字形到轮廓字形的自动转换,大大简化了轮廓字形库的制作过程。用字形轮廓法保存的字形,在还原输出时可以通过线性变换改变大小,并且根据输出设备分辨率和字形大小对线性变换后的字形轮廓线作进一步的变形处理,最后经过点阵化将轮廓线还原为输出终端可以显示的图片。轮廓字形经过还原可以在不同的设备上得到大小不同且保持字形原有设计特征的高质量字形。与点阵式字形相比,轮廓字形大大压缩了字形存储占用的空间,还原字形质量高,而且生成过程简单。但是,轮廓字形还原过程中的点阵化和变形处理降低了字形显示的速度。

2.2 结构式字形压缩法

与西方文字不同,汉字是一种非字母化,非拼音化的象形文字。汉字的最小构成单位是笔划,由数十种笔划组成了数百

高锐智 硕士生,主要研究方向为中文信息处理和图像处理。华成英 副教授,主要研究方向为电子技术及微机应用。

种结构块，再由这数百种笔划结构块按照一定的方位关系组成了数万个汉字，一般称这些笔划结构块为“部件”。汉字的信息量主要由部件及其组合来体现[3]。汉字字形结构式压缩方法正是根据汉字结构化特点提出的压缩方法，其核心思想是把字形信息转换为笔划或者部件形状和字形结构两部分信息分别保存，字形还原时用两部分信息组合得到整字字形。结构式压缩法按照组字基本元素（或称为组件）的不同，可以分为笔划组合式和部件组合式两大类。笔划组合式方法将笔划作为组成字形的基本元素；而部件组合式方法将偏旁、部首和独体字等部件作为组成字形的基本元素。结构式字形的生成过程主要包括：建立组件模板和获取组件结构关系两个步骤。建立组件模板的过程是通过分析字形库中包含的大量字形的拓扑结构和形状特征，得到组成字形库中字形所需的笔划或者部件的形状信息。结构关系的获取首先要根据现有模板将字形拆分为若干模板组件的组合，再参照原字形获得这些组件还原字形的结构信息。由于汉字笔划组字形式复杂，不同字体笔划形状差别大，建立组件模板的工作十分困难，需要设计者拥有丰富的字形设计经验并且耗费大量的时间，因此实现字形的自动生成难度很大。结构式字形的还原首先要根据字形的组成在模板中找到相应的组件，再根据结构信息调整组件的位置和形状，最后再对组件进行变形和点阵化处理，得到输出终端可以显示的图片。与字形轮廓法的字形相比，结构式字形用组件代替了整字轮廓，避免了字形信息的重复存储，进一步压缩了字形数据，结构式字形通过对组件轮廓的线性变换改变字形大小，因而可以获得更好的字形变换效果。但是，结构式字形模板中组件数量的多少会直接影响字形的还原效果，组件越少，还原字形细部特征的能力越差，还原字形的质量也越差，而且结构式字形的还原过程更复杂，因而速度更慢。

3 汉字字形结构式压缩方法的实现

3.1 参数图形学方法

现有的结构式字形多采用参数图形学的方法描述组件和字形结构信息。这种方法提取组件拓扑结构和各部分形状的特征作为组件形状控制参数，通过调整参数实现组件在字形重组中的结构和形状还原。D. E. Knuth 等对西文参数式字型技术进行了全面深入的研究，并且在文[8]中介绍了西文参数式字形设计系统 METAFONT 的原理和使用方法。但是，汉字笔划字形比西文字形更复杂，变化更多样，如果沿用西文字形的参数化方法，不能得到高质量的字形还原效果。文[4,5,7]分别提出了三种汉字结构式字形设计和表示方法。这三种方法都从汉字书写方式入手，采用汉字字形骨架为组件形状控制的基本参数，同时加入笔划宽度和笔划端部（起笔、笔锋和转折）的形状特征作为辅助参数。选择骨架作为基本参数是因为字形骨架可以反映字形的拓扑结构，能够代表笔划书写过程中的走笔路径，与控制笔划粗细的宽度参数共同作用就可以还原笔划字形骨干。汉字字形起笔、笔锋和转折处的形状集中体现字体特点，并且与笔划类型和字形结构紧密相关，因而其形状复杂、特征多变。要得到高质量的还原字形必须在字形骨架的基础上加入这部分形状信息。这种形式的结构式字形，可以还原得到高质量的汉字字形，进一步压缩了字形数据。但是，这种形式字形的设计需要对字形骨架的精确描述，并且需要具有丰富的字形设计经验的字形设计人员来完成端部的形状控制，字形设计制作工作效率难以提高。

3.2 基于笔划抽取和聚类的方法

针对参数图形学方法存在的字形制作困难的问题，文[1]中提出了一种以笔划抽取和聚类为基础的结构式字形压缩方法。这种字形结构式压缩方法增加了组件模板的信息量，由完整的笔划轮廓组成组件模板，将组件结构信息简化为笔划的位置和尺度。这样的结构式字形表示形式降低了组件结构信息获取的难度，但是同时增加了建立笔划轮廓模板的难度。为了能够简单高效地生成结构式字形，这种方法采用对现有轮廓字形进行结构式转换的字形生成方法。字形压缩过程分三步完成：对现有轮廓字形进行笔划轮廓抽取；对抽取结果聚类生成笔划轮廓模板；拆分字形轮廓为笔划轮廓再进行模板匹配，并且获得字形还原结构的信息。从现有轮廓字形中抽取笔划轮廓是这一方法的创新之处，其好处在于：轮廓字形的生成和笔划抽取都可以由计算机自动完成，提高了笔划轮廓模板的生成效率；对轮廓字形进行笔划抽取可以得到高质量的笔划轮廓；轮廓字形的字体丰富，可以转换生成多种结构式字形。通过笔划轮廓聚类模板，可以在保留笔划轮廓形状特征的前提下压缩重复的笔划轮廓信息，从而提高了字形数据的压缩率。与基于字形骨架和参数式字形的结构式压缩方法相比，这种方法，不需要在字形设计时加入大量的先验知识，从而简化了模板设计的工作，能够实现汉字结构式字形的自动生成。

3.3 笔划抽取

汉字笔划抽取是上述结构式字形压缩方法的重点和难点，也是汉字信息处理中的一个重要课题，其算法主要分为两类[3]：通过细化抽取笔划；直接抽取笔划。结构式压缩方法中笔划轮廓抽取的目的是从字形轮廓中得到独立和尽量完整的笔划轮廓。而细化本身是一个易产生形变而且费时的过程，所以我们采用不经细化直接抽取的方法。笔划轮廓抽取的难点在于消除字符中笔划的粘连与交叉。在这类算法中，凹点信息是处理的关键，文[6]提出的抽取方法取得了较好的效果，并且在各种汉字字型技术中被广泛应用。这种算法基于字符边缘的凹点，并假设这样的凹点成对出现。但是这种假设是不严格的，因为字形中笔划的艺术效果会产生不成对出现的凹点。因此算法中需要加入附加准则删除多余凹点。文[1]中提出了一种文[6]的改进算法，这一算法利用了字形轮廓的两个特征：笔划交叉或是粘连伴随着字形轮廓上凹点的成对出现；字形轮廓同一笔画段节点的曲率一致。基于这两个特性，算法从节点曲率的角度出发，确定凹点的连接关系，进而得到笔划段。这种算法避免了文[6]的算法中直接求取凹点对的不确定性，可以通过笔划段的抽取与连接得到完整封闭的笔划轮廓信息。

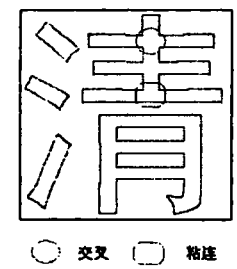

图1 抽取难点

3.4 笔划聚类

笔划轮廓模板的聚类是上述结构式字形压缩方法的另一个难点。对笔划轮廓样本的聚类要通过两个步骤完成，首先将

满足拓扑结构相同和端部形状特征相似的笔划轮廓聚成一类，再从各类中选择具备该类典型形状特点的笔划轮廓作为模板。模板中的笔划轮廓应该满足以下三个要求：完备性，即模板中的笔划轮廓能够组合还原指定字符集中的所有字形；不可替代性，即模板中的每一个笔划轮廓都代表一种拓扑结构或者端部形状特征；典型性，即模板中的笔划轮廓具备某一类笔划轮廓的形状特征的共性。这三个要求保证了压缩后的结构式字形在获得高压缩率的同时保持良好的还原质量。笔划轮廓的聚类过程首先需要提取笔划轮廓的特征，再选用适当的算法在特征空间对笔划轮廓进行聚类并选择模板，其中笔划轮廓特征的选择对聚类的结果影响最大。我们可以提取的笔划轮廓特征有多种，其中包括四边码特征、网格特征和密度特征等。我们还可以通过霍夫变换或者小波变换提取笔划轮廓的其它特征。笔划轮廓的傅立叶频谱是笔划轮廓形状的频域特征，它的低阶和高阶部分分别代表了笔划轮廓的概貌和细节。为了在聚类中突出笔划轮廓的形状特征，我们选择提取笔划轮廓的傅立叶频谱作为聚类的基本特征。

3.5 系统实现

采用以上压缩方法，我们设计和实现了结构式字形计算机辅助设计系统，它可以简单高效地将轮廓字形转换为结构式字形，实现了字形的自动生成。

系统包括字形数据转换模块、笔划轮廓抽取模块、笔划轮廓聚类模板生成模块和字形拆分重建模块四个组成部分。图2描述了系统的工作流程和各模块的输入输出。

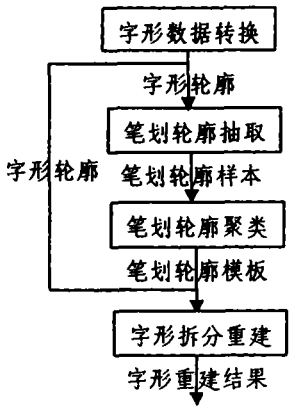

图2 系统总体框架

现有轮廓字型技术的字形轮廓信息格式各不相同，不同的系统平台提供的调用接口差别也较大，为此我们设计实现了字形数据转换模块。字形数据转换模块的主要功能是将不同技术的轮廓字形信息转换为平台独立的、格式一致和满足结构式压缩处理要求的轮廓字形信息。增加这一模块可以使我们的软件方便地移植到不同的系统环境下，并且减少软件升级的复杂性。我们现已开发了在 Windows 平台上，基于系统调用的 TrueType 字体数据转换程序。在开始结构式字形生成的操作前，首先要通过数据转换模块得到统一格式的数据。这些数据主要是字形的轮廓点信息以及其他字形还原的辅助信息。

笔划轮廓抽取模块的主要功能是应用文[1]中提出的笔划轮廓抽取算法从字形轮廓中得到独立和完整的笔划轮廓，并将得到的笔划轮廓保存作为下一步聚类的样本。这一步骤可以由计算机自动完成，其正确率已经超过98%。

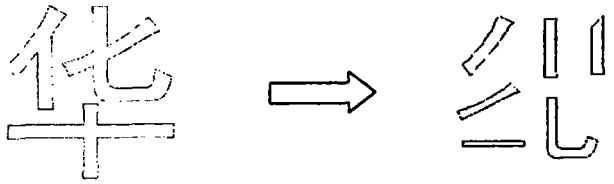

图3 笔划抽取界面

笔划轮廓聚类和模板生成模块对由笔划抽取获得的笔划轮廓样本进行聚类，并且从每一类中选择笔划轮廓产生笔划轮廓模板。我们选择 Mean Shift 算法对笔划轮廓的傅立叶频谱进行聚类。用 Mean Shift 算法对笔划轮廓进行聚类，聚类速度快，可以通过聚类参数的调整实现对模板类别数的控制，并且可以在聚类的过程中完成笔划轮廓模板的选择。图4是我们从聚类结果中随意选出的两类，从中可以看出聚类和模板选择的效果。

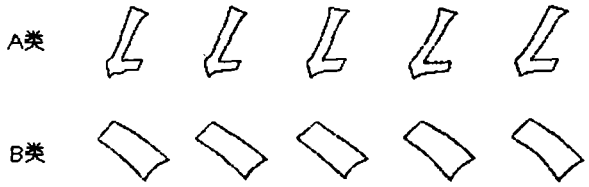

图4 笔划聚类结果(每一类中的第一个为模板)

字形拆分重建模块的功能是利用笔划抽取算法拆分字形轮廓为笔划轮廓，参考笔划轮廓模板获得并保存字形结构信息。字形结构信息的获得使用图形匹配算法，其目的是在模板库中的笔划轮廓中找到使字形还原质量最高的模板并获得用此模板还原字形的结构信息。为了获得最好的字形重建效果，我们提供了人机交互的重建调整功能。操作人员可以参照原字形方便地调整笔划轮廓的位置和尺寸，得到满意的重建效果，并保存结构信息。笔划轮廓模板和字形还原的结构信息就构成了结构化压缩存储的字形信息。

字形结构式压缩过程中各个模块的输入和输出数据主要是形状轮廓信息，为了获得高质量的还原字形，需要大量轮廓样本，系统处理的轮廓信息的数据量很大。我们采用文件作为各模块间的接口，简化了各模块数据结构的相关性，方便了数据的保存和转移。

4 实验结果

我们应用结构式字形辅助设计系统对由 TrueType 黑体字转换得到的轮廓字形进行结构式压缩。我们随机选择了三组字形数据，它们分别包含100，1000和2000个黑体汉字字形。我们分别对每一组字形进行结构式压缩，在字形压缩过程中采用了相同的笔划抽取和聚类算法，并且选择相同的聚类参数。从表2的实验结果可以看出：笔划类别数随着字形数目的增加而增加，但是增加的比例小于字形数目增加的比例，这说

表2 对 TrueType 黑体的压缩结果

字形数目	100	1000	2000
原始数据(字节数)	26298	268665	537756
笔划类别数	206	980	1662
结构式字形数据(字节数)	12807	167916	346853
压缩率(%)	48.7	62.5	64.5

明笔划轮廓模板中包含的笔划轮廓数目随着字形样本的增加趋于稳定;三组字形都获得了较大的压缩比。在图6中给出了三个字形的还原结果,通过与原字形的对比可以看出字形还原的质量比较高。

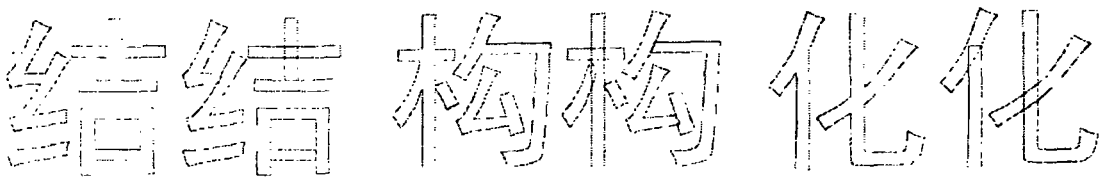

图5 字形还原结果

结论 汉字字形结构式压缩方法利用了汉字字形的特点,能够充分压缩汉字字形中的重复信息。对汉字轮廓字形进行结构式压缩,将字形轮廓信息转换为笔划轮廓模板和字形结构信息,可以在轮廓字形压缩的基础上进一步压缩字形的存储空间,并且得到高质量的还原字形。基于笔划抽取和聚类的结构式压缩方法充分利用了现有轮廓字形的笔划形状信息,简化了组件模板的生成,实现了结构式字形的自动生成,提出了汉字字形结构式压缩技术研究的新方向。我们设计实现的结构式字形计算机辅助设计系统适用于各种的系统平台和字形数据格式,能够实现轮廓字形数据到结构式字形数据的自动转换。它有效地提高了字形设计和制作的工作效率,为汉字结构式字形的应用提供了基础。我们的实验结果证明了汉字字形结构式压缩算法可以获得较高的数据压缩率和高质量的字形还原效果。汉字字形结构式压缩方法同样适用于与汉字字形有相同特点的日文和韩文,因此我们正在将结构式字形压缩的研究成果推广到CJK(中文、日文和韩文)字型技术。

参考文献

1 宋晓丹,罗予频. Outline字体结构式压缩算法及其实现. 中文信息学报,2002,16(3):52~57
2 彭寿全,黄可. 汉字信息处理. 成都:电子科技大学出版社,1994
3 孙星明,胡运发,等. 结构化汉字信息处理. 长沙:国防科技大学出版社,2001
4 赵恒,金通光,王国瑾. 骨架汉字字形存储与显示技术. 中文信息学报,1997,11(1):37~43
5 樊建平. 智能汉字字形设计技术及一个试验性系统ICCDS. 中文信息学报,1990,4(3):1~11
6 Ma Xiaohu, Pan Zhigeng, Zhang Fuyan. Automatic generation of Chinese Outline font based on stroke extraction. Journal of Computer Science and Technology, 1995, 10(1): 42~52
7 Lim Soon-Bum, Kim Myung-Soo. Oriental character font design by a structured composition of stroke elements, Computer Aided Design, 1995, 27(3): 193~207
8 Knuth D E. The METAFONT book. Addison-Welsey, Reading Mass. 1986

(上接第54页)

这方面已经做了很多的工作,建立了民族类和一个包含800个民族左右的民族知识库,已经完成了比较基本的工作。但是,本体还需要进一步的完善,首先属性和关系集还不完全,我们会在今后的工作中不断将其补充完整;其次,也是非常重要的一点,公理库要添加更多的公理,以保证知识的一致性;另外,民族知识库还缺少很多民族的知识,就是已有的民族的知识也需要不断的补充和更新,我们可以通过在网上进行知识挖掘等方式,来补充我们的知识库。

参考文献

1 Harvard大学. ad2000项目. 网址 http://www.ad2000.org
2 Bowden P R, Halstead P, Rose T G. Extracting Conceptual Knowledge from Text Using Explicit Relation Markers. In: N. Shadbolt, K. Ohara, G. Schreiber, eds. Advances in Knowledge Acquisition. Lecture Notes In Artificial Intelligence, Springer-Verlag, Berlin, 1996, 1076: 147~162
3 Hahn R, Schnattinger K, Romacker M. Automatic Knowledge Acquisition from Medical Texts. Text Knowledge Engineering Lab, 1996
4 Lu R Q. New Approaches to Knowledge Acquisition. World Scientific Publishers, 1992
5 Cycorp. Features of CycL. 网址 http://www.cyc.com
6 曹存根. 面向专家的知识获取. 北京:科学出版社,1998
7 Welty C. The Ontological Nature of Subject Taxonomies. In: Proc. of the First Intl. Conf. (FOIS'98), June 6-8, Trento, Italy. 317~327
8 Guarino N. Formal Ontology and Information Systems. In: Proc. of the First Intl. Conf. (FOIS'98), June 6-8, Trento, Italy. 3~15
9 Guarino N, Welty C. Ontological Analysis of Taxonomic Relationships. In: Intl. Conf. on Conceptual Modeling. Springer-Verlag LNCS Vol. 1920, Oct. 2000. 210~224
10 Guarino N, Welty C. A Formal Ontology of Properties. In: Proc. of EKAW-2000: The 12th Intl. Conf. on Knowledge Engineering and Knowledge Management. Spring-Verlag LNCS Vol. 1937, Oct. 2000. 97~112
11 Smith S, Mark D M. Ontology and Geographic Kinds. Proceedings of the International Symposium on Spatial Data Handling (SDH'98), Vancouver, Canada, July, 1998. 12~15
12 《中国大百科全书》之民族卷. 中国大百科全书出版社
13 Chaudhri V K, Farquhar A, et al. The Generic Frame Protocol 2.0: [SRI International Technical Report]. 1997
14 赵锦元,戴佩丽. 世界民族通览. 中央民族大学出版社,2000
15 Cao Cungen. Extracting and Sharing Medical Knowledge. Journal of Computer Science and Technology, 2002, 3
16 任新建. 略论中国民族关系史上的文化交流和整合. 中国传统文化网中华文化研究通讯栏目导航,1999(7)
17 张德海,曹存根,张宇翔. 国家和城市知识获取与本体论分析. 中国人工智能学会第九届全国学术年会暨中国人工智能学会成立20周年庆祝大会,2001. 366~370
18 唐素勤,曹存根. 智能教学系统:综述与改进. 中国人工智能学会第九届全国学术年会暨中国人工智能学会成立20周年庆祝大会,2001. 1129~1132