古文字自动识别过程及其程序实现

张霄军　陈小荷

【摘　要】计算机硬件的发展使得大规模古文字字库建设成为可能,《古文字诂林》等大型古文字释类工具书的编纂出版为古文字库建设提供了丰富的资源,人工智能技术的不断发展为古文字自动识别算法的演进提供了条件。众多古文字学家和考古学家对古文字结构、异体、义释、语法等方面的大量研究工作也使得古文字自动识别有了理论上的依据。因此,在硬件、资源、技术和理论四方面都取得长足进步的同时,古文字自动识别技术的研究和开发也就提上了议事日程。古文字自动识别的一般过程为:(1)进入 OCR 系统,获取识别图像;(2)对图像预处理,获取文字轮廓;(3)对文字进行切割,并提取其结构特征;(4)单字识别;(5)自动识别后处理;(6)必要的人工文字校对,识别结束。本文运用数字图像识别与处理原理,在 Visual C++下编译了部分程序代码,可以实现以上步骤,并具有一定扩展性。

【关键词】古文字;自动识别;程序化

一、大规模古文字自动识别实现的可能性

随着中文信息处理技术的发展,古文字自动识别技术也势必成为研究的重点,而大规模古文字自动识别的实现则依赖于计算机硬件的升级、人工智能技术的发展、基础资源的建设和理论研究的深入。如今,这四方面条件都取得了长足的进步,大规模古文字自动识别的实现也就可以预期了。大规模古文字自动识别技术对于我国古文字考证、出土文物鉴定、古文字辞书编纂等都有着重要意义。

1. 计算机硬件的升级

计算机硬件是指构成计算机系统的所有物质元器件、部件、设备,以及相关的工作原理与设计、制造、检测等技术的总称。元器件包括集成电路、印刷电路以及其他磁性元件、电子元件等。第一代计算机以电子管作为主要元器件,第二代计算机以晶体管作为主要元器件,这两代计算机体积大、功耗大、故障率高、运算速度低、不能用于文字处理。从第三代计算机开始以集成电路作为主要元器件,集成电路的发展大大促进了计算机体系结构和硬件的完善。目前普遍使用的是第四代计算机,其主要元器件采用大规模集成电路,具有运算速度快、容量大、体积小、功耗少、可靠性高、应用范围广的特点,可以广泛地应用于文字和文本处理;计算机系统的部件和设备包括中央处理器(central processing unit, CPU)、存储器、输入输出设备和电源等。中央处理器是计算机内部对数据进行处理并对过程进行控制的部件,由运算器、控制器等组成。早期计算机为了节省成本,一般采用串行运算器,运算速度慢,后来逐渐采用并行运算器,大大加快了计算机运算速度,现在的计算机普遍采用浮点运算器,扩大了数据处理的应用范围。现代计算机普遍采用微程序控制器(micro - programmed control unit, MCU),可以实现不同计算机间指令兼容问题。现代计算机的中央处理器有向微处理器发展的趋势,即随着大规模集成电路技术的迅速发展,芯片集成度越来越高,中央处理器可以集成在一个半导体芯片上,甚至一个芯片上能集成多个处理器,功能也不断增强。存储器是用来储存程序所需的数据和指令信息。过去曾经使用磁心存储器和磁鼓存储器,存储容量有限,现代计算机根据不同的功能、结构与工作原理,存储器的介质也不相同,主要有半导体存储器、磁盘存储器、磁带存储器和光盘存储器等。输入输出设备也由简单的批次输入(纸带输入机、软盘输入机)发展成为交互式输入(键盘、鼠标、触摸屏)和语音、

[作者简介] 张霄军,南京师范大学文学院博士生,陕西师范大学外国语学院讲师。研究兴趣为中文信息处理。
陈小荷,南京师范大学文学院特聘教授,博士导师。研究兴趣为中文信息处理。

文字、图像输入设备等,由单一的显示输出设备发展为印刷输出设备、语音输出设备和绘图仪等。此外,计算机制造、检测和维护等技术也日新月异,完全能适应大规模文本与文字处理的需求。

2. 人工智能技术的发展

人工智能是研究解释和模拟人类智能、智能行为及其规律的一门学科,其主要任务是建立智能信息处理理论,进而设计可以展现某些近似于人类智能行为的计算系统。人工智能的研究已经有近50年的历史,发展是曲折的,目前在专家系统、机器翻译、机器视觉和问题求解等方面的研究已有实际应用。近年来对人工神经网络的知识表示、常识推理、机器学习和分布式人工智能等基础性研究也取得了可喜的进展。初期的人工智能技术都是基于物理符号机制和启发式求解的连接机制,近年来有人提出无需表示、无需概念的智能观,对逻辑在人工智能中的作用、知识与概念化、认知与学习、认知与感知、计算智能与人工智能的关系等问题开展了有益的辩论。此外,多学科交叉、人机一体化等观点也影响着人工智能的研究。

人工智能计算系统的实现依赖于程序设计语言和计算机程序设计方法。程序设计语言的好坏不仅影响到程序使用是否方便,还涉及到程序员所写程序的质量。程序设计语言的发展经历了从低级到高级的发展阶段,低级语言包括字位码、机器语言和汇编语言,其特点是与特定的机器有关,功效高,但使用复杂、烦琐、费时且易出差错。高级语言的表示方法要比低级语言更接近于待解问题,其特点是易学、易用、易维护。程序设计语言的发展趋势是模块化、简明性、形式化、并行化和可视化。模块化是指不仅语言具有模块成分,程序由模块组成,而且语言本身的结构也是模块化的。简明性是指所涉及的基本概念不多,成分简单,结构清晰,易学易用。形式化指要发展合适的形式体系,以描述语言的语法、语义和语用。并行化指发展具有合适并行成分的并行语言。可视化是指要发展"所见即所得"的程序设计语言;程序设计方法是针对某一领域或某一领域的特定一类问题所用的一整套特定的算法。程序设计的发展可以归结为从顺序程序设计到并发程序设计、并行程序设计和分布程序设计,从非结构化程序设计到结构化程序设计,从过程式程序设计到非过程式程序设计、逻辑式程序设计、函数式程序设计、对象式程序设计以及可视程序设计、文化程序设计等,从低级语言工具到高级语言工具。

3. 基础资源的建设

古文字字库建设和汉字大字符集的编制是大规模古文字自动识别的最基本的基础资源建设。《古文字诂林》的顺利出版是我国古文字字库建设的辉煌成果,也为实现和完善汉字大字符集提供了素材和佐证。

《古文字诂林》全书约1400万字,汇集了古今中外几百位学者、专家对中国古文字的考释和论证,搜集了近16万个古字形,是迄今为止涉及古字形最多、考释资料最全的研究古文字的特大型工具书。《古文字诂林》数据库的构造既方便了古文字字库的提取,又兼顾了汉字大字符集的扩充,是大规模古文字自动识别的最重要的基础资源。

《古文字诂林》数据库由三个库构成:字头对象基本属性数据库、字形库和文本资料库。字头对象基本属性数据库详细记录了部首分类属性、拼音属性、古隶定字笔画属性、楷定字笔画笔顺属性和楷定字四角号码的编码属性,共记录了9832个《古文字诂林》字头数据,并向用户提供了检索要素。字形库由五部分组成:篆书字形库、古隶定字形库、古文字字形库、扩充的GBK汉字字形库以及排版专用的字形库。其中古文字字形库是由甲骨文、金文、陶文等八大类古文字字形组成,字形来自《甲骨文编》、《金文编》、《古陶文字徵》等十五部研究著作,经扫描、修补、分类整理而成。文本资料库是由以字头为基本记录单元的文本文件汇集而成,每个文本文件中标有字头标记,下有"字形记录"与"释义记录"两部分,相当于文本知识库。文本文件具有标记性语言的特点,方便计算机自动存取文本。

除了《古文字诂林》数据库提供的古文字字库,《小篆字库》、《商周金文数字化处理系统》和《战国楚文字数字化处理系统》等古文字数字化工具书都提供了相应的古文字字库,可以应用于古文字自动识别处理。

有了字库,还需要有符合国际标准的字符集。从目前情况来看,计算机用汉字字符集已由GB2312-80(6763个字符)、GB13000.1(20902个字符)、GB18030(27564个字符)发展到ISO10646-2000(70275个字符)。而关于汉字古文字在国际标准字符集中编码问题的国际标准化组织(IRG)会议则通过了(1)M22.9决议:IRG接受古汉字兴趣组的报告,认为古汉字应当作为独立于CJK统一汉字的文种而单独编码;(2)

WG2M45.34 决议:扩大 IRG 的工作范围,新的工作范围将包括古汉字和 CJK 笔画的研究;(3)SC2M13 -05 决议:批准扩大 IRG 的工作范围,新的工作范围将包括古汉字和 CJK 笔画的研究。

此外,古文字结构特征库、古文字句法规则标注集、古文字词类标注集和古文字义项标注集等作为大规模古文字自动识别的基础资源建设,现在也应根据研究现状投入建设。

4. 理论研究的深入

大批考古文字资料的发现使古文字理论研究,尤其是古文字结构字形研究取得了很大进展,出版了一批高质量的研究论著。

甲骨文字的考释和断代研究等有新的收获,一批甲骨学史研究的论文和著作相继问世。对于甲骨文结构字形特征的研究,张桂光先生的《甲骨文形符系统特征的探讨》等论著具有很高的参考价值;有铭铜器不断出土使金文研究有了重要的发展,如围绕陕西眉县杨家村发现的窖藏青铜器及其铭文就发表不少高水平的论文。在收集新出金文、编纂金文索引和古文字信息化处理方面也有许多值得重视的新成果,张再兴先生的《西周金文文字系统论》等论著很具有代表性;战国秦汉文字研究方面成就非常突出,随着新出土的战国文字资料(尤其是楚简)的公布,在学术界形成了战国文字研究热,这几年也出版了一大批研究校读战国文字资料的论著和一批反映战国秦汉文字研究成果的文字编。古文字研究的繁荣,一方面主要是由于新资料的不断发现而影响巨大,新出古文字资料的整理研究取得很大成就;另一方面,世纪之交对百年来古文字研究的回顾和反思也促使一批综合性的资料整理研究和学术史专题研究取得比较突出的成果。

从大规模古文字自动识别的角度看,这种繁荣的背后也存在着许多值得关注的问题。如在甲骨文和金文研究方面,疑难文字的考释工作进展不大;在古文字构形研究方面,一些关系汉字发展演变和构形规律的重要现象还缺乏系统全面的研究;最重要的是,用于信息处理的古文字构形研究尚未开展。因此应鼓励和支持一些学者在这个领域开展长期而艰苦细致的研究工作。

二、古文字自动识别过程

汉字识别是指用计算机提取汉字特征,使其与机器中预先存放的特征集匹配判别,将汉字自动转换成某种代码(例如国际区位码)的一种技术。这里,汉字特征可以是其构形特征,也可以是语义特征,也可以是语法特征,甚至可以是好几种特征的集合。由此可见,对于现在业已人为考释识别的古文字的研究尤其重要,对这些古文字的特征加以概括总结使之成为未考释的待识别的古文字的特征集。同时,古文字通常是以拓片的形式加以识别的,因此,古文字识别同时又是一个图像识别问题。图像识别是对处理后的图像进行分析,在分割的基础上选择需要提取的特征,并对某些参数进行测量,以便对这些特征进行匹配归类。在这里,图像特征可以是形状特征,也可以是纹理特征,对古文字拓片图像而言,则形状特征主要是文字的结构特征,纹理特征则主要是灰度和点阵特征。由于目前人为考释识别出的古文字数量有限且没有进行专门的纹理特征分析,因此图像识别主要还是依赖结构特征。综上所述,古文字自动识别既涉及图像识别,又涉及汉字识别,是两者的综合。一般来说,古汉字自动识别应包括以下步骤:(1)进入 OCR 系统,获取识别图像;(2)对文字图像做预处理,获取文字轮廓;(3)对文字进行切割,并提取其结构特征;(4)单字识别;(5)自动识别后处理;(6)必要的人工文字校对,识别结束。一般来说,对于识别结果正确的古文字要归纳入库,再次提取其特征,为其他待识别的古文字提供特征模板。

步骤(1):这是文字图像获取的过程。图像获取就是图像的数字化过程,也是将图像采集到计算机中的过程,它主要涉及成像及模数转换技术。目前图像获取设备主要有黑白摄像机、彩色摄像机、扫描仪、数码相机等,此外,显微摄像设备、红外摄像机、高速摄像机、胶片扫描器等专用设备也被用来获取图像。对于古文字图像获取而言,常用的设备就可以达到获取目的了。目前常用的 OCR 系统是扫描输入系统,文字拓片经扫描仪转化成为图像文件。有些发掘出的带有古文字的文物不宜制作拓片,我们可以采用照相或摄像的形式获取文字图像。

步骤(2):这是图像预处理的过程。我们获取的图像一般为带灰度值的数字信号,在预处理过程中应将256色灰度图处理成二值(0,1)的。图像获取时,在光电扫描过程中纸张的质量、油墨的质量都会产生污点、飞白、断笔、交连等干扰,称之为"噪声",在预处理过程中要将这些噪声去除。去除噪声后的二值化文字点阵逐层剥去边缘上的点,变成比画宽度只有一个字节的文字骨架图形,这叫做图像的细化。基于细化的识别方法以比画骨架为基础,将待识别的字符的笔段细化为骨架后再作进一步的分析和处理。通常在特征提取之前,还要对文字图像进行规范化。所谓规范化,就是把文字尺寸变换成统一大小,纠正文字位置(平移),文字比画粗细变换等文字图形的规格化处理。文字图像的细化和规范化统称为"抽取轮廓"。图像预处理包括二值化、平滑化(去噪声)和抽取轮廓等。

步骤(3):这是特征提取过程。笔段特征和笔画特征是汉字结构的本质字形特征,汉字笔画形态多变,因此在汉字识别时通常将每一笔画分成形态基本不变的若干笔段。笔段特征具有明显的直观性、紧凑性和普遍性,有利于建立通用性较强的文字识别系统,但它对于噪声比较敏感而不易提取;笔画特征具有明显的抗干扰性、方向性和普遍性,此外,由于任何一个汉字的每一笔画间都存在着一个相对的位置关系,我们可以利用笔画间的特征信息来实现整字的识别。笔画特征的难点在于笔画分类的粗细以及笔画特征点(端点、折点、歧点和交点)的判定。无论是笔段特征提取还是笔画特征提取,都要将二值化后的文字图像分解,这叫"文字切割"。

步骤(4):这是文字识别过程。文字识别方法可以分为统计模式识别、结构模式识别以及人工神经网络方法等。文字识别是一个复杂的过程,任何一种单一的方法都不能取得令人满意的结果。人工神经网络方法是新近发展起来的一种模式识别方法,但如果将其应用于特征复杂的文字识别,网络规模将非常巨大,结构会很复杂,很难实用化。目前常用的是统计模式识别和结构模式识别相结合的决策树识别方法。文字识别实际上是文字特征分类的问题,决策树实际上就是多级分类器,决策树的每个非终节点都包含一个独立的单级分类器。也就是说,决策树将复杂的文字结构特征统计分成了一个个独立的识别系统,每个独立的识别系统识别一类文字的结构特征,解决一部分问题,那么,一个决策树各个分支所解决问题的总和就是该决策树所解决的问题,也就是一个或一类汉字的识别结果。

步骤(5):这是识别后处理过程。在进行初级识别后需要进行后处理,以进一步提高识别率。后处理可以是基于词法分析的,也可以是基于语义判断的,还可以是基于语法分析的。因此,对于古汉语,尤其是甲骨文、金文等文本要作详细的分析,总结其词法、语义和句法特征,编纂古汉语词典、句典等并将其形式化,为古文字信息处理服务。

步骤(6):这是人工校对过程。任何一个古文字自动识别系统识别的正确率都不会是100%,因此,自动识别结束后要辅以人工校对。其实,就目前研究现状而言,我们并不指望计算机能够准确地自动识别出所有待识别的古文字,只要该系统能够为我们提供几个可供选择的字项及各自概率,然后由专家根据经验来判定,那就足够了。因此,对于古文字自动识别而言,人工校对过程其实才是决策阶段,必不可少。

三、古文字自动识别的程序实现

任何一个文字自动识别系统都是由程序来实现的,而程序的实现又依赖于所需数据库的建设。就古文字自动识别而言,所需的基本数据库包括:(1)古文字字库(字符集);(2)古文字结构特征库;(3)古文字句法规则标注集;(4)古文字词类标注集;(5)古文字义项标注集等。但目前这几个数据库的建设程度参差不齐,数据库的规模也大小不一,因此,要实现大规模古文字自动识别,尚需假以时日,尚需各位有志同仁做些扎扎实实的基础资源建设方面的工作,尚需考古学界、古文字学界、语言学界、词典编纂学界以及计算机学界通力合作,更需有关部门的高度重视、政策鼓励。基于目前的研究现状,本文用Visual C++语言实现了古文字自动识别的部分步骤,具体包括:

(1) 程序名称:detect.cpp。该程序用于获取图像的预处理,包括行的检测,以及旋转角度的检测等。

（2）程序名称：common_gfx. h，common_gfx. cpp。该程序用于获取图像的预处理，包含局部自适应阈值二值化的算法以及相关的滤波函数。

（3）程序名称：segimage. h，segimage. cpp。该程序用于获取图像的预处理，定义并实现了字符间分割的函数。

（4）程序名称：character. h，character. cpp。该程序用于文字特征提取与识别。

以上程序都是在基于假定一切条件和数据库都存在的前提下编写，因此含有调用数据库的函数命令。以上程序都具有良好的移植性和可扩充性，可以用于程序直接调用、嵌套或修改。

四、结　论

把计算机引入古文字考释领域，探讨计算机与古籍整理研究的关系这一课题，已经引起了越来越多的人的关注和兴趣。这是代表人类现代文明的计算机技术与代表人类古代文明的古文字考证的一种奇妙结合，这种结合始于上世纪80年代初期，到了80年代中期就随着中文信息处理系统走向实用化和计算机的普及而呈现出不断扩展其广度和深度的态势。大规模古文字自动识别是其典型的趋势之一，在这一趋势出现之初，很有必要对该课题的宏观研究和管理，以避免大规模的重复劳动。

主要参考文献

[1] 朱敬国．手写甲骨文在线识别的模糊数学模型[A]．第二届中文信息处理国际会议论文集[C]．北京：清华大学出版社，1987.

[2] 张普．汉语信息处理研究[M]．北京：北京语言学院出版社，1992.

[3] 计算机科学技术百科全书(选编本)[M]．北京：清华大学出版社，2002.

[4] 康才畯，江荻，戴亚平．一种基于构件的藏文识别算法[A]．中国民族语言工程研究新进展[C]．北京：社会科学文献出版社，2005.

[5] 沈康年．《古文字诂林》数据库[A]．辞书与数字化研究[C]．上海：上海辞书出版社，2005.

[6] 张桂光．甲骨文形符系统特征的探讨[A]．古文字论集[C]．北京：中华书局，2004.

[7] 张再兴．西周金文文字系统论[M]．上海：华东师范大学出版社，2004.

[8] 苏彦华等．Visual C++数字图像识别技术典型案例[M]．北京：人民邮电出版社，2004.

[9] 黄德宽．从转型到建构：世纪之交的汉字研究与汉语文字学[J]．语言文字应用．2005(3)．