中文电子出版系统的硬件和软件

一、引　言

早期的计算机系统主要用于科学计算，而今天计算机系统处理的数据中有80%为文字信息。近20年来，文字和图形处理技术得到了迅速发展，现在有灰度层次的照片信息正在迅速成为计算机系统中的另一重要数据类型。汉字是象形文字，有一个很大的字符集，这一特点给研制先进的中文电子出版系统带来一系列的技术难题。电子出版系统将根本改变中国的印刷工作，而中国有大约20000家印刷厂、3 000家报社，以及一大批企事业单位，潜在的市场起码达几十亿元。技术难关和市场需求恰恰是引起新的突破的推动力。

二、汉字字形在计算机中的表示形式

汉字与以拉丁字母为基础的西文不同，包含一个很大的字符集，比较常用的就有近7 000字。报纸杂志的版面丰富多采，至少需要40种不同的字体和20种不同的字号。精密照排系统要求高分辨率，一个10磅的正文小字至少需用100×100点表示，而一个63磅的报纸大标题字将用600×600点表示。假如我们把所有这些不同字体、不同字号的汉字以高分辨率点阵形式存入计算机，将占大约30亿字节的存储空间。所以，汉字字形存储量大的问题是发展先进的中文电子出版系统的主要障碍。近15年来出现了多种描述汉字字形的方案，本文将分析这些方案的优缺点。

1. 全点阵方案

汉字字形直接用点阵描述。同一汉字的不同字号将由不同大小的点阵来表示，因而将占不同的存储区域。对于每英寸300线（300DPI）的低分辨率打印系统，若容纳4种字体、7种字号（7磅到21磅），每种字体含6 763个汉字，则字形所需存储量就达84M字节。

全点阵方案的优点是：

（1）完全没有失真；假如对每种大小的点阵字形都作精心的修改，则可以达到最高质量。

（2）易于实现。现在至少有一半以上的中文电子印刷系统采用点阵表示字形的方案。

此方案的主要缺点是：

（1）汉字字形所占的硬盘空间太大，从而留给用户文件的空间太小，必然造成使用不便。应该强调，上面所述的4种字体和7种字号是远远不够的；另外，300DPI只是低分辨率，不能满足出版要求。

（2）从硬盘上读字形点阵的速度太慢，由于频繁访盘读字形点阵，使一页版面点阵的形成时间大大增加，结果使原来能一分钟输出8页的激光打印机（例如佳能LBP-SX），实际上降为一分钟或几分钟才输出一页。

可以采取下述措施以便在一定程度上缓解上述矛盾：

（1）把常用的汉字字形点阵放入掩膜ROM。

（2）在控制器的RAM中增加一个“事先加载”区（download area），把最常用的字形点阵事先送入该区，这样避免了用到时再读盘。

（3）只存一种或两种字号的点阵，然后用软件变倍的方法得到其他字号的点阵；当对点阵进行放大时，可采用若干平滑算法以改善边缘的锯齿。

（4）选用高速CPU，高速硬盘，在主机和控制器之间选用DMA，而不是PIO接口。

全点阵方案的致命弱点是不灵活，很难从一种表示获得各种变形。显然，此方案非常不适合精密照排系统。

2. 黑白段描述法

60年代末，德国Hell公司的Digiset CRT照排机上首先采用黑白段描述方法。此后

不少照排系统沿用这一方法。黑白段描述法只记录字形水平方向上每条线的各个黑段和白段的长度，而不存储整个点阵。字形点阵可以毫不失真地复原，但有下述缺点：

(1) 压缩倍数很低。存储量只减少一半或一半多一点，仍不得不用大硬盘来存放汉字字形，而硬盘的存取时间限制了向输出设备提供字形点阵的速度。

(2) 不易产生变形字，也不易对字形作旋转等操作，而这些是报纸、杂志、地图和广告编排中所渴望的。

3. 华光电子出版系统中使用的轮廓向量和笔锋描述法

1976 年，我们提出了一种描述汉字字形和进行文字变倍的新方法。我们把汉字笔划区分成规则笔划和曲线形式的不规则笔划；前者指横、竖、折三类笔划，据统计，占汉字笔划的一半以上。规则笔划由直线段、起笔笔锋、收笔笔锋和转折笔锋等组成，这些笔锋是规则笔划中主要的变化部分，但这些笔锋的形状变化都是很有限的。我们可以把这些有限数量的笔锋的形状事先存入计算机中。这样，只需少量信息就可以表示规则笔划的始点 x，y 坐标、长度、宽度、倾斜度和各种起笔、收笔和转折笔锋；对于不规则笔划，则用向量信息描述笔划的轮廓。

因为不规则笔划表示成向量形式，而规则笔划的压缩信息也可转换成一连串向量信息，因而只要对向量串中的每个向量进行对应的变倍，就很容易对字形作放大或缩小。这样，可以从一种基本字号得到不同的字号。但是为了保证变倍后的文字质量，需采取一些重要措施。我们用不同的变倍方法来处理下面两种情况：

(a) 对于规则笔划的宽度

变倍后的宽＝［变倍前的宽×变倍值］

这里变倍值＝需要的字号/基本字号

〔A〕表示 A 的整数部分；

(b) 对于除规则笔划的宽度部分外的任何向量

变倍后的向量的 Δx＝［变倍前向量终点 x 坐标×变倍值］－［变倍前向量始点 x 坐标×变倍值］

变倍后的向量的 Δy＝［变倍前向量终点 y 坐标×变倍值］－［变倍前向量始点 y 坐标×变倍值］

这里〔A〕表示 A 的整数部分。

实质上，方法 (b) 先对每个节点变倍，然后用减法得到变倍后的 Δx，Δy；而方法 (a) 直接对向量的长度进行变倍。方法 (b) 需要更多的计算，但舍入误差不会积累。

假如方法 (a) 用于对不规则笔划的一串向量进行变倍，则舍入误差的积累将引起不可容忍的失真；假如方法 (b) 用于对规则笔划的宽度部分进行变倍，则变倍后的规则笔划的宽度将无法精确控制。例如，一个宋体 10.5 磅汉字的两笔横在 96×96 基上具有同样的两格宽，现需要缩小到 9 磅小字，由于 y 位置不同，采用方法 (b) 变倍后的两笔横可能具有不同的宽度：其中一笔变成一格；而另一笔仍保持两格。这是不能容忍的。但若用方法 (a) 对三类规则笔划的宽部分作变倍，则无此问题。

区分规则笔划和不规则笔划不仅能提高压缩倍数，更重要的是能保证变倍后的文字质量。

本方案的压缩倍数很高。4 种基本字体，21 种不同字号，7000 汉字仅占 3.2MB 存储空间。

4. Postscript 曲线轮廓描述法

上述向量轮廓描述具有一定局限性，当变倍时，特别是字形变得很大时，边缘将表现出折线的痕迹而不是真正的曲线。近年来版面描述语言 Postscript 逐渐被普遍接受，正在成为国际标准，在 Postscript 中，折线、圆弧和三次曲线相结合描述字形的轮廓；而且字符和图形采用同样的描述方法。这是描述大的字形的最好方法。

把三次曲线表示的轮廓转换成点阵是一件很费时间的操作。对于西文系统，由于当前使用的字体、字号不多，西文字母又只有几十个，所以只需转换一次，然后把点阵在高速 RAM 中暂存，以后就可以反复使用。由于汉字字数太多，这样做是不现实的。由此可见，中文系统对字形转换速度要求是很高的。

三、把向量轮廓转换成点阵的算法和设备

1976 年我们设计了一套快速算法，用来把华光系统中采用的字形压缩信息高速复原成高分辨率点阵，该算法后来成为欧洲专利 EP0095536 的主要内容之一。复原步骤如下：

(a) 把十分紧凑的规则笔划压缩信息以及多种形式的不规则笔划轮廓向量信息转换成统一形式的向量，用带符号的 Δx，Δy 对表示。

(b) 把统一形式的向量串转换成标记点阵，对于每个点，标记点阵中有两位标记指示该点的下述四种情况之一：

(1) 黑段的开始 (B 标记)；

(2) 黑段的结束 (E 标记)；

(3) 孤立黑点 (I 标记)；

(4) 空白 (V 标记)。

(c) 把标记点阵转换成最终点阵。

在复原步骤 b 中，我们设计了只用加、减和比较操作，易于硬件实现的算法，能用较少拍节把向量转换成一串最邻近的阶梯点；还设计了一套算法，能在笔划交错等各种复杂情况下正确地写入标记点阵的两位标记。在复原步骤 c 中，我们设计了一拍内完成多个点的转换算法。

在常规计算机上，用软件实现上述算法，速度是很慢的。采用最新的 RISC 或 CISC 芯片，能显著提高速度，但仍不能完全满足精密照排系统对字形产生速度的要求。对于中文照排系统而言，字形复原速度是一尖锐问题。在华光Ⅳ型中，我们采用微处理器与专用超大规模集成电路 (VLSI ASIC) 相结合的做法。专用超大规模集成电路是专门针对上述复原步骤 b 和 c 设计的，内含一系列的寄存器、计数器、多路转换器、复杂的译码电路以及各种控制信号的产生电路，等价于约 400 片常规的中、小规模集成电路。专用芯片与微处理器并行工作，专用芯片内的操作也高度平行。这样，一拍内可完成常规计算机上十几拍甚至几十拍才能完成的操作。结果，汉字字形复原速度高达 710 字/秒 (对于 100×100 点阵)。设计专用超大规模集成电路是一件困难的任务，任何细小的错误都

会造成芯片掩膜的返工。完全逼真的模拟有时很难实现，我们在设计过程中，采用了若干推理的方法，尽量事先验证设计的正确性，结果，没有做实验，也没有做模拟，一次成功地完成了芯片的研制。

基于超大规模集成专用芯片的华光Ⅳ型光栅图象处理器（RIP）是系统的核心硬件，其主要功能如下：

（1）解释版面描述语言；

（2）把字形压缩信息复原成点阵；

（3）对字形作变倍，这里 X 和 Y 方向的变倍值可不同，从而产生长字和扁字；

（4）对字形作倾斜和旋转处理；

（5）产生各种变形字和作字形修饰，例如空心字、灰色字、阴影字、边框字等；

（6）处理花边、图形和照片等；

（7）产生一页点阵并送入输出缓冲供扫描用；

（8）控制输出设备的动作。

四、华光系统的排版软件

系统支持两种风格的排版软件：批处理（或称基于源语言）方式和交互式（或称直接操作）方式，并使两种风格的软件能融合在一起，互相补充。排版软件由下列子系统组成。

1. 批处理排版软件

华光批处理排版软件具有下述特点：

（1）语言简单。

（2）功能强。

（3）自动化程度高，能自动产生复杂版面。例如，只需指出积分的范围、积分的上、下限，软件就能自动产生美观规范的积分；也即用户只需指出“排什么”，而不必指示“如何排”。

（4）完整的语法检查功能，能迅速和准确地指出写错的注解。

（5）能在屏幕上显示清样，也能把一页的清样在针式打印机上输出。屏幕显示、针打、激光打印机和精密照排机四者的输出版样完全相同。

华光批处理排版软件在功能上具有下述特点：

（1）能方便地排数学公式。

- 能自动地正确安排公式与周围文字的位置关系；
- 能方便地排多层复杂的分式；
- 能方便地排各种积分式；
- 能排各种方程式，并能将方程号自动对位；
- 处理各种大算符，并能将其上下附加内容自动居中和对齐；
- 能排各种复杂的行列式和矩阵；
- 能自动排多层复杂的根式；

·能自动排多层复杂的角标；

·方程式自动在最佳点拆行；

·能排阿克生码，及处理各种科技符号；

·方程组自动拆页，并可自动将方程组大括号拆开；

·能自动上下附加符号。

(2) 能排化学式，包括各种反应式、化学键、苯环和原子等。

在数学式和化学式排版方面做到全书统一、格式规范、绝无参差不齐、前后不同、大小不一等弊病。

(3) 用菜单能一次任意改变整本书的版心大小、正文字体字号、正文行距、页码、书眉及标题字体字号、格式；即一次录入的内容，可按不同要求方便地排出各种开本和格式的版面。

(4) 中外文混排，英文自动分音节，加连字符。

(5) 自动将随文注内容排在页末，并在一次排版时允许多种随文注符格式。

(6) 自动填写全书目录中的页码。

(7) 排字典时，自动抽取词条作书眉。

(8) 能全书竖排，并自动排竖排书眉、页码以及自动排竖排随文注；能在同一版面上横竖混排；只要改变菜单选择就能将录入内容由横排改竖排或竖排改为横排；改竖排时自动将标点、括号、英文、数字、着重点和线以及脚注符等转换为竖排格式。

(9) 能分多栏排、等距分栏或不等距分栏，并能自动换栏及拉平；能多栏对照式排版。

(10) 很强的表格排版功能。

·能排各种跨栏或不跨栏的无线表；

·能排各种复杂的表格，包括有斜线项或项中又套表格等十分复杂的有线表格；

·无论有线或无线表格，甚至是十分复杂的有线表都能自动拆页；

·有线表格拆页时能按需要自动重复排表头。

(11) 能接受其他排版软件的排版结果，自动安插在用户认为合适的位置上。

(12) 有定义宏注解功能。

批处理注解式语言的缺点是不够直观，不能边输入边看结果。因此对于需要及时看到版面的出版物采用交互式排版，如报纸、复杂表格、线路图、化学式等，以补充注解式排版的不足。华光系统有一个统一的版面描述语言，称为华光 PDL，各个独立开发的排版软件的输出文件都用华光 PDL 的格式。有一个华光 PDL 的解释程序，用来把不同软件编排的结果汇总在一页上，实现了注解式和交互式排版的结合，从而达到使用方便、生产效率提高的目的。

2. 报纸组版软件 NPM

NPM（Newspaper Page Makeup）系统是一交互式软件，专门用于满足报纸编排的要求。采用超高分辨率竖形大屏幕显示器，操作人员可以方便地划分报纸版面，灌入文字，修改正文和标题；可随意增加和移动花边和网纹；可作文图合一的编排和输出；大屏幕上显示的是多字体、多字号、带花边和网纹、文图合一的清样，所有的操作均直接

针对这一逼真的清样进行。

3. 交互式图形排版软件 HD（基于 AUTOCAD）
4. 交互式表格和流程图排版软件
5. 交互式乐谱排版软件
6. 交互式国际象棋、象棋、围棋和扑克牌排版软件
7. 图片扫描和编辑软件
8. 多窗口集成排版系统

五、多窗口环境下的集成组版软件 WITS

WITS（Window-based Integrated Typesetting System）是一个集成的组版软件系统，它由一系列用于组版的工具组合而成，其中包括一个组版主系统，多个专用组版工具（如数学、化学、表格、乐谱、棋牌等）以及一些其他辅助工具（如字模工具及第三方开发的工具等）。WITS 中的工具是在一个多窗口的环境下，按照统一的内部共享数据结构以及接口协议开发完成的，这些工具可以在多窗口的环境下同时工作，互相通信，从而有效地构成了一个集成的组版系统。

WITS 系统具有以下特点：

（1）高度集成，易于维护与扩充

系统中所有的工具都是建立在一个多窗口的环境上，它们之间不仅可以通过文件传输数据，还可以利用环境的支持进行内部信息的共享与传递。此外，系统采用的是开放式设计，各个工具都是在一个统一的环境下，按照统一的协议独立开发而成的。因此，降低了整个系统的复杂性，易于维护，并可在不影响原有工具的前提下，不断增加新的或引入第三方开发的工具以进一步完善系统。

（2）强大的组版功能

WITS 系统不仅具有一个功能强大，可排各种复杂版面的组版主系统，还包括一系列支持各专项组版的工具，因此，它可以完成组版阶段的各项任务以及任意复杂版面的组版。

（3）界面统一友好，使用方便灵活

系统中的工具全部采用统一的、国际流行的窗口界面风格，并辅之以 WYSIWYG 的显示及联机 HELP 的功能，使得整个系统风格统一，操作直观，界面美观友好，易于掌握与操作。

（4）排版格式规范

系统中的工具全部以专业出版的质量作为设计标准，并在排版时以自动排版为主，对版面格式实行自动控制，保证不同人的组版结果，其格式是一致的。此外，在很多工具中都提供了尺子、背景格以及捕捉（snap）等辅助功能，使得排版格式更加准确。

（5）自动保持文章格式与内容的一致性

保持文章格式与内容的一致性是指在修改文章时，不能使原有的格式变形或内容丢失。WITS 作为一个交互的系统，其组版工具能在文章修改时自动调节相关的内容或格式（包括在不同的页），以保证文章的一致性。

(6) 适用范围广

WITS 系统不仅可应用于专业出版，而且也适用于办公室自动化应用。就组版对象而言，它可以用于报纸、杂志、书刊等多种出版物的组版工作。此外，WITS 系统还对多种少数民族文字的处理提供了潜在的支持，使得系统能够很容易地用于少数民族出版物的组版。

WITS 系统是一个符合国际电子出版系统发展趋势的新一代组版系统，它的开发完成为今后进一步的发展打下了良好的基础。