**A Prototype Document Image Analysis System for Technical Journals**

**面向科技期刊的文档图像分析原型系统**

**George Nagy, Rensselaer Polytechnic Institute Sharad Seth, University of Nebraska at Lincoln Mahesh Viswanathan, IBM**

快速计算电子数据存储和访问的技术期刊标准图书馆的要求。一个中型研究图书馆订阅约2000种期刊，每种期刊平均每卷约500页，每年总计100万页。虽然这篇文章是由照排机以1，270 dpi(每英寸点数)输出到胶片上的，但在300 dpi的激光打印机或显示器上复制对大多数读者来说是勉强可以接受的(至少对文本和一些艺术来说是这样)。在300 dpi下，每页包含大约600万像素(图片元素)。在10:1的保守压缩比下(使用现有的传真方法)，整个期刊集每年产生80千兆字节。虽然这个卷远远超出了单个工作站的存储能力，但对于库文件服务器来说是可以接受的。(当然，即使没有压缩，未格式化的文本每页也只需要大约6千字节，但它不是技术材料的可接受载体。)

一篇10页的文章可以通过高速网络传输，并以图像形式打印或显示，比步行到附近的图书馆花费的时间少得多。此外，虽然大多数研究图书馆都有计算机，但你可能要等几天才能通过传真或快递获得馆际互借。因此，有充分的动机来开发数字化技术材料的电子分发系统。

然而，即使材料是以数字图像形式提供的，也不是每个人都能方便地使用高速线路、激光打印机或2,100 x 3,000点显示器。我们展示了智能文档分割如何将电子浏览带到仅配备调制解调器和个人计算机的读者的范围内。

对于当前基于知识的图像处理算法的研究，文档分析构成了一个具有适当难度的领域。 这很重要，因为文档分析本身是一种临时应用程序：毫无疑问，最终信息生产者和消费者必须以数字方式联系起来。 但是，识别页面上重要信息块的位置和范围也有实际优势。 这也适用于分割。

智能文档分割可以在大多数用户的范围内带来电子浏览。作者展示了如何通过文档处理，分析和解析图形句子来实现这一目标

**Glossary 词汇表**

**与或图(或树)**。一种解决方案策略的表示，其中从开始节点到解决方案节点的路径需要遍历或节点的任何分支和与节点的每个分支。在两人游戏中使用的相关最小-最大搜索中，从开始节点到解决方案节点的路径在最小节点取最低成本分支，在最大节点取最高成本分支。

**Bitmap位图**。将点映射到二进制像素数组的图像的数字表示。

**Branch-and-bound分界线**。 一种搜索技术，可避免某些路径导致比目前获得的最佳解决方案成本更高的解决方案。

**CClTT** (Comite Consultatif International de Telegraphie et Telephonie)。 颁布传真编码和传输标准的国际组织。 CCITT Group 3 压缩方法用于误码率相对较高的信道，例如电话网络。 第 4 组标准专为公共数据网络等低错误率信道而设计。 代码的压缩率是图像的位图表示中的位数除以编码表示中的位数。

**Compiler tools编译器工具**。 为特定应用程序生成词法和句法分析程序的程序，包括计算机程序的汇编、解释、编译和翻译。 在 Unix 系统中，Lex 是一种流行的词法分析工具，而 YACC（又一种编译器编译器）用于解析上下文无关语言。

**Document interchange formats文档交换格式**。编码组织文档的布局和内容组织的标准，以便将文档从一台计算机系统转移到另一个计算机。流行的例子包括文档内容AR - Chitecture（DCA），文档交换格式（DIF），文档样式语义规范语言（DSSSL），Office Document Architecture（ODA）和标准通则的标记语言（SGML）

**Drop cap掉线**。超大的大写字符跨越多条印刷线，通常用于开始文章或部分。

超文本。文本或其他媒体的数据结构，包括概念相关项目之间的链接（指针）。这些联系允许相当大的纬度选择遍历许多信息的策略。链接是预设的或由数据用户创建的。

**Kerning**。通过删除它们之间的一些空间，单独的字符比常用放在一起。

**Layout布局**。与内容相比，文档中的物理（例如，几何和印刷）组织文本和图形物质组织

**Leading**。 文本连续行之间的分隔，最初是通过在文字行之间插入铅条来实现的。

**Lexical analysis词法分析**。 每个标记（单词）都属于一个词汇类别，该类别决定了哪些标记可以在它之前或之后。 在英语中，词汇类别的例子是动词、名词和代词。 词法分析确定标记的类别。

**Ligature**。 两个或多个符号，例如 ，打印为单个图案。

**OCR**（光学字符识别）。 将印刷或书面材料转换为计算机可读 (ASCII) 代码的技术。 oprical 一词用于将其与磁性墨水字符识别 (MICR) 区分开来。 OCR 系统通常包括光学扫描仪、按适当顺序定位字符的预处理器、模式分类器和上下文后处理器。

**Page icon**页面图标。 将打印页面细分为与 重要布局单元 相对应的 嵌套矩形的缩小比例表示。 它是分解的 X-Y 树的图形显示。

**Point点**。 打印机的测量单位，大约等于 1/72 英寸（美国、欧洲和比利时点略有不同）。 1 个pica（十二点活字)等于12 points。 符号 10/12 表示 10 点类型，带 2 点leading（空白）。

**PostScript**。 由 Adobe Systems 引入的一种编程语言，允许对计算机打印机、显示器和数码照排机的排版文本和插图进行独立于设备的规范。 这种语言通常被称为页面描述语言。

**句法**。 与语法同义：确定词汇标记（例如单词）的允许配置的正式规则。 解析（句法分析的一种形式）确定一串单词是否构成一个合法的句子。 语法所接受的语言是（通常是无限的）所有合法句子的集合。 文法通常表示为一组重写规则或产生式，允许将产生式左侧的符号串替换为右侧的字符串以生成合法的句子 . 不会在左侧单独出现的符号，因此不能被替换的符号称为终结符； 所有其他符号都称为非终结符。 形式上，文法是由起始符号、定界符、一组终结符号、一组非终结符号和一组产生式组成的五元组。

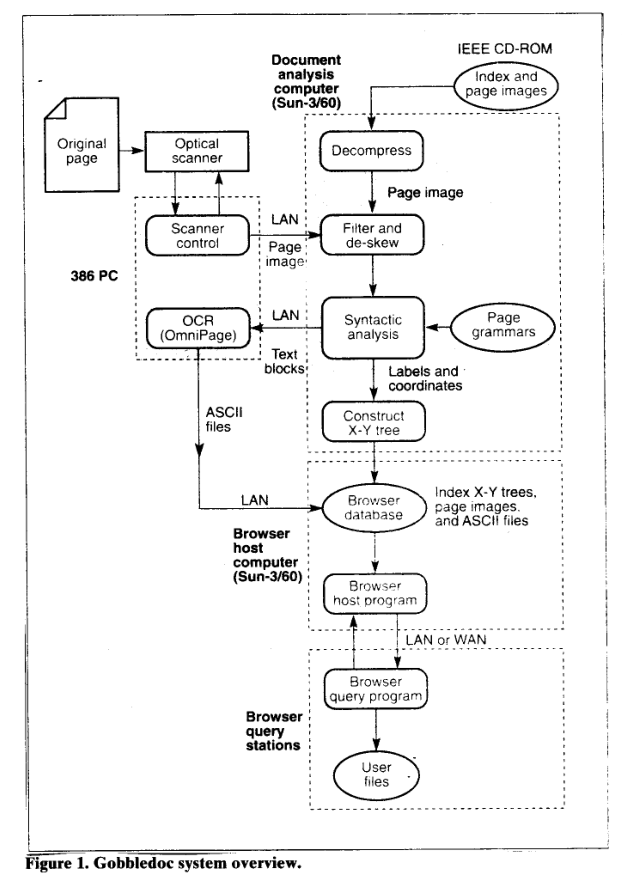
**TIFF**（标记图像文件格式）。 压缩或未压缩形式的彩色、灰度和黑白图像的一系列流行格式。

**VGA**（视频图形适配器）。 IBM 于 1987 年推出的 480 x 640 像素彩色屏幕显示器的格式约定。

**X-Y 树**。 每个节点对应一个矩形块的空间数据结构。 每个节点的子节点表示父块在特定（水平或垂直）方向上的细分位置。 在 VLS1 设计中通常通过极坐标图表示相同的组织。 用于等距（曼哈顿）空间细分的其他流行的分层数据结构包括 K-D 树和四叉树。

**智能页面分割的优势**

* 可以在 PC 显示器上再现较小的块，而不会发生水平滚动或丢失易读性
* 选定的块可以比通过网络浏览的整个页面更快地传输
* 布局分析有助于 OCR 保留阅读顺序
* OCR 难以处理的选定块的密钥输入比重新加密整个页面成本更低
* 将文本与图形和半色调区分开来可以提高图像压缩效率。
* 高对比度和半色调区域之间的直线边界被保留用于数字复印



根据其逻辑结构标记页面图像，而无需借助光学字符识别（请参阅“智能页面分割的优势”侧栏）。

OCR 的研究始于世纪之交。 1914 年首次展示了替代电报员和辅助盲人阅读器的发明，自 1955 年以来，高速 OCR 系统已商业化用于特殊印刷表格和固定间距打字稿。 数字化页面分析领域始于 60 年代，但直到最近相关技术才足够成熟，可以转换复杂的排版文档，例如技术文章。 主要的催化剂包括准确、快速、廉价的页面扫描仪； 高速计算机网络； 为数字化页面图像提供足够的存储和处理能力； 和数字传真的压缩标准。

文档图像分析需要许多软件工具的组合。 Gobbledoc 是我们在过去五年在伦斯勒理工学院开发的系统，它将多个现有组件与一种新的图像分解方法集成在一起。 图 1 显示了从原始文档到用户计算机屏幕的信息流。 页面要么在本地扫描，要么从 CD-ROM 中获取。 它使用句法方法进行分割和标记，并存储在 X-Y 树中。 文本块的图像被传输到个人计算机并使用 OCR 进行识别。 产生的 ASCII 输出（包括一些用转义引号标记的特殊字符）链接到 X-Y 树的相应节点。 然后所有数据都存储在浏览器主机上。 浏览器从远程客户端工作站激活，用户请求的信息显示在客户端终端上。

在 Gobbledoc 中，图像处理、文档分析和 OCR 操作在获取文档时以批处理模式进行。 在解释了文档图像获取过程之后，我们描述了必须输入系统以处理一系列页面图像的知识库。 我们展示了我们的 X-Y 树数据结构如何将二维页面分割问题转换为一系列可以使用传统编译器工具解决的一维字符串解析问题。 接下来，使用句法分析将每个页面划分为标记的矩形块。 标记文本的块由 OCR 转换以获得辅助 (ASCII) 文档表示。 但是这样的符号文件更适合计算机搜索而不是人工访问文档内容。 由于在 OCR 过程中丢失了太多的视觉布局线索（包括一些特殊字符），我们的系统保留了原始块图像以供人类浏览。 最后，我们考虑特定于文档图像的存储、网络和显示问题，并描述我们的原型浏览器。

**预处理**

我们通过以 300 dpi 扫描每页将其转换为数字形式。 该速率足以保留所有重要的空白区域。 由于整个 X-Y 树方法对歪斜非常敏感，我们仔细对齐扫描仪上的每一页。 我们认识到，这在生产环境中是不可能的，因为系统可以采用其他人已经证明的一种出色的歪斜校正方法（由软件或硬件执行）。

**示例页面**。 我们还从 University Microfilms 为 IEE-IEEE 准备的 CD-ROM 数据库中获取样本页面。 IEEE 每年分发大约 130,000 页的 300 dpi 图像形式的 CD。 其他专业协会也不甘落后。 IEE-IEEE 数据库以 CCITT Group 4 压缩格式存储。 这是二维修改修改读取 (MMR) 代码，作为在极低错误率公共数据网络上运行的传真终端的标准而开发的。目前，我们使用软件解压页面，但可以使用用于此目的的标准芯片。 由于我们无法控制 CD 页面的歪斜程度，因此我们对它们进行去歪斜

**噪音**。 我们的文档包含由库存中的有缺陷的纤维、不完善的印刷、影印和数字化引起的噪音斑点。 （高质量的期刊相对没有噪音，但我们在未涂布的纸上使用影印本。）这种噪音不会打扰人类读者，但会使自动分析复杂化。 为了应对噪音，可以构建足够健壮的页面语法以忽略斑点。 这是可行的，但很乏味。 相反，我们在初步通道中过滤掉所有小于给定尺寸的斑点。 分析完成后，将恢复所有斑点。 因此，大小阈值可以非常大。 在 i 和 j 上丢失几个句点或点不会影响布局分析，并且在任何文档组件提交给 OCR 或显示供人工检查之前，它们会被恢复。

**布局编码。**分析特定文档族的第一步是对区分该族与其他族的信息进行编码。 所需的知识库非常具体。 它可以以 if-then 规则、形式定义语言、框位置的几何树、专家系统或形式语法的形式明确说明——或隐式嵌入到处理程序中。 截至 1991 年 7 月，Computer 中专题文章标题页的一些约束示例显示在随附的侧边栏中。 我们的实验中使用的 IEEE 模式分析和机器智能交易 (IEEE-PAMI) 中的页面规则类似，稍后在图 6 中显示。

特定出版物的系统已经成功地展示在报纸页面、商业信函、文章和和技术期刊目录、专利申请、简历、带有预先指定布局的打印表格、乐谱、地图、工程图、电路图，甚至国际象棋期刊。 已经开发了一个复杂的基于知识的系统，该系统展示了不同组件之间完全交互的有用性，用于邮政地址定位和解释。 这些领域的最新成就在文件分析会议记录中有所描述。

**句法分析**。 正如每种编程语言都有其固定的规则集一样，每种出版物都有预先确定的布局约定。 这些约定规定了与页面上的逻辑实体相对应的块的大小、位置、间距和排序，如计算机的侧边栏所示。 因此，满足假定的布局约定的文档总是可以在没有 OCR 的情况下被解析为不同的组件，就像语法正确的程序代码可以被解析为有意义的构造而无需语义理解一样

句法分析同时完成分割和组件识别。 特定于出版物的文档语法是对给定出版物中的文章可以采用的所有合法页面格式的正式描述。 原则上，对页面的解析将揭示该页面在语法规则下是否可接受，如果是，则必须为各个组件分配哪些标签以解释有效的图形“句子”。

尽管我们使用主要为形式语言开发的编译器工具，但文档信息的句法分析显示了许多分析自然语言的困难。 布局规则可能不足以识别每个文档组件。 例如，带有方程式的文本行可能会从根本上改变预期的行间距。 因此，我们必须确保微小的偏差仅具有局部影响。 此外，现代编程语言的语法是从一开始就建立的，而文档语法必须间接推断，如后面所讨论的。 （一项永无止境的任务：期刊经常换上新面孔。）

**块语法**。 特定期刊的文档语法由一组块语法组成。 每个块语法将一个块水平或垂直细分为一组子块。 因此，应用整个文档语法的最终结果是将页面细分为嵌套的矩形块。 这种细分可以用 X-Y 树 的数据结构来有效地表示（图 2）。 块语法本身也以树的形式组织：用于细分每个块的块语法由上一级解析的结果递归确定。

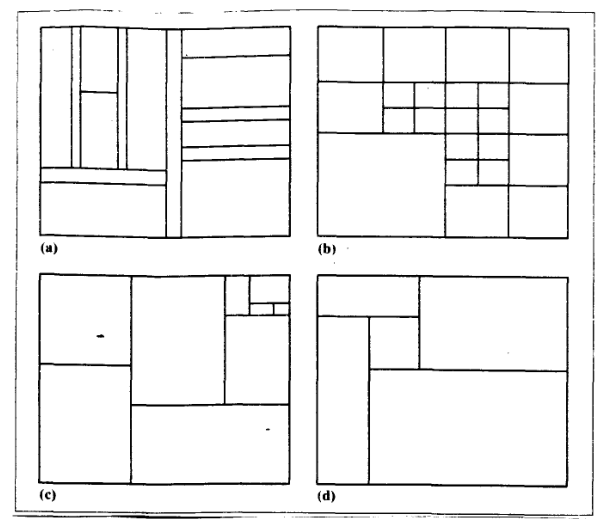


图 2. 将矩形分层细分为矩形：用于页面分割的 X-Y 树（a）； 用于比较或组合多个图像的四叉树（b）； 用于快速搜索的 K-D 树 (c)； 只能通过连续的水平和垂直细分无法获得的矩形块的平铺示例 (d)。

这种方法有效地将困难的二维分割问题转化为一组可管理的一维分割问题。

从理论上讲，句法形式主义很好理解，并且可以使用复杂的软件进行词法分析和符号串的解析。 因此，每个块语法都作为传统的字符串语法来实现，它对一个称为块配置文件的二进制串进行操作。 块轮廓是块内黑色区域的阈值垂直或水平投影。 块轮廓中的零对应贯穿整个块的空白区域，因此是细分位置的良好候选。

用块语法表示整个页面的结构可以大大简化问题。 但是每个块语法本身是一个复杂的结构。 它必须适应许多替代配置。 例如，为了将标题块与署名块分开，块语法必须提供不同数量的标题行和署名，以及由字母的上升和下降引起的间距变化。 为了简化设计过程，每个块语法按照从配置文件特征中提取的句法属性分几个阶段构建。

**句法属性。**

* （水平/垂直）块配置文件是一个二进制字符串，对于每个仅包含白色像素的水平或垂直扫描线，它包含一个零； 否则它是1。
* 黑色原子是最大的全一子串。 它是当前块配置文件的最小不可分割分区。 白色原子是全零子串
* 黑色分子是一系列黑色和白色原子，然后是黑色原子。 白色分子是将两个黑色分子分开的白色原子。
* 实体是已分配类标签（标题、作者、图标题）的分子。 它可能取决于排序关系。

**句法属性**。

块语法的第一阶段对块轮廓的 1 和 0 进行操作。 由 1 或 0 组成的字符串称为原子。 原子根据其长度分为几类。 一串交替的黑白原子是一个分子。 分子的类别取决于它包含的原子的数量和种类。 最后，分子根据它们的出现顺序转化为实体。 选择原子、分子和实体这三个词是因为它们并不特定于特定的出版物或细分。

决定解析的句法属性是实体内原子的大小和数量，以及页面上实体允许出现的实体的数量和顺序。 表 1 显示了包含标题和署名的页面片段的水平轮廓的预期变化。 将符号分配到更大的单元是通过重写规则或产生式来完成的。 这些规则考虑了预期的可变性。

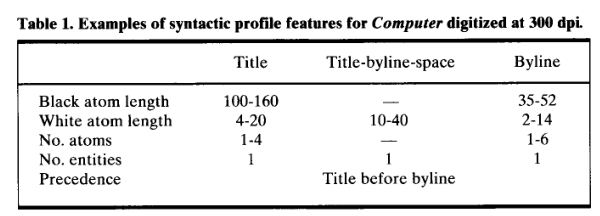
第一行中的条目对应于以 300 dpi 扫描的专题文章中标题和署名行的最大和最小高度（以像素为单位）。 第二行显示了前导的预期轮廓高度。 必须有 1 到 4 行标题行和 1 到 6 行作者行，署名行在标题下方。 每个页面只有一个标题块和一个署名块。（根据已有的数据设定的规则）

在图 3 的大幅简化示例中，顶部显示了典型的水平块轮廓（应该是中间长串的0000111110001111000....）。 原子长度被缩短以显示整个水平块轮廓。

在第 1 阶段，一连串的 1 或 0 根据它们的长度压缩成原子。 例如，2 到 3 次的黑色运行称为 a。 在第 2 阶段，原子被分组为分子。 （由两个 b,v 或 b,w 类型的原子组成的0~3个黑白序列，后跟一个 b 类型的黑色原子，标记为 A。）在第 3 阶段，分子被解释为文档实体。

图片

5为0的像素高度为5，空白行可以为3（v）、4或5（w）个像素高



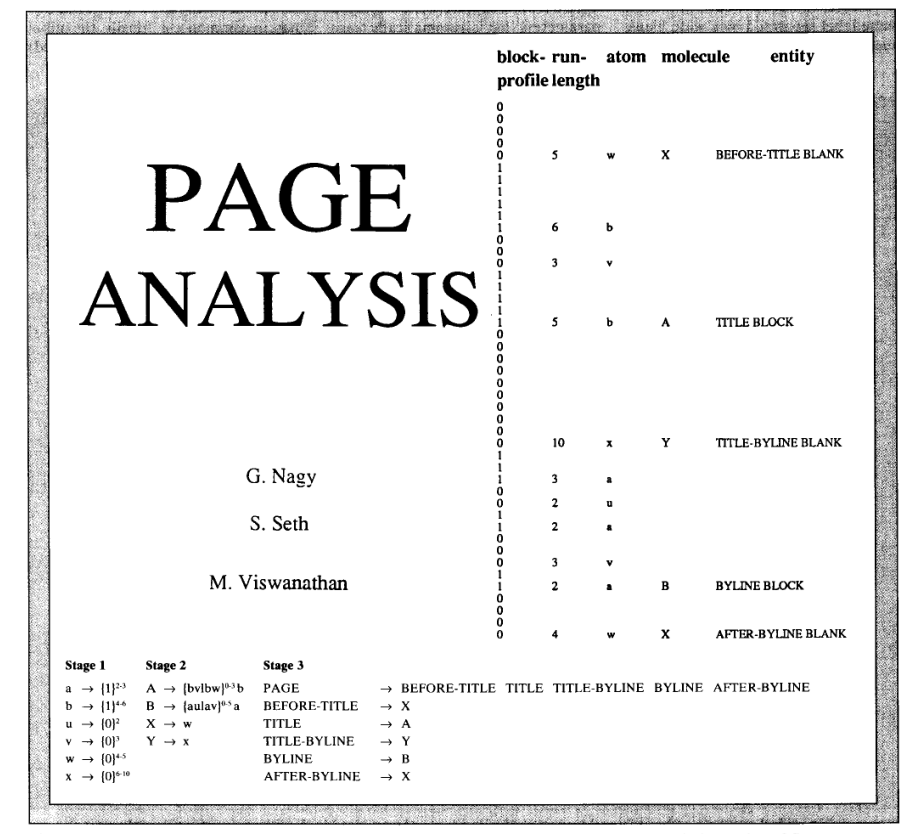


图3.标题页的第一个垂直细分的简化块语法的插图类似于计算机的标题页。

底部的块语法指定标题块（分子 A）可能有 1 到 4 行打印，由空行分隔。 每行打印（原子 b）可以是四到六个像素高，每个空白行可以是3（v）、4或5（w）个像素高。 （由于标题块和作者块中空白行的高度范围重叠，因此将空白行的高度分为两个范围。）此块语法的第 3 阶段确保标题块在署名块之前。

请注意，分子 X 有时被解释为标题前空白，有时被解释为行后空白，这取决于它相对于其他分子的位置。 第 4 阶段（未图示）合并由宽空白分隔具有相同标签的连续实体，例如文本段落。

基于这属性的块语法通常对单个出版物。（如 IEEE-PAMI）有数百个版本，直接在二进制配置文件字符串上定义每个块语法，虽然理论上等效，但在实践中太麻烦了。

**语法规范**。

手工构建块语法仍然非常耗时，并且是我们当前方法的最大弱点。 可用的信息来源是对扫描副本样本的计算机辅助测量、出版商的样式手册以及在页面格式化程序中为给定出版物定制的宏定义。 我们还开发了布局参数的表格表示，可自动转换为块语法。

我们文档布局的表格表示的有吸引力的替代方案包括 Office 文档架构 (ODA)、标准通用标记语言 (SGML) 和文档样式语义规范语言 (DSSSL) 文档间变换标准。 然而，从这些规范中提取出我们分解所需的交替水平和垂直细分的标准并不容易。

与高级语法相比，低级块语法往往更通用（表现出较少的依赖于发布的可变性）。 因此，可以为一系列新的出版物重用现有的块语法。 例如，只有几种常用的段落格式。 理想情况下，初始参数将在实际操作中进行调整（学习）； 这是一个持续研究的主题。

**通用排版规则**

* 打印的线条大致水平。
* 字符的基线是对齐的。
* 每一行文本都以单点大小设置。
* 上升，下降和首部有一致的高度。
* 在罗马字体中，衬线是对齐的。
* 字体(包括斜体或粗体等变体)不会在一个单词内改变。
* 在一行文本中，单词和字符空间是单一形式的，单词空间比字符空间大。
* 段落中的文本行间距均匀。

**页面分解**。

我们现在展示如何将块语法概念是如何被扩展到分割和标记整个页面的。 块语法旨在将给定块的每个子块解释为特定（标记）实体。 在子块级别，块语法为摘要、标题块、署名（逐行）块、参考（引用）条目和图标题特性分配标签。 这些较低级别的实体中的每一个都可以具有对应于不同格式的替代语法。 在块的解析过程中，唯一的标签与每个子块相关联。只有那些没有提供块语法的块被分配了标签，才被认为是正确的分段和标记。

块语法指定的所有其他非终结符都等同于下层块语法的起始符号。 与每个块语法的四个阶段不同，即使在原则上，这组块语法不能组合成一个单一的主语法，，因为在每个级别都必须提取一组新的块配置文件，并且位置 每个块边界的大小取决于上面级别的解析。

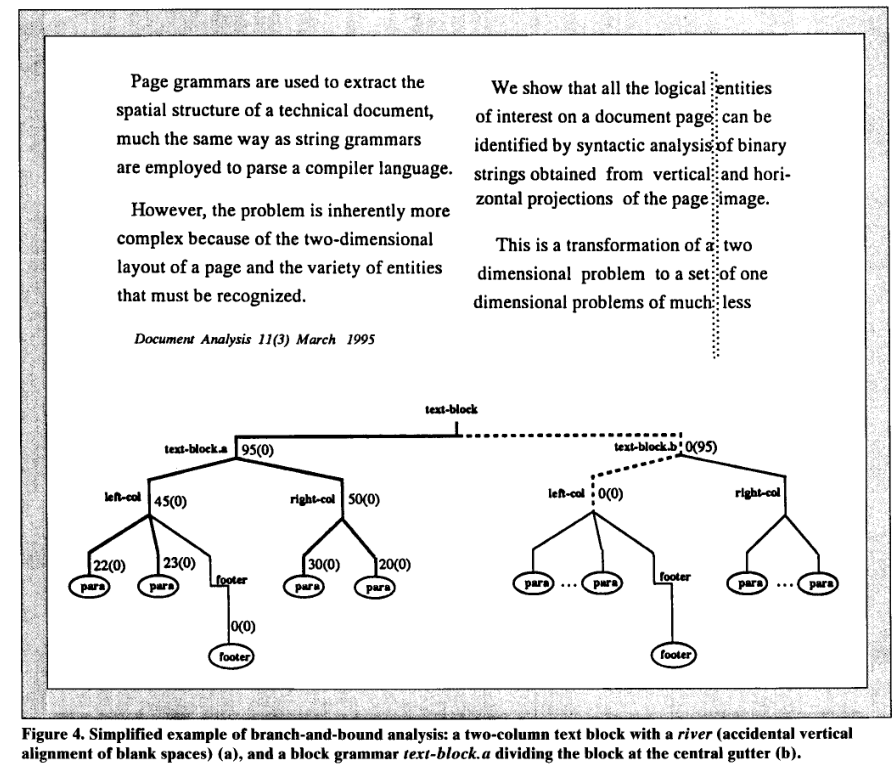
通过解析递归生成的块轮廓，几十种不同类别的对数分量被隔离和识别。细分页面的递归算法验证了深度优先遍历中标签赋值的正确性。只有当给定块及其所有嵌套子块的解析成功时，它才会成功返回。该算法可以依次尝试几个替代块语法中的每一个。例如，如果不知道给定页面代表文章的哪一部分，可以用两个块语法来解析:一个是为标题页设计的，一个是非标题页设计的。因此，解析可以被认为是“与或”树:在每一层，每个子块都必须通过一组可选语法中的一个来标识。

替代语法的另一个原因是块配置文件可能有两个合法的解析。 当一个块的标签不能从它自己的配置文件中确定时，就会发生这种情况，而只能在下面的级别。 在这种情况下，必须使用替代语法在多个假设下解析块。

**错误控制**。

页面的某些部分可能是意外的格式，没有提供替代语法。因为这会在某个时候产生不可识别的子块，所以整个解析将会失败。因此，对“与或”方法进行了修改，以避免灾难性的失败，并有效地确定最佳可能的标记。“最佳”被定义为标记块的最大累积面积。如果页面的最大标记区域不能超过当前下限，分支定界策略可以避免解析子块，即使子块被完全分割和标记(图4)。子块语法left-col正确地将左列分为两个段落和一个页脚，但是页脚语法失败了，因为它需要更小的类型。语法右列（right-col）正确处理右列，因此现在95%的区域(除了页脚块)都被标记了。

图4。分支定界分析的简化示例：具有河流的两列文本块（意外垂直对齐的空间垂直对齐）（a），和块语法text-b1ock.a，划分中央排水沟处的块（b）。



在尝试改进此过程中，第二个替代的顶级语法，Text-B1ock.b（为较窄的排水沟设计）应用于整个块并在河流处将其分割。然而，现在，left-col失败，因为文本行没有穿过中央装订线，所以块不能再规定的行间距分割成水平条。在没有尝试树的最右边分支的情况下放弃搜索，因为即使河右右侧被正确标记，整体结果也将不如早期的解析。搜索树的分支上的数字显示了识别区域的百分比，括号中是适用的下限。

**排印惯例。**

我们的方法使用文档语法形式的特定于公共的知识库。我们还研究了基于排版约定而不是基于出版物特定知识的分段技术。以前的方法通常在处理程序中嵌入典型的图形约束。因此，如果没有广泛的实验，就很难采用现有的程序并对其能力形成清晰的概念。进一步的进展可能有赖于通过对这种信息的公开编纂来发展更加一致和全面的知识库。在侧栏“通用排版规则”中显示了一些用拉丁字母的派生词为技术期刊设置的通用排版知识的例子我们用于光学字符识别的商业软件结合了这样的排版规则来定位和等排一个文本块中的字符。

**OCR**

我们没有开发自己的光学字符识别，而是考虑了商业产品。高端系统由硬件和软件组成。有些只需要额外的处理器板。低端系统只是软件，通常运行在具有扩展内部存储的电脑上。一些光学字符识别系统可以针对特定的字体进行训练。

我们从Caere Corp .选择OmniPage系统是因为它在技术期刊中使用的字体和尺寸方面具有卓越的性能(无需用户培训)。然而，由光学字符识别系统产生的ASCII版本不如键入版本准确。不寻常的字体可能会影响系统，但大多数分类错误都是由于错别字造成的(尤其是小字体、字距、粗体、非下划线和斜体字体)。公式和等式有时会出错。带有首字母的行(如本文的第一个字母)、大的数学符号或上标可能会完全漏掉。有些错误似乎不合理，可能会重复多次。

在其他应用程序中，准确性至关重要，文本将使用拼写检查和人工后期编辑进行更正。然而，即使没有昂贵的后处理，通过光学字符识别获得的ASCII版本不仅对自动搜索有用，而且因为它可以转换成标准的文档交换格式或直接复制到用户的文字处理器或桌面出版系统中。(我们把定价、版权法和瘟疫这些难题放在一边。)

**文本图像压缩。**

我们还试图使用光学字符识别程序的输出来提供文本图像的高度压缩版本。只存储每个字符模式的第一个图形实例。当随后的模式被具有相同字母标签的光学字符识别器识别时，原始模式和新模式的位图被比较，如果它们匹配，则仅存储指向原始位图的指针。所需的匹配程度决定了再现的保真度。早期的符号匹配方案与所有以前的原型模式进行了位图比较。相反，我们使用高效的光学字符识别系统进行模式识别。在以300 dpi扫描的密集全文本打印页面上，初步结果显示压缩率是CCITT Group 4的两倍。

**文档浏览器**

远程图像浏览系统通常使用某种形式的渐进编码和传输。最初，仅显示低分辨率图像。随着附加数据的传输，更多的细节随之出现。这样的方案对于打印文档几乎没有用，因为低分辨率版本几乎难以辨认。取而代之的是，我们根据需要显示页面的一部分，使用之前获得的细分为有意义的块。

浏览器系统由两个程序组成，这两个程序在不同的计算机上进行通信。一个是主机程序，它处理来自远程用户的请求，并管理数字化文档的数据库，标记为X-Y树，以及X-Y树的文本叶节点的ASCII呈现。另一个是响应用户请求的(查询站)客户端程序。用户请求的标记图像部分显示在客户站。为了便于移植，我们还实现了浏览器的X-Windows版本(参见图5)。

主机程序由服务器和数据库管理器组成。服务器处理主机和查询站之间的通信。查询站由客户端和用户界面组成。该接口将用户的请求传递给客户端。客户端与主机建立连接并提交请求。数据库管理器解析服务器在指定端口接收的请求，并将必要的数据发送到查询站。然后，从服务器接收的数据被传输给用户。

用户从弹出式菜单中选择所需的页面。将显示页面的几何表示(一个缩小的标记为X-Y的树，只有叶节点)。然后通过在几何表示中“鼠标”该块来重新寻找所需的块。使用ASCII按钮，用户可以显示ASCII版本的文本块或300 dpi的图像。

**换行**。大多数技术印刷材料都是按列或行排列的，足以在480 x 640像素的VGA屏幕上以300 dpi显示。然而，有些类型(通常是标题和摘要)会遍布整个页面，如果没有笨拙的水平滚动，就无法以图像形式显示。当文本块宽于预设宽度时，我们会将分割过程延续到行级别。然后我们找到字间空白，在每一行文本中，它比最宽的字符间空间更宽。每个文本行在一个或多个单词边界处被划分，连续的段在一个接一个地显示。这是图像中文本的换行形式，而不是符号形式。

**联系**。双文本-图像表示提供了增强文档访问的机会。例如，光学字符识别系统可以识别文本和图形标题中“图形”一词的所有实例。由于我们知道每个图形和图形标题的位置，我们可以推导出每个插图的图形号。然后，当用户反驳文本中提到某个图形时，按下鼠标按钮，该图形就会在单独的窗口中出现在屏幕上。或者，屏幕上任何文本段落中引用的所有图形都可以以降低的分辨率同时显示。

类似的概念也适用于引用的参考文献。同样，ASCII版本需要自动设置链接，但是引用可以从图像或ASCII文本显示中调用。虽然我们只试验了同一文档中的链接，但概念可以扩展到多个文档。这里，图像处理仅在识别图像中识别的关键词的坐标和将页面分割成块中起作用。信息的进一步增强属于超文本领域。

**实施**。解压缩、去歪斜、噪声消除和语法分析程序在Sun-3/60上运行。整个页面的未压缩位图只存储一次。轮廓提取例程通过每次解析后存储在X-Y树中的X和Y坐标来访问所需的块。每个阶段的解析器都是由Unix编译器实用程序Lex和YACC生成的C语言程序。一个Unix shell程序控制不同级别之间的递归。ASCII文件的生成和存储是一个离线过程。微视扫描程序、OmniPage光学字符识别系统和光盘访问程序运行在一台400万字节内存的英特尔80386电脑上。

远程浏览器的主机和客户端程序也在Sun工作站上运行。所有浏览器交互都是在线过程。原型系统已经过测试，主机位于伦斯勒理工学院，查询站位于内布拉斯加大学，反之亦然。一台主机可以同时服务多个查询站。然而，我们的浏览器不具备操作系统所必需的复杂的信息检索和书目导航设施。目前，我们的浏览器只包含几十个页面，可以从简单的菜单中选择。此外，对于认真浏览的人来说，它的响应时间太慢了。然而，它确实展示了与文档图像分析相关的几个功能，这些功能对于通过工作站完全访问技术库是理想的。

**例子。**

图6再现了《IEEE模式分析和机器智能交易》中的一页，它是从光盘中提取的，并打印在Apple- Writer的PostScript表示上。在为页面生成标记的X-Y树之前，图像会被去偏斜和过滤。X-Y树包含页面中所有线段的坐标以及分配的标签。示例中对应于文本的12个子帧被转换为TIFF(标记图像文件格式)，并使用网络文件传输协议传输到电脑。在示例中，这些包括页码、页眉、标题、署名、标题、关键词、章节标题、两个文本段落图、图形标题、脚注和页脚。OmniPage在每个子图像上分别运行，相应的ASCII输出文件返回给Sun系统。

当浏览器被激活时，用户可以连续选择出版物名称、数量、发行和页码。浏览器控制按钮显示在下一页图7的左上角。最终选择提示显示所选页面的X-Y树的标记地理度量表示，如屏幕左下角所示。现在，用户可以在已处理块的ASCII和图像格式副本之间进行选择。多个请求被接受，从图7中可以明显看出，图7显示了标题、脚注和抽象块的ASCII格式。请注意，脚注块中的“手稿”一词被错误识别，摘要第七行中的“是”一词也是如此。OmniPage记录了字符的数量和拒绝的数量，但是当然不能计算替换错误的数量。

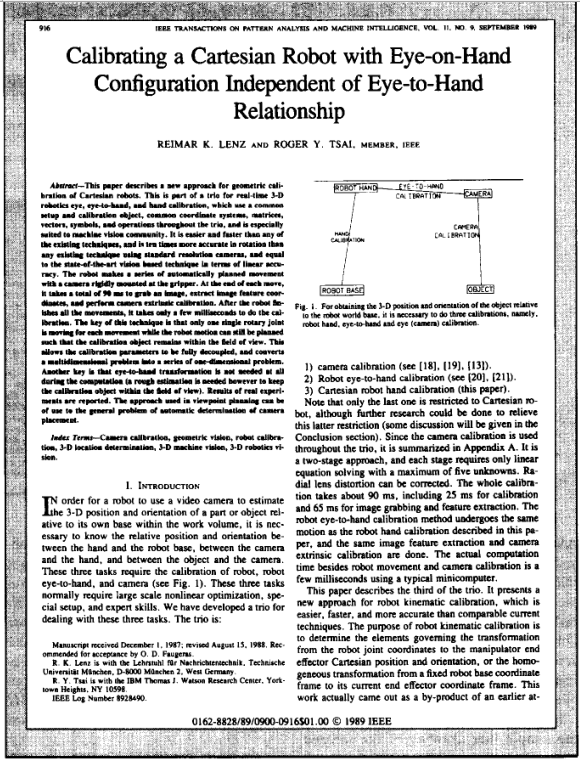


图6。从IEEE光盘数据库中提取的标题页。去歪斜版本用于光学字符识别，并且图像也被过滤用于句法分析

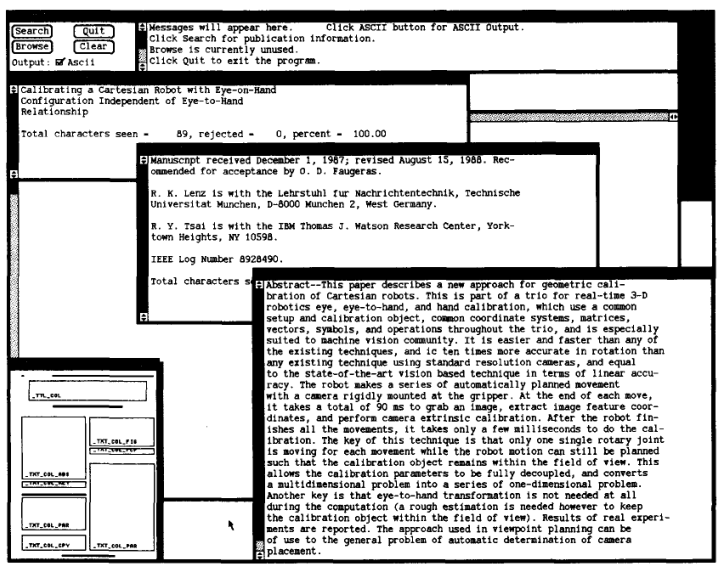


图7。OmniPage处理的三个ZEE-PAMZ文本块的ASCII显示。

图8显示了示例页面的300 dpi版本的简单线条画。由于页面图像目前只以二进制形式存在于IEEE光盘上，因此照片也会以极高的对比度显示。如果图像尺寸超过屏幕尺寸，可以滚动或缩放图像。图9显示了文本块的图像和ASCII显示-图标题

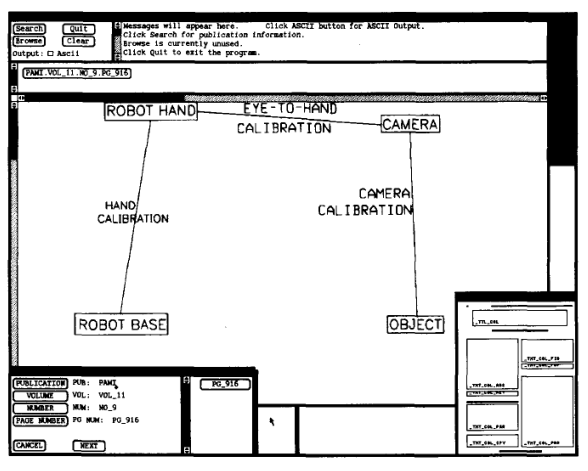


图8。图形块的图像显示。通过单击几何图形表示的右栏(右下角)中最上面的块来选择该块。如果数字足够小，它会以全300 dpi分辨率显示在屏幕上；否则，它会缩小或滚动。

**实验结果**。

我们已经成功地处理了《国际商用机器公司研究与发展杂志》的21页复印件和《PAMI电气与电子工程师协会》的20页复印件。由于这些页面用于开发和改进页面语法，因此我们从每个日志中随机选择12页进行独立测试。PAMI的9页被完美地分割并贴上标签。三个犯了小错误。在IBM Journal的12页中，有7页被完美分割，3页有小错误，2页遗漏了大约四分之一的页面。所有的错误都可以通过对程序块语法的简单修改来纠正，但是很明显，为了达到可接受的性能，需要几个设计和测试周期。当算法失败时，也可以调用类似于光学字符识别中后期编辑的交互式步骤。由于系统本身可以标记故障，因此整体吞吐量不会受到很大影响。

处理每个新页面所需的两到三分钟的大部分时间都消耗在递归配置文件提取中。算法的编码很少考虑效率；例如，我们尽可能使用shell脚本。我们最近在一台32 x 32处理器的DAP(分布式处理器阵列)计算机上实现了一个支持文件提取的算法。初步比较表明，提取2,000 x 3,000像素图像的水平和垂直文件所需的时间减少到十分之一。

尽管技术文档的ASCII表示对于许多目的来说是足够的，但是对于人类的访问来说，忠实地再现原始布局是非常需要的。这种再现不仅对于计算机表示尚未标准化的图形、方程和表格是必要的，而且与仅保留部分信息的光学字符识别输出相比，保留原始布局和排版提高了易读性。

我们已经演示了一个基于语法文档分析和光学字符识别的系统原型，它可以提供对存储的技术文档的有用的远程访问。使我们的系统区别于其他系统的两个方面是:(1)特别适用于印刷品的X-Y树数据结构，以及(2)在越来越多的修饰级别上对图像块进行句法分析。

我们现在正在将我们的原型浏览器与伦斯勒理工学院图书馆信息系统接口。自1988年以来，该系统已经包括所有IEEE索引数据，包括摘要。Infotrax为每篇潜在感兴趣的文章提供出版物的标题、卷、期和页码。我们的样本文档已经准备好了相应的信息索引。

除了用于信息检索，我们还打算调整我们的系统，为自动或交互索引提供输入。一旦文章被处理，只有标题、作者和摘要等相关字段的ASCII版本才会被转发给索引站。正文将保留供后续全文搜索使用，而不是仅搜索选定的索引项。

较长期的研究目标包括开发改进的方法来获取特定于出版物的知识库，可能包括某种形式的学习。然而，我们也在研究，仅使用通用排版知识，可以在多大程度上分析文档。

**总结：**

本文展示了我们的 X-Y 树数据结构如何将二维页面分割问题转换为一系列可以使用传统编译器工具解决的一维字符串解析问题。 接下来，使用句法分析将每个页面划分为标记的矩形块。

通过X-Y树，将矩形分层细分为矩形，然后通过块语法对每个块进行标记（页码、页眉、标题、署名、标题、关键词、章节标题、两个文本段落图、图形标题、脚注和页脚）

**块语法**。 特定期刊的文档语法由一组块语法组成。 每个块语法将一个块水平或垂直细分为一组子块。 因此，应用整个文档语法的最终结果是将页面细分为嵌套的矩形块。 这种细分可以用 X-Y 树 的数据结构来有效地表示（图 1）。 块语法本身也以树的形式组织：用于细分每个块的块语法由上一级解析的结果递归确定。

**X-Y树**：每个节点对应一个矩形块的空间数据结构。每个节点的子节点表示父块在特定（水平或垂直）方向上的细分位置。（X-Y树包含页面中所有线段的坐标以及分配的标签）

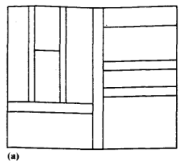


图1. 将矩形分层细分为矩形：用于页面分割的 X-Y 树

至于为什么使用X-Y树：从原始文档到用户计算机屏幕的信息流，页面要么在本地扫描，要么从 CD-ROM 中获取。它使用句法方法进行分割和标记，并存储在 X-Y 树中。

用户从弹出式菜单中选择所需的页面。将显示页面的几何表示(一个缩小的标记为X-Y的树，只有叶节点)。然后通过在几何表示中“鼠标”该块来重新寻找所需的块。

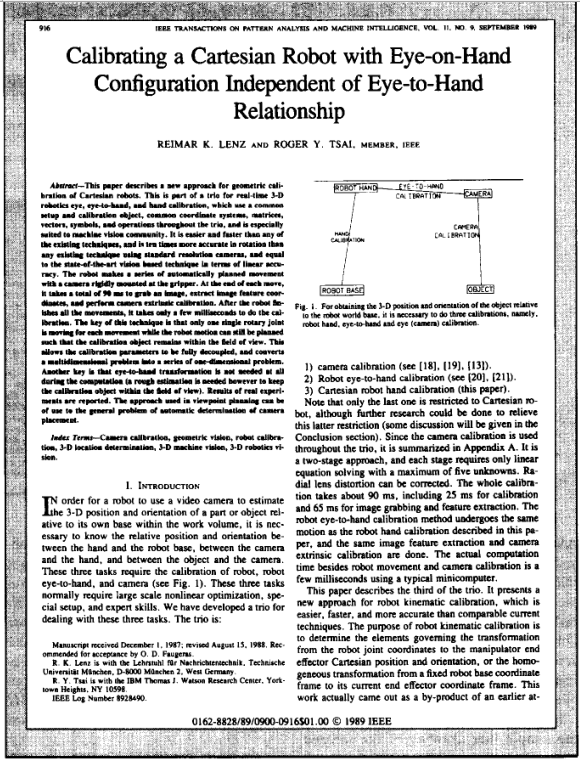


图6。从IEEE光盘数据库中提取的标题页。去歪斜版本用于光学字符识别，并且图像也被过滤用于句法分析

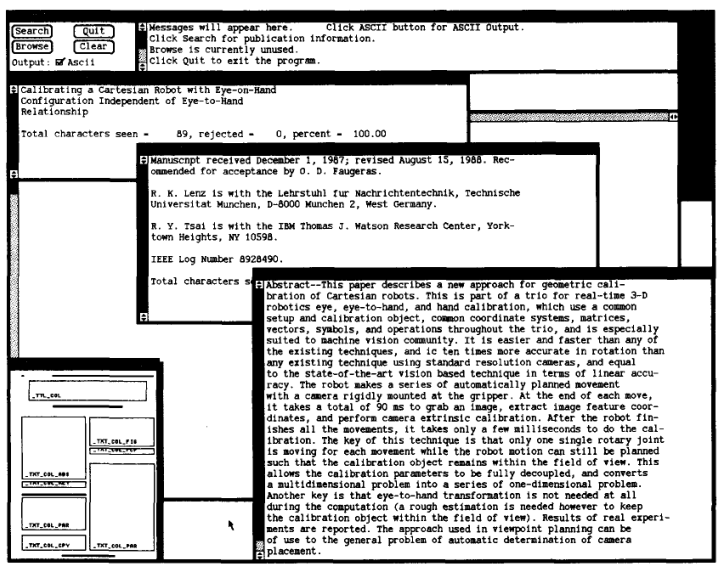


图7。OmniPage处理的三个ZEE-PAMZ文本块的ASCII显示。

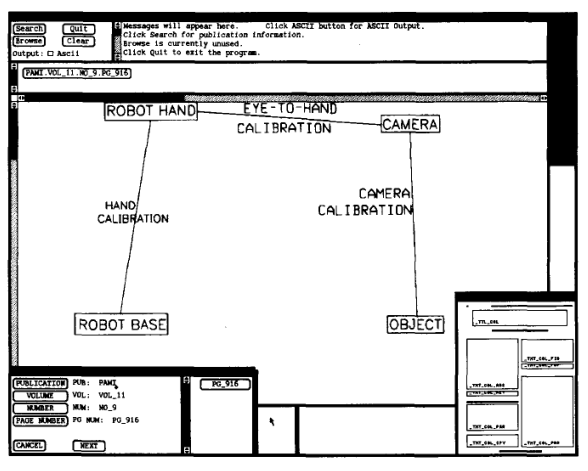
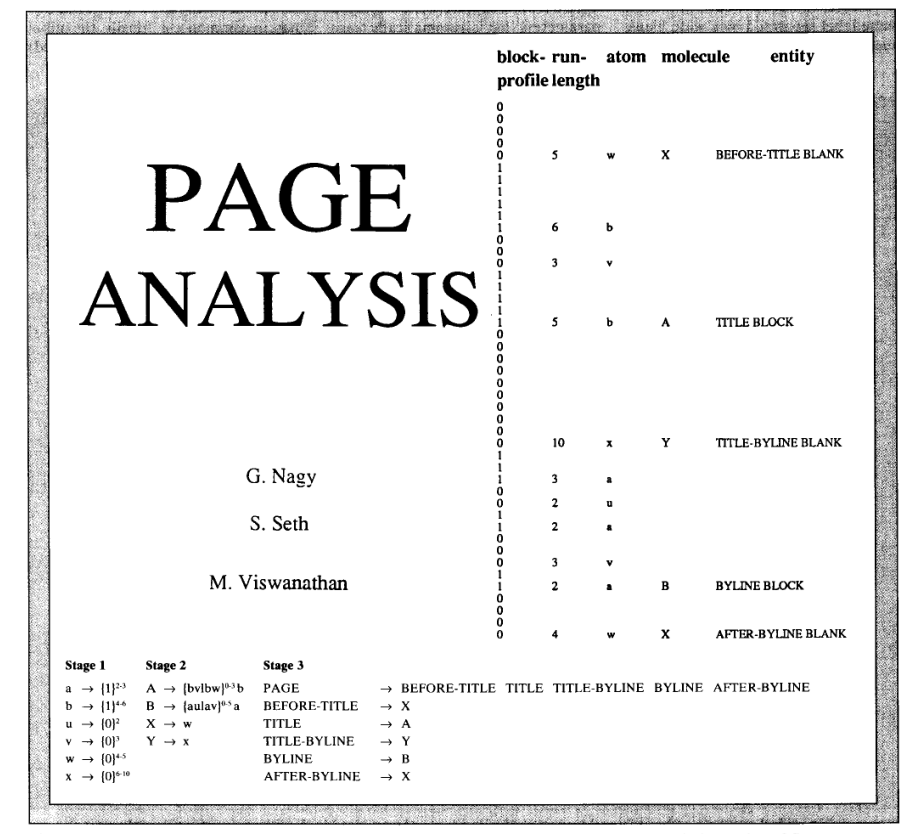


图8。图形块的图像显示。通过单击几何图形表示的右栏(右下角)中最上面的块来选择该块。如果数字足够小，它会以全300 dpi分辨率显示在屏幕上；否则，它会缩小或滚动。

**块语法实现细节：**

每个块语法都作为传统的字符串语法来实现 ，在称为块配置文件的二进制字符串上进行操作。块配置文件是块内黑色区域的阈值垂直或水平投影

* 第 1 阶段，一连串的 1 或 0 根据它们的长度压缩成原子
* 在第 2 阶段，原子被分组为分子
* 在第 3 阶段，分子被解释为文档实体。（第 3 阶段确保标题块在署名块之前）
* 第 4 阶段（未图示）合并由宽空白分隔具有相同标签的连续实体，例如文本段落。



顶部显示了典型的水平块配置文件（中间长串的0000111110001111000....）

如果不知道给定页面代表文章的哪一部分，可以用两个块语法来解析:一个是为标题页设计的，一个是非标题页设计的。

下图为使用不同的块语法解析：

