

[12] 发明专利申请公开说明书

[21] 申请号 00119542.5

[43]公开日 2002年2月20日

[11]公开号 CN 1336604A

[22]申请日 2000.8.1 [21]申请号 00119542.5

[71] 申请人 复旦大学

地址 200433 上海市邯郸路 220 号

共同申请人 上海金鑫计算机系统工程有 限公司

[72]发明人 施伯乐 张 亮

王 勇 陈智峰

印 峻 陈国梁

舒韵宏 焦宇翔

[74] 专利代理机构 上海专利商标事务所

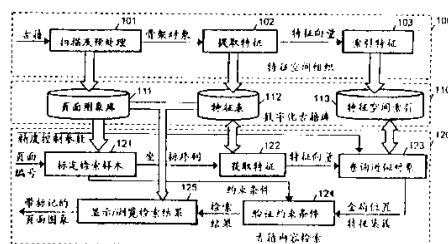
代理人 任永武

权利要求书 2 页 说明书 19 页 附图页数 6 页

[54]发明名称 中文古籍数字化及内容检索自动化方法和系统

[57] 摘要

本发明建立以视觉相似性为基础的计算机古籍内容检索方法和检索系统。设计按照古迹书写规则所确定的对象线性序编号位置特征、页面编号和页面内各对象的几何坐标构成的页面特征、多级重心分划区域笔画因素累计值形态特征和相应的提取技术;提出任意检索点标定方法和提高匹配精度的约束验证技术;创立以及允许检索者在检索阶段利用搜索精度控制参数权衡查全率与查准率,实现即时选择对象形态至对象语义映射的动态调整方法。发明中优化组合图象处理、高维特征空间索引和上述技术,用通用计算机及相关外部设备,实现软/硬件合一的中文古籍数字化和内容检索系统。达到直接在页面图象上自动完成的、支持任意检索点的计算机古籍内容检索技术效果。



权 利 要 求 书

1. 一种中文古籍数字化及内容检索的方法，其特征在于，它由一次性完成的特征空间组织（100）处理阶段和可多次重复的相继的古籍内容检索（120）处理阶段组成；

所述的特征空间组织处理（100）阶段包括以下步骤：

通过扫描和预处理模块（101）产生页面图和将它存入页面图象库（111），同时通过骨架传给后续的提取特征模块（102）以将页面图象中的对象分解为独立图象的有序集合；

通过提取特征模块（102）将所述对象的有序集合分离成页面特征，对象全局位置特征和形态特征向量并将这些特征保存在特征表（112）中；

通过索引特征模块（103）组织所述全局位置特征和形态特征向量并保存于数据结构特征空间索引模块（113）；

通过数据结构特征空间索引模块（113）对形态特征向量进行视觉相似性聚类以及排除与检索点不相似的文字符号图象；以及

通过调整精度控制参数确定特征空间索引模块（113）中的搜索范围，以将其全局位置特征反馈；

所述内容检索（120）阶段包括以下步骤：

通过标定检索样本模块（121）设定页面图象的页面坐标和坐标序列的顺序以形成检索样本，并将座标序列的顺序作为约束条件传给验证约束条件模块（124）；

通过获取特征模块（122）将页面坐标序列作为条件从特征表（112）中确定页面图象的具体对象，以获得与对象相对应的形态特征向量；

通过近似查询模块（123）以形态特征向量为参考点寻找最近邻元素以构成参考点的相似对象集合；并将对应的相似对象集合组成全局位置特征集簇传递给验证约束条件模块（124）；

由验证约束条件模块（124）根据所述约束条件检验集簇元素的有效组合，以形成检索结果；以及

通过显示/浏览检索结果模块（125）将检索结果显现在检索者的客户机屏幕（206b）上并根据所述搜索范围的全局位置特征反馈结果对精度控制参数调整。

2. 一种实现如权利要求 1 所述的检索方法的系统，包括服务器（200a）和客户机（200b）；所述服务器包括中央处理机（202a）、随机存储器（203a）、硬盘（204a）、键盘（205a）、显示器（206a）、网络连接设备（207b），扫描仪（208）和指示设备（209a）；所述客户机包括中央处理器（202b）、主存储器（203b）、硬盘或只读存储器（204b）、键盘（205b）、显示器（206b）、网络连接设备（207b）和指示设备（209b）；其特征在于：所述服务器(200a、200b)的硬盘（204b）含有永久性存储计算机操作系统检索应用软件、页面图象和特征空间索引文件；它们由下列模块组成：

扫描和预处理模块（101）；
提取特征模块（102）；
索引特征模块（103）；
数据结构特征空间索引模块（113）；
标定检索样本模块（121）；
获取特征模块（122）；
查询近似模块（123）；
验证约束条件模块（124）；和
显示/浏览检索结果模块（125）。

3. 如权利要求 2 所述的系统，其特征在于：所述的扫描仪（208）是一种数字化的扫描仪。

4. 如权利要求 2 或 3 所述的系统，其特征在于：所述服务器（200a）和客户机（200b）连接于网络（210）。

5. 如权利要求 4 所述的系统，其特征在于：所述的网络（210）是广域网。

6. 如权利要求 5 所述的系统，其特征在于：所述客户机（200a）与服务器之间的通信遵循 HTTP 协议。

7. 如权利要求 4 所述的系统，其特征在于：所述网络是局域网。

8. 如权利要求 2 或 3 所述的系统，其特征在于：所述服务器（200a）和客户机（200b）是同一台机器；网络连接设备（207a、207b）采用 loopback 适配器。

9. 如权利要求 2 或 3 所述的系统，其特征在于：所述服务器的计算机操作系统是 Windows95/Windows98 (Microsoft 商标)、MacOSC Apple 商标)、Unix 的各种版本之一。

说明书

中文古籍数字化及内容检索自动化方法和系统

本发明涉及高速、且以内容为其目的的中文古籍文献数字化及在数字化古籍页面图象中直接实现内容检索的自动化方法和系统。

古籍作为人类文化遗产的重要组成部分，具有极高的学术研究和艺术欣赏价值。由于其珍奇、稀有的特点，古籍的上述价值无法在大范围内为公众所利用，即使在严格限定的范围内，古籍原件的安全性和可持续保藏性依然难以保障。对古籍文献的发掘和有效利用已成为各国数字化图书馆(Digital Library)工程的主要目标之一。迄今为止，提出的各种古籍数字化和数字化媒体的利用方式可归纳如下：

标引加图象浏览方式。首先以预定的分辨率扫描古籍页面，消除噪声后作为古籍页面的数字化媒体(简称“页面图象”)保存于大容量存储装置(常用光盘)中。图书馆或博物馆专业人员对页面图象标引(如按部/类/属/目分类、书名、著者时代、著者姓名、著作方式、出版年代、出版地、出版者、版式、行款、批校者、题跋者、藏印、封面、扉页、序文、前/后添加页、凡例、目录、图、附录、跋等)，作为页面图象的附加信息并建立相关索引，保存在存储装置中备查。检索者利用数据输入设备(键盘或鼠标)，通过系统提供的有限数量的检索点(常见的是书号、部/类/属/目分类、书名、著者时代、著者姓名)检索古籍，然后浏览全书或部分页面的页面图象，也可根据预先标引信息浏览古籍的页面图象中的封面、前/后添加页、扉页、序文、凡例、目录、图、附录、跋等。系统一般还提供了浏览过程中可控制页面的进退和图象放大/缩小等辅助功能。这种方式的主要问题在于：

标引项目不会很多；

检索点不会多于标引项目；

标引项目难以覆盖检索者的特定检索目标；

除检索点外，页面图象中的大部分内容只可浏览，不能达到古籍内容检索的效果。

附带文本文件加文本文件全文检索方式。首先根据古籍制作与之对应的文本

文件(如人工键盘录入), 然后应用全文检索技术对该附带文本文件实现文字内容检索, 最后再由对应关系调出页面图象。这种间接方式在其必不可少的附带正文文件的生成阶段, 正文文本与古籍原稿内容的同一性判定、字符集规模、特殊符号处理、自动化程度等方面存在着图书馆或博物馆业务无法接受的制约条件; 这些问题致使中国专利申请公开说明书 CN-1151558A 中提出的基于文本文件形式的信息检索方法和系统无法应用于以图象为其实质的古籍页面的内容检索应用。另外, 通假字在古籍中的广泛使用, 也使全文检索技术对古籍内容检索缺乏必要的能力。

光学字符识别加文本文件全文检索方式。该方式用光学字符识别(OCR)技术生成古籍对应的文本文件和检索对象, 然后应用全文检索技术对该附带文本文件实现文字内容检索, 最后再由对应关系调出页面图象。然而, 由于古籍出版年代、版本形式不同, 古籍用字差别巨大, 无法建立包括所有古今字词的词典; 更由于中文古籍中毛笔手书汉字笔画模糊、不规范、笔画间/部件间的相对位置不稳定、笔画倾角/相对长度不稳定、书写风格差异、软笔笔画变形等诸多因素, 难以完成软笔手书字体的准确识别。中国专利申请公开说明书 CN-1165571A 中提出了一种生成与检索对象形状相似的文字串(如“中间决算”与“牛间决算”）、对每种可能的变形分别应用一次文本文件全文检索的方法, 以回避错误识别给检索带来的上述问题。但是, 该方法对古籍而言是无能为力的。因为文字串的变形数是随文字串长度以指数规律增长的。例如, 设每个字的平均变形数为 k , 文字串长度为 n , 则可能的变形文字串总数为 k^n 。因此, 该方法在算法上缺乏可伸缩性(Scalability), 反映到应用中, 是缺乏实用性。OCR 作为附加文本文件生成工具的另一个重要缺陷是: 古籍文字/符号对象(以下简称“对象”)的语义在 OCR 识别阶段已“冻结”, 即对象的图象已确定性地映射到一个文字。检索者在检索过程中没有任何能力改变已被附加文本文件制作者冻结的语义映射。在以毛笔手书为主要特征的中文古籍作品中, 手写字体的变化、页面纸质的污损都不可避免地导致对象的语义无法唯一确定, 需要检索者根据工作目标即时地做出选择, 例如确定查全率和查准率的折中。这一要求无法被基于 OCR 的古籍内容检索方法所满足。

总之, 对于以毛笔手书汉字为其主要特征的中文古籍作品, 其内容检索问题十分困难。目前尚无有效的、直接的内容检索方法和系统。

本发明的目的是提出一种直接在页面图象上自动完成的、基于视觉相似性的、任意检索点的计算机古籍内容检索新方法。

本发明的又一目的是提出一种允许检索者在检索阶段即时选择对象形态至对象语义映射的动态调整方法。

本发明的另一目的是提出一种可以与目前图书馆常用的标引方法配合使用的查询/浏览相结合的古籍检索工具。

本发明的再一目的是使用通用计算机及相关外部设备，建立能够实现上述方法技术效果的中文古籍数字化和内容检索系统。

本发明中，基于视觉相似性的计算机古籍内容检索方法，其特征在于，由特征空间组织和内容检索两个相继阶段构成；特征空间组织为古籍中的内容(对象及其序列关系)生成其特征聚类，建立易于根据视觉相似性快速查找近似对象的索引结构；内容检索是利用该索引结构，自动地快速获得所有与检索者给定对象视觉内容近似的其他对象；对于待处理的古籍，特征空间组织阶段一次性地完成，而内容检索阶段可根据检索者的要求多次重复。

本发明中利用了图象处理、特征提取、高维特征空间索引、任意检索点标定、特征快速匹配和约束验证等技术，其特征在于：优化组合这些技术，利用通用的计算机和外部设备，实现直接在页面图象上自动完成的、基于视觉相似性的古籍内容检索；对从属于优化方法的按照古迹书写规则所确定的对象线性序编号位置特征、页面内对象几何布局的页面特征、多级重心分划区域笔画因素累计值的对象形态特征定义和对这些特征提取；对任意检索点标定和对提高匹配精度的约束验证；以及检索者在检索阶段利用搜索精度控制参数权衡查全率与查准率，实现即时选择对象形态至对象语义映射的动态调整。

以上述特征和处理方法为核心，用通用计算机及相关外部设备，建立能够实现新技术效果的、软/硬件合一的中文古籍数字化和内容检索系统。

本发明由检索者在古籍页面图象上随意地标定检索对象，可提供任意的检索入口点，完全满足检索者的特定检索目标的需要；由于检索对象直接出自于页面图象，无须考虑同一性判定、字符集规模、特殊符号处理、通假字、词库等问题，自动化程度高、操作简便，易于图书馆工作人员使用；利用“近似匹配”的技术路线，摆脱了由“识别”方法引进的额外困难——即目前尚不能完全准确的由对

象形态到对象语义的抽象过程；利用搜索精度控制参数权衡查全率与查准率的动态语义映射选择机制适应了中文手写字体变化和古籍污损的工作环境；发明中提出的采用多级重心分划区域笔画因素累计值的对象形态特征，是一种特征提取的优化实施方案，它正确体现了手写文字的视觉内容，即相对灵活的笔画分布密度所表达的文字/符号。其中，以对象高度和宽度的最大值为单边长的正方形位图规格化方法较好地保持了对对象的宽高比特征；依区域重心对位图作多级分划较好地解决了手写文字里笔画/部件间的相对位置偏移的问题；基于笔画因素的特征构成对软笔手写汉字笔画不均匀、笔画模糊、倾角/相对长度缺乏规律等现象，都有较强的容错能力，也便于对古籍中的非文字符号对象的统一处理。本发明的方法还能够与目前图书馆常用的标引方法配合使用，形成查询/浏览相结合的古籍检索工具。

下面结合附图说明本发明的实施例。

图 1 是系统结构和基本处理流程图；

图 2 是系统硬件结构的方框图；

图 3 是检索方法总体流程图；

图 4 是特征空间组织流程图；

图 5 是内容检索流程图；

图 6 是流程图中使用符号意义说明；

图 7 是二值位图纵向投影示意图；

图 8 是平滑用辅助栅格；

图 9a 和 9b 是加注分列标记的页面图象；

图 10a 和 10b 是从列中划分对象的处理结果；

图 11 是分划出的对象例；

图 12 是图 11 的细化位图；

图 13 是图 12 经规格化后的位图；

图 14 是图 13 基于重心的一级和二级区域划分例；

图 15 是横、竖、撇、捺笔画因素定义；

图 16 是一级区域和二级区域编号规则；

图 17 是图 14 的撇、捺笔划因素在一级划分区域中的分布以及横、竖笔划因

素在二级划分区域中的分布图；

图 18 是根据精度控制参数调整和确定搜索范围的处理示意图。

现参照图 1 说明本发明的系统检索方法的基本处理流程。应注意：图 1 中的两个处理单元标定检索样本 121 和显示/浏览检索结果 125 作为程序文件单独或整体存储在图 2 的硬盘 204b 中；其余各个方框图表示的处理单元作为数据文件或程序文件单独或整体存储在图 2 的硬盘 204a 中。

本发明中的检索方法与技术由特征空间组织 100 和古籍内容检索 120 两个相继处理阶段构成，前者产生的数字化古籍库 110 为后者提供基础。特征空间组织阶段 100 一次性完成，古籍内容检索阶段 120 可根据检索者的要求多次重复。

古籍经过扫描和预处理模块 101，一方面产生页面图象存入页面图象库 111 以备用户浏览，另一方面页面图象中的对象通过骨架传给后续的提取特征模块 102 被分解为独立图象的有序集合。存入库 111 中的页面图象可以是原始扫描结果(如彩色图象或灰度图象)，保持古籍原有的视觉形象和风格；也可以是经过预处理加工后的清晰图象，获得较好的可读性。对象有序集合又被提取特征模块 102 分离成转换为三类特征：页面特征、对象的全局位置特征和形态特征序列。这些特征保存在特征表 112 中。模块 102 提取的全局位置特征和形态特征向量由高维空间索引特征模块 103 加以良好的组织并保存于数据结构特征空间索引模块 113 中。除了对特征向量对象的数学表达的视觉相似性聚类之外，特征空间索引结构 113 的另一个职能就是及时排除与检索点不相似的文字/符号图象，加速搜索查询点的视觉相似的对象。这是古籍内容检索实现高速化的基础。

内容检索阶段 120 采用按示例查询的工作方式。标定检索样本模块 121 支持检索者在所浏览的页面图象上随时、任意地标定对象，记录客户端指示设备 209b 点击页面图象时的页面坐标和该坐标序列的顺序，形成检索者的检索样本。坐标序列的顺序作为约束条件传给验证约束条件模块 124。页面坐标序列本身被获取特征模块 122 用来作为条件从特征表 112 中确定页面图象中的具体对象，获得与对象相应的特征向量。查询近似对象模块 123 以得到的特征向量为参考点，在特征空间索引 113 中的寻找最近邻元素，构成参考点的相似对象集合。该模块 123 同时将与检索样本中所有对象对应的相似对象集合组合成全局位置特征集簇交给验证约束条件模块 124。由模块 124 根据得到的约束条件检验集簇元素的有效组

合，形成检索结果。这些结果由显示/浏览检索结果模块 125 以醒目的方式显现在检索者的客户机屏幕 206b 上。供用户浏览和观察其上下文。

检索者可以通过调整精度控制参数的取值来获得查全率与查准率的权衡。精度控制参数仅仅是由用户指定的系统给定的线性区间中的一个点。系统将其取值作为一个参数，确定在特征空间索引结构 113 中的搜索范围，把范围内点（近似的候选对象）的全局位置特征返回。由于检索者可以立即通过显示/浏览检索结果模块 125 观察反馈结果，并能够再次对精度控制参数进行调整，重复古籍内容检索阶段 120 的过程，观察变化的效果，所以精度控制参数的具体取值既不要求准确也不要求特殊技术和技巧。

现参照图 2，图中例示了用以实施本发明的系统硬件结构。它们是连接于网络 210 的服务器 200a 和客户机 200b。服务器 200a 用于数据和页面图象的存储、维护、管理、检索以及检索结果的传输。其硬件系统是由总线 201a 联系在一起的通用计算机结构，包括具有运算和控制输入输出功能的中央处理机 202a、保存程序和运算中间数据的随机存取存储器 203a、永久性存储计算机操作系统、检索应用软件、页面图象、特征空间索引文件等内容的硬盘 204a、用以键入命令与参数的键盘 205a 和显示命令反馈结果的显示器 206a、网络连接设备 207a、古籍页面数字化的扫描仪 208 和功能选择与辅助定位设备即指示设备 209a；客户机 200b 负责人机界面的操作、送出查询浏览需求及显示浏览查询结果。其硬件系统是由总线 201b 联系在一起的通用计算机结构，其中包括具有运算和控制输入输出功能的中央处理机 202b、保存程序和运算中间数据的主存储器 203b、永久性存储计算机操作系统、检索应用软件等内容的硬盘 204b1（或只读存储器 204b2）、用以键入命令与参数的键盘 205b、显示页面图象和命令反馈结果的显示器 206b、网络连接设备 207b、帮助指定显示器 206b 上屏幕位置的指示设备（如鼠标器、手写笔）209b；服务器和客户机通过网络连接设备 207a、207b 经由网络 210 联系起来，互通信息。

作为上述实施方案的另一种特例，网络 210 可以是广域网（WAN，如 Internet）。在被称作浏览器/服务器模式的系统结构中，客户机 200a 和服务器 200b 之间的通信遵循 HTTP 协议。客户机 200b 通过指定服务器 200a 的统一资源定位器（URL）地址来指定某个 Web 页，然后帮助检索者准备检索/浏览请求，传送

请求至服务器 200a，并接受服务器 200a 传来的页面图象及相关信息（如 JAVA 小应用程序）；服务器 200a 存放以 HTML 语言编写的超媒体文件，它有一个 HTTP 守护进程，它接收客户机 200b 提出的请求并做出响应，每当该进程接收到一个请求时，就创建一个新的子进程为该请求服务，完成合法性检查，针对客户机的请求进行处理并制作数据，包括使用 CGI 程序对数据进行前期和后期处理，然后，把处理好的页面图象等发送给提出请求的客户机 200b。

作为上述实施方案的又一种特例，网络 210 可以是局域网（LAN）。

作为上述实施方案的又再一种特例，服务器 200a 和客户机 200b 可以是同一台机器，此时没有网络 210、网络连接设备 207a、207b，采用 loopback 适配器；总线为 201a、中央处理机为 202a、随机存取存储器为 203a、硬盘为 204a、键盘为 205a、显示器为 206b、扫描仪为 208、指示设备为 209a。

在另一实施方案中的客户机，可以采用移动计算设备（如笔记本电脑、PDA 等）。

服务器的操作系统可以是 Windows95/Windows98（Microsoft 商标）、MacOS（Apple 商标）、Unix 的各种实现版本如（IBM 的 AIX 或自由软件 Linux），不要求多窗口和图形人机界面，但应支持 HTTP 访问协议；客户机可以采用上述任何一种操作系统，但同时要求多窗口和图形人机界面，以及支持 HTTP 访问协议；当采用客户机/服务器在一台计算机上的实施方案时，操作系统取客户机端的配置；当客户机是 PDA 等手持设备时，该手持设备的操作系统或其等同物应支持 HTTP 访问协议。

下面进一步具体说明本发明检索方法的流程特点和所采用的技术。

本发明的视觉相似性的计算机古籍内容检索方法由一系列的技术单元有机组合而成。各个技术单元可以采用公知的技术方案实现，也可以用本发明提出的技术方案实现，以换取较高的执行效率。组合这些技术单元实现直接在页面图象上自动完成的、基于视觉相似性的古籍内容检索技术效果的检索方法是本发明的主要内容，从属于检索方法的一些关键技术是本发明的又一内容。图 3 是检索方法的总体流程图，图 4、图 5 是图 3 的详细流程图。图 6 是流程图中使用符号意义说明。

如前所述，检索方法由特征空间组织 100 和古籍内容检索 120 两个相继处理

阶段构成。特征空间组织阶段 100 由古籍信息服务供应商预先一次性完成。其生成结果，即图 1 中的数字化古籍库 110 保存在图 2 中服务器端的硬盘或光盘 204a 中。古籍内容检索阶段 120 可根据检索者的要求多次重复，它利用硬盘或光盘 204a 中存储的数字化古籍库。两个阶段 301 和 302 不必在时间上连续，仅要求保证如图 3 给出的顺序即可。

现结合图 4 进一步说明特征空间组织阶段。特征空间组织的目的如前所述是古籍中的内容（对象及其序列关系）生成其特征聚类，建立易于根据视觉相似性快速查找近似对象的索引结构。特征空间组织阶段的基本步骤如下：

1. 扫描古籍页面 101a

通过可见光或其他光源按照古籍页码编号逐页扫描古籍，得到其数字化彩色或灰度图象。对保存完好的古籍，可采用普通平板式扫描仪，对于被火损或其他原因损坏的古籍，可用远红外或其他光源照射，显现被遮掩的文字。

2. 预处理 101b

为突出古籍内容、克服扫描误差、分离前景对象和背景噪声、获得对象，在正式构造特征空间索引 113 之前，进行版面倾斜校正、噪声消除、二值化和列/对象的分划、对象细化等预处理工作。可用标准的中文光学字符识别(OCR)技术的预处理手段或组合图象处理的功能，必要时需少量的人工干预实现。以下给出一些实施例。

(1) 色彩和灰度处理

由扫描步骤 101a 得到的数字化古籍页面图象可以是彩色或灰度的。这样做的目的是为了最大限度地保持古籍的原貌，便于用户观赏。为后续步骤的处理需要，供提取特征的页面图象应该转换为黑白两色的，即所谓的二值图象或位图。供用户观赏的页面图象仍可保持原来的色彩或灰度。

彩色图像一般表达为 RGB 或其他色彩空间(如 YIQ)的点集。从图象压缩的角度来看，采用非 RGB 色彩空间的方案的情况更为普遍。因为这些方案将图像的主要特征集中于空间中的某一个坐标轴上，对该轴上的灰度图象进行处理，能够基本体现图象形态。在中文古籍内容检索领域中，采用上述方案将彩色图象转变为灰度图象仍能保持文字/符号对象的形态。

一种具体实施方案是将彩色图象分解为 Y、I、Q 三个分量，再将其中的 Y

分量作为灰度图像留作进一步的处理。Y 分量包含了原始图像的主要信息。YIQ 与 RGB 间的转换关系为：

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.322 \\ 0.211 & -0.523 & -0.312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \quad \begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.0 & 1.176 & 0.763 \\ 1.0 & -0.411 & -0.677 \\ 1.0 & -0.964 & 1.487 \end{bmatrix} \begin{bmatrix} Y \\ I \\ Q \end{bmatrix}$$

灰度图象经过二值化成为二值位图。二值化的关键是确定合适的阈值。一种选择方法是根据灰度直方图确定整体阈值。设灰度级数目为 G ，图象的像素总数为 n ，第 k 级灰度 ($1 \leq k \leq G$) 的像素数为 n_k ，统计图象在灰度级 ($1 \leq k \leq G$) 处的出现频率

$$p(k) = \frac{n_k}{n}, \quad k = 1, 2, \dots, G$$

并以 $p(k)$ 为纵坐标， k 为横坐标作图，得到图象的灰度直方图。中文古籍的灰度直方图一般是双峰的，两个尖峰分别代表了前景和背景像素。灰度阈值可取在双峰之间的波谷处，例如取值 $1 \leq g \leq G$ 。根据灰度阈值 g 将灰度图象 IMG_g 转变为二值位图 IMG_b ：

$$IMG_b(i, j) = \begin{cases} 1, & IMG_g(i, j) \geq g \\ 0, & IMG_g(i, j) < g \end{cases}, \quad \begin{matrix} i = 1, 2, \dots, R \\ j = 1, 2, \dots, C \end{matrix}$$

其中， R, C 分别是图象像素矩阵的行和列数。

对于多峰的灰度直方图，可采用局部阈值二值化方法。

(2) 版面校正

扫描获得的页面图象会因为古籍原稿置放角度的不准而发生偏斜，影响后续处理。大多数情况下，偏斜的角度不会太大。设偏离正常位置（如垂向）的范围是 $[-A, +A]$ 。以 a 为增量，从 $-A$ 起旋转二值位图，按下面方法计算投影密度，直至 $+A$ 。记录有最大投影密度的二值位图作为校正图。

参照图 7，首先，将某一旋转后的位图（图 7 的上半部分）沿纵向投影，得到图象前景像素的水平分布（图 7 的下半部分）。令投影宽度为 W ，则平均线高度

$$h = \frac{\sum_i \sum_j IMG_b(i, j)}{W}$$

在水平分布的平均线上计算投影密度

$$\rho = \sum_k \frac{n_k}{W_k},$$

上式中, n_k 是平均线上的第 k 个连续段中高于 h 的点数, W_k 是这些点在平均线上的投影宽度。选择平均线上的投影而不是所有水平分布的投影有助于减少页面图象上下横线和边界的影响。

(3) 消除噪声

使用平滑技术消除二值位图中残留的孤立点, 平滑笔划边沿。平滑过程是图象处理技术中低通滤波的应用。

一种简单的实施方案是采用如图 8 所示的 3×3 栅格决定象素 x_0 的取值。若以 x 表示象素 x 取值为 1 (前景色), 以 $\sim x$ 表示象素 x 取值为 0 (背景色), 则象素 x_0 平滑后的结果是:

$$\begin{aligned} x_0 = & \sim x_0 [x_3 x_7 (x_1 + x_5) + x_1 x_5 (x_3 + x_7)] + \\ & x_0 \sim [\sim (x_3 + x_7) \\ & (\sim (x_4 + x_5 + x_6) + \sim (x_1 + x_2 + x_8) + \sim (x_1 + x_5) + \sim (x_6 + x_7 + x_8) + \sim (x_2 + x_3 + x_4))] \end{aligned}$$

(4) 对象分割

汉字 OCR 的行、字切分技术可以直接用于对象分割。以下是另外一种较为简单的对象分割方法。它分为分列、分词和调整三个接续步骤。如前所述, 二值位图 IMG_b 的宽度为 C , 高度为 R , 在坐标 (i, j) 处的象素记为 $IMG_b(i, j)$, $IMG_b(i, j) = 1$ 表示该点为前景色。

A. 分列

令第 j 列上像素总数为

$$C_j = \sum_{i=1}^R IMG_b(i, j),$$

C_j 构成的水平分布图光滑后的结果记为

$$S_j = \frac{1}{d} \sum_{d=0}^{\mu-1} C_{j+d}, \quad (j = 0, \dots, C - \mu).$$

其中, μ 为光滑步长。 S_j 的最大值、最小值和两者之差分别记以:

$$M = \max\{S_j\}, \quad m = \min\{S_j\}, \quad D = \max - \min$$

再令: $Th = m + \alpha D$, 其中 α 是阈值参数, 一般取 0.1 或 0.2。求出 $S_j = Th$ 的 j 值 $j_0, j_1, \dots, j_{2n-1}$, 这些值依序两两组对, 即: $p_k = (j_{2k} + j_{2k+1})/2$, $0 < k < n$, 可得到页面

的列分隔线序列 p_k 。如图 9a 中的虚线所示。为提取易于处理的对象列，在分词之前还应排除竖线。具体方法是：计算平均列宽 $\delta = (p_{n-1} - p_0) / (n - 1)$ ，如果两相邻列分隔线 (p_k 和 p_{k+1}) 间距小于 0.1δ ，则认为此两相邻列分隔线之间是古籍的列分隔竖线，当将它们之间填为背景色并且用 $(p_k + p_{k+1}) / 2$ 替代这两条列分隔线。图 9a 经过排除竖线后，得到图 9b 中已分割开来的白色条块。

B. 分词

将得到的对象列视为原始页面图象，调换步骤 A. 中的行、列标记。可以得到各个对象的基本划分。具体结果见图 10a。

C. 调整

自动分割区域有时存在少量的误判结果，分割技术应提供图像反馈，供处理人员手工调整分割区域。这是用图 1 中服务器端的指示设备 209a 选择删除/增加功能，然后点击相应对象或位置。例如，删除图 10a 顶部由原古籍外边框导致的一条无用分割线后得到正确的对象分割，如图 10b 所示。一个分割完成的对象图例如图 11 所示。

(5) 细化

将对象的二值位图转化为线宽为单像素的骨架图象，以减少因笔画宽度差异对特征提取的影响。细化算法如下：

i. $I'' = \text{IMG}_b$;

ii. Do

a. $I = I''$;

b. 扫描 I 中的所有像素，形成新位图 I' 。对 I 中像素 x_0 ，考察其如图 8 所示的邻域，若 C_1 成立，则 I' 中相应位置置 1；

c. 扫描 I' 中的所有像素，形成新位图 I'' 。对 I' 中像素 x_0 ，考察其如图 8 所示的邻域，若 C_2 成立，则 I'' 中相应位置置 1；

Until $I = I''$;

iii. 返回 I'' .

$$\begin{aligned}
C_1 = & x_0 \sim x_1 \sim x_2 \sim x_3 x_4 x_5 x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + \\
& x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 \sim x_4 x_5 \sim x_6 x_7 \sim x_8 + \\
& x_0 \sim x_1 \sim x_2 \sim x_3 x_4 x_5 \sim x_6 x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 \sim x_4 x_5 x_6 x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 x_4 x_5 x_6 x_7 \sim x_8 + \\
& x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 \sim x_6 x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 x_6 x_7 \sim x_8 + \\
& x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 x_6 x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 \sim x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + \\
& x_0 \sim x_1 x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 \sim x_4 x_5 x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 x_4 x_5 x_6 \sim x_7 \sim x_8 + \\
& x_0 \sim x_1 x_2 x_3 x_4 x_5 x_6 x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 x_4 x_5 \sim x_6 x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 \sim x_4 x_5 x_6 x_7 \sim x_8 + \\
& x_0 \sim x_1 x_2 x_3 x_4 x_5 x_6 x_7 \sim x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 \sim x_4 x_5 \sim x_6 x_7 x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 x_4 x_5 \sim x_6 x_7 x_8 + \\
& x_0 \sim x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 \sim x_4 x_5 x_6 x_7 x_8 + x_0 \sim x_1 \sim x_2 \sim x_3 x_4 x_5 x_6 x_7 x_8 + \\
& x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 \sim x_6 x_7 x_8 + x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 x_7 x_8 + x_0 \sim x_1 \sim x_2 x_3 \sim x_4 x_5 x_6 x_7 x_8 + \\
& x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 x_6 x_7 x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 x_5 \sim x_6 x_7 x_8 + \\
& x_0 x_1 \sim x_2 \sim x_3 x_4 x_5 \sim x_6 x_7 x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 x_5 x_6 x_7 x_8 + \\
& x_0 x_1 \sim x_2 \sim x_3 x_4 x_5 x_6 x_7 x_8
\end{aligned}$$

算法结束时的位图即为细化后的骨架图象。算法中的条件

$$\begin{aligned}
C_2 = & x_0 \sim x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 \sim x_1 x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + \\
& x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 x_1 \sim x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + \\
& x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 \sim x_8 + x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + \\
& x_0 x_1 \sim x_2 x_3 x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 x_6 x_7 \sim x_8 + x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + \\
& x_0 x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 x_1 x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + \\
& x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 \sim x_8 + x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + x_0 x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 x_7 \sim x_8 + \\
& x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 x_6 x_7 \sim x_8 + x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 \sim x_7 \sim x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + \\
& x_0 x_1 \sim x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 x_8 + x_0 x_1 \sim x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 x_8 + x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + \\
& x_0 x_1 \sim x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + \\
& x_0 x_1 \sim x_2 x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 \sim x_7 x_8 + x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 \sim x_6 \sim x_7 x_8 + \\
& x_0 x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 \sim x_7 x_8 + x_0 x_1 x_2 x_3 x_4 x_5 \sim x_6 \sim x_7 x_8 + x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + \\
& x_0 x_1 x_2 \sim x_3 \sim x_4 \sim x_5 x_6 x_7 x_8 + x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 \sim x_6 x_7 x_8 + x_0 x_1 x_2 x_3 x_4 \sim x_5 \sim x_6 x_7 x_8 + \\
& x_0 x_1 x_2 x_3 \sim x_4 \sim x_5 x_6 x_7 x_8
\end{aligned}$$

(6) 规格化

为消除手写体对象尺寸和位置变化的影响，规格化各对象的骨架图象。例如，图 13 是图 12 的骨架图象的规格化位图，外框表示新位图的边界。

规格化方法是选择骨架图象的高度和宽度的最大值作为单边长，作一正方形位图。然后将骨架图象置于该正方形位图正中。称上述正方形为 MBS(Minimal Bounding Square)。与使用外接矩形 MBB(Minimal Bounding Box)的常规规格化方法相比，这里的规格化方法保持了对象的宽高比。不易导致细长对象在特征提取时出现偏差。

3. 特征提取 102

本方法针对单册古籍定义和提取三类基本特征，即：页面特征、对象的全局位置特征和形态特征。如果将同一人誊写的多卷古籍组合在一起处理，只需添加书籍标识。上述特征描述了古籍内容。

在模块 102 中，每个对象已从页面图象中分离出来，每个对象都已具备明确的页面内几何坐标和尺寸范围。下面具体定义所述三类基本特征及其提取方法。

定义 1 对象的全局位置特征（GLF）是该对象在一册古籍的页面中的线性序编号。

只要能保证对象与其全局位置特征是 1-1 对应的，定义中的线性序可采取任意形式。例如，全局位置特征的提取方法可按照古籍的誊写习惯（页码从小到大，页内从右向左、各列自上而下），获得由扫描及预处理模块得到的各对象的全局位置特征。对于复杂版面布局，全局位置特征的提取方法可先利用递归曲线如 Hilbert 或 Piano 曲线扫描版面区域，然后各区域内部再按常规方式处理。

定义 2 古籍的页面特征（PF）由页面编号和页面内各对象的几何坐标构成。

页面特征描述了由页面中对象的几何布局关系。

对象的形态特征刻画了对象的视觉语义。进而，除去多音字外，一个汉字的书写唯一决定了该字的语言学语义。换言之，通过对汉字形态的比较，可以实现文字、符号语言学语义的近似匹配。任何中文 OCR 中的汉字特征提取技术均可作为对象形态特征的提取方法。

然而，在以毛笔手书汉字为特征的中文古籍中，存在很多可变因素影响汉字部件及其构成笔划的提取。例如，笔划粗细不均匀、部分笔划模糊或欠落、同一文字的多次出现时笔划/部件间的相对位置偏移、笔划倾角/相对长度变化等，都会影响对象在视觉意义上的匹配。需要开发容错能力较强的特征提取技术。注意到“方块汉字部件部位和比例的固定划一是长期以来汉字书法艺术的结晶”这一事实，以下给出一种在多级质心分划区域中统计笔划因素累计值的形态特征描述及其提取技术。它对中文古籍里存在的上述变化因素有较强的容错能力。

定义 3 对象的形态特征（MF）是其图象在多级质心分划区域中笔划因素分量的累计值。

形态特征的提取方法如下：

首先，根据对象的重心对其 MBS 作多层分划。每个区域的分划点定为该区域中对象前景点（附图中的黑点集）的重心。深一层的分划在浅一层的基础上递归进行。图 13 的一、二层分划的具体方式如图 14 所示。

然后，统计各区域中笔画因素，分类累计后形成特征向量。所谓笔画因素，是指可构成横、竖、撇、捺四种笔画的基本元素，其点阵排列如图 15 所示。相对于完整笔划，基于笔划因素的特征构成对软笔手写汉字笔划不均匀、笔划模糊、倾角/相对长度缺乏规律等现象都具有较强的容错能力，也便于对古籍中非文字符号对象的统一处理。从对象的位图中提取笔划因素方法简便，存在多种实施方案。例如，分别以四种笔画因素为结构元素（Structure Elements），应用数学形态学方法，对图 13 的前景点（图中方框内的黑点）作腐蚀（Erosion）运算，得到四种笔画因素在方框内的分布。将笔划因素的提取方法作用于图 14 中，可得到分划区域里的笔划分布，再用区域中所有前景点的像素数除之，得到笔画因素在各区域中的分布密度。注意到汉字中横竖笔画的出现频度大大高于撇捺笔画，同时为降低特征空间的维数，提高索引及检索的效率，对撇捺笔画因素的统计可以较横竖笔画浅一个层次，即区域分划中可少分解一层。一种具体方式为横、竖笔画因素均用二层区域分划，撇、捺笔画因素均用一层分划。图 17 中例示了两层分划区域中横、竖笔划分布和一层分划区域中撇、捺笔划分布。采用图 16 的区域编号规则，古籍中所有对象的形态特征向量张成了 $16 \times 2 + 4 \times 2 = 40$ 维特征空间。空间中的向量 f 由以下公式计算：

$$f(i) = \sum_{1 \leq k \leq i} \frac{h(k)}{p_2(k)}, \quad f(16+i) = \sum_{1 \leq k \leq i} \frac{s(k)}{p_2(k)}, \quad f(32+j) = \sum_{1 \leq k \leq j} \frac{p(k)}{p_1(k)}, \quad f(36+j) = \sum_{1 \leq k \leq j} \frac{n(k)}{p_1(k)}$$

$$i = 1, 2, \dots, 16 \quad j = 1, 2, 3, 4$$

上式中， $p_1(k)$ 和 $p_2(k)$ 分别为特征提取前位图一级和二级划分区域 k 中的像素点数， $h(k)$ 、 $s(k)$ 、 $p(k)$ 、 $n(k)$ 分别为横、竖、撇、捺笔划因素在位图区域 k 中的黑像素点数。

采用多级质心分划区域笔划因素累计值的对象形态特征，较好地体现了手写字的视觉内容，能以相对灵活的笔划分布密度来表达文字/符号。在特征空间中定义某种度量（或称距离），可形成向量空间。一种度量是公知的欧氏距离。在形成的特征向量空间中，对象的形态特征向量构成了特征空间中点的坐标。因此，形态相似对象的特征点自然形成了聚类，而有差异的汉字的特征点间有较大的距

离。

至此，古籍的特征已提取完毕，古籍页面特征、对象的形态特征和全局位置特征保留至特征表 112。即特征表由多个形如（页面编号、页内几何坐标、全局位置特征、形态特征）的四元组组成，多个的数目是扫描预处理模块 101 确定的对象个数。

4. 特征空间索引 113

实际应用中，生成的特征空间一般具有维数高、特征点数量多等特点。需要设计与应用目标对应的空间索引结构，合理组织所有的特征点，以较小的存储开销换取快速的信息查询。原理上讲，所有的空间索引方法(如 R-树及其改进方法、X-树、SR-树、PK-树等)都能成为特征空间索引结构的实施方案。然而，部分索引算法的性能如 R-树会随空间维数的增大而急剧下降。此处给出 SR-树的优化实施方案。关于 SR-树内部的实现及其性能分析，请参阅相关论文和软件包说明。

A. 数据结构

定义数据项 $E_i=(MF_i, GLF_i)=(f_i, GLF_i)$ 。 f_i 是特征空间中点 i 的坐标，也就是对象 i 的形态特征向量； GLF_i 是对象 i 的全局位置特征。

B. 创建 SR-树

调用函数 `new_HnSRTreeFilePath, Dimension, DataSize, BlockSize, SplitFactor, ReinsertFactor`。生成一棵空 SR-树返回之，返回数据类型 `HnSRTreeFile`。

调用中的输入参数的意义和取值如下表：

参数名	类型	参数意义	取值
Path	字符串	保存 SR-树的数据文件名	古籍名.idx
Dimension	整数	特征空间维数	40
DataSize	整数	特征点相关属性 GLF 字节数	2
BlockSize	整数	数据块大小（字节）	8192（系统缺省值）
SplitFactor	整数	数据库最小利用率（百分之）	40（系统缺省值）
ReinsertFactor	整数	重新插入因子（百分之）	30（系统缺省值）

C. 插入数据项

根据 B.返回的 SR-树对象 `File`，调用其方法 `Store(...)`将数据项 $E_i=(f_i, GLF_i)$ 插

入 SR-树。具体步骤是：

HnSRTreeFile File;

File.Store(Point, Data)。

其中的参数的意义和取值如下表：

参数名	类型	参数意义	取值
Point	HnPoint &	特征空间中点坐标的存放地址	对象的形态特征向量 f
Data	HnData &	特征空间中点属性的存放地址	该对象的 GLF

5. 处理流程控制

古籍处理采用循环方式完成。在一幅页面图象中，对每个对象施行 102 至 113 的处理，一页内的对象是否处理完成在图 4 的 105 中判断。如果本页还有其他对象，则重复上述过程，否则转次页处理。一册古籍是否已完全转化为数字化古籍库 110 在图 4 的 106 中判断。

现结合图 5 说明内容检索 120 处理阶段。内容检索必须在被检索古籍已完成特征空间组织 100 步骤之后进行。对于所建立的一套特征空间索引结构，检索者可执行任意次数的内容检索参见图 3。内容检索的目的，是利用特征空间组织所得到的索引结构，快速获得所有与给定对象视觉内容相似的其他对象。内容检索的基本步骤如下：

(1) 读取精度控制参数 501

检索者通过人机交互方式调整检索精度控制参数。此参数仅代表概念上的“严格”和“宽松”，取值的确定无需任何背景知识。参数取值一般分为多级，各级所对应的距离阈值可由发明实施人按由零到大单调增加方式任意设定。一种实施方式是设定 11 级，第 0 级规定距离阈值为零，表示严格匹配；第 10 级为最宽松的精度控制条件，规定距离阈值为 1；其间按 0.1 的增量逐步增大距离阈值。由于内容检索可多次执行，检索者可参照上次检索结果动态调整精度控制参数，对下一次的查全率和查准率给予新的权衡，满足其需要。精度控制参数影响近似对象查询 123 在特征空间索引 113 中的搜索范围。

(2) 打开起始浏览页面 502

检索者可通过输入任意的页面编号调出相应的页面图象或结合通用的标引方法进入某个页面。直接输入页面编号的方案最为简单。与标引方法配合使用的方

案较为实用。这不仅与图书馆和古籍光盘库现行的检索方式协调一致，而且所形成的二级检索模式更便于处理大量书写风格各异的古籍文献。标引方法提供的检索点引导检索者在数字图书馆或光盘库发现候选的古籍卷宗，基于视觉相似性的内容检索方法进一步为检索者在卷内发现目标提供帮助。

（3）标定检索对象 121

在显示的页面图象上，检索者利用指示设备 209b 如鼠标或手写笔点击对象，设定或调整对象顺序。标定检索样本模块 121 记录指示设备给出的页面编号、页面内几何坐标并根据检索者设定的顺序在页面图象上标记代表该顺序的自然数。可配合浏览控制机制，在多页中标定检索对象。当检索者启动检索时，模块 121 根据上述对象的页面编号、页面内几何坐标序列和坐标序列的顺序形成检索样本。页面编号和坐标集合传给获取特征模块 122，坐标序列的顺序作为约束条件传给验证约束条件模块 124。之后对每个检索样本的成员对象实施 122 至 123 处理，在 506 步骤中进行后处理并判断循环结束。

（4）获取检索样本的形态特征向量 122

根据检索样本之成员对象的页面标号和页面内几何坐标从特征表中获得该对象的形态特征向量。获取方法取决于特征表之页面内几何坐标的组织方式。页面图象经对象分割后，每个对象都有一个包含它的矩形（参见 2（4））。如果对象的页面内几何坐标由该矩形的中点坐标给出，则应该在页号相同的情况下，先在特征表中根据欧氏距离计算与样本成员位置最接近的点，然后再从该表项获得对象的形态特征向量；如果对象的页面内几何坐标由该矩形的对角点坐标给出，则应该在页号相同的情况下，先在特征表中检验矩形是否包含样本成员位置，然后再从包含样本的表项中获得对象的形态特征向量。前一种方法对每个对象节省一对坐标的存储空间，后一种方法在比较时可以避免乘除运算，执行速度较快。在古籍内对象数目较多或检索样本长度一般较短时，使用前一种方法有利。

（5）近似对象查询 123

相对于某个样本成员对象，在特征空间索引中依照最近邻原则查找其视觉相似的对象集合。具体做法是，设由 123 得到的形态向量是 v ，由 501 读取的搜索精度控制参数是 r ，则应用以下的 A~B 得到相似对象的全局位置特征 GLF 的集合。

A. 根据参数 r 设定范围边界。

对特征空间的每一维，设其变动范围是 W ，则首先设定检索范围宽度

$$w = \begin{cases} \varepsilon & r = 0 \\ W \times r/s & 0 < r \leq s \end{cases}$$

其中， ε 是一个十分小的数，一般取值 0.0001，对应严格搜索的情况。 s 是 r 的最大取值。如果按照前述读取精度控制参数步骤中所述， $s=10$ 。

然后，参照图 18 调整检索范围的位置，得到在特征空间该维上一个包含检索点 x 且位于 W 内的区间 w ，使得 x 尽可能地位于的 w 中点。记 w 的边界分别为 a_i 和 b_i 。

利用 SR-树程序包中 HnRect 的方法 SetRange 设定检索范围，即对第 i 维 `rect.SetRange (a_i , HnRange::INCLUSIVE, b_i , HnRange::INCLUSIVE, i)`。

其中，HnRange::INCLUSIVE 是软件包中定义的常数。

B. 范围查找 (Range Search)。

根据 A.中设定的检索范围，逐个从特征空间索引 113 中返回相似对象的全局位置特征 GLF，形成该样本成员的相似对象集合。具体算法如下：

- i)调用 HnSRTreeFile 对象 File 的 GetFirst 方法，返回第一个近似对象的 GLF；
- ii)将该 GLF 并入结果集合
- iii)反复调用 HnSRTreeFile 对象 File 的 GetNext 方法，返回下一个近似对象的 GLF。将该 GLF 并入结果集合，直至返回参数中 Key.isValid()测试为假。

(6) 处理查找结果 123

对检索样本的所有成员对象，它们的近似对象的 GLF 集合被汇集成一簇，传给验证约束条件模块 124。

(7) 验证约束条件 124

所谓约束条件，即是在 121 中检索者所标定对象元素的相对顺序。具体验证过程如下：

- A. 令检索样本包含 M 个成员对象，按其相对顺序依次记以 e_1, e_2, \dots, e_M ，从 506 得到的簇中的 M 个 GLF 表记以 L_1, L_2, \dots, L_M
- B. 将 L_1 作为 L ，用下标 i 从 2 以增量 1 循环至 M ，执行 C
- C. 对 L 中的每一个元素 e ，设其 GLF 为 j ，如果 L_i 中不存在 GLF 为 $j+i-1$ 的对象，则将 e 从 L 中删去

D. 循环结束时 L 中保留的结果就是检索结果的第一个元素列表。

(8) 在页面图象上标记检索结果 508

逐个从 127 的检索结果中取出首元素的 GLF，以此为索引，查找特征表 112，确定检索结果首元素的页面编号和页面内坐标。在页面图象上粘贴附加标记如红色圆点，标示由此开始的连续 M 个对象。当本页从偏移量开始不足 M 个对象时，从次页首部开始标记剩余对象。

(9) 页面图象显示/浏览 125

设立首项标记、前项标记、后项标记、末项标记等跳转按钮，结合普通的首页、前页、后页和末页浏览按钮，提供检索者观察检索结果和观察其上下文的功能。

说明书附图

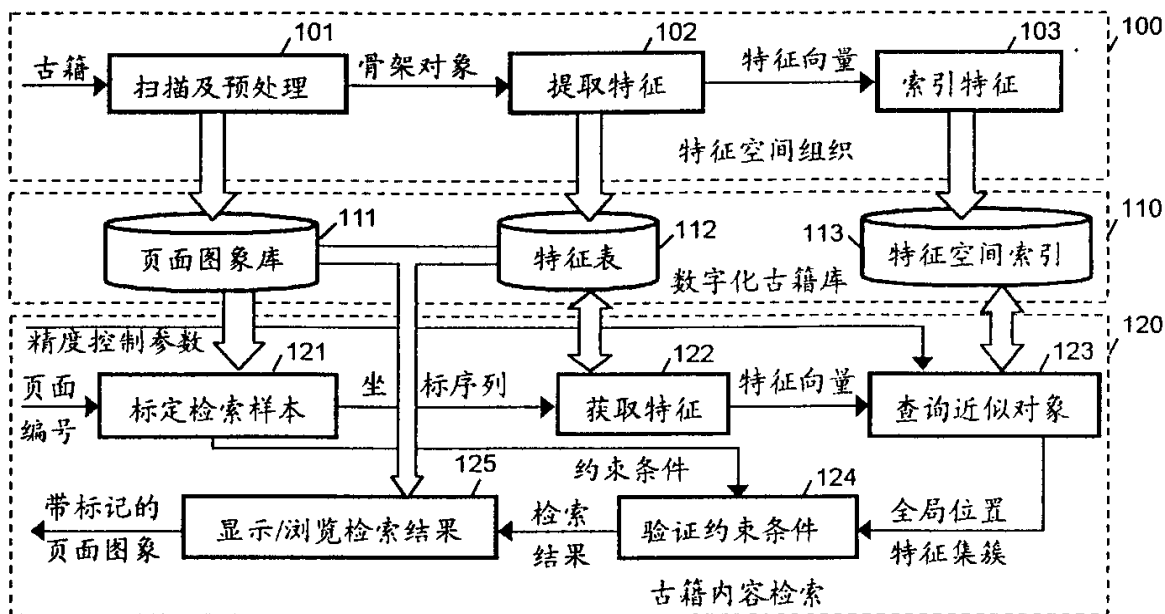


图 1

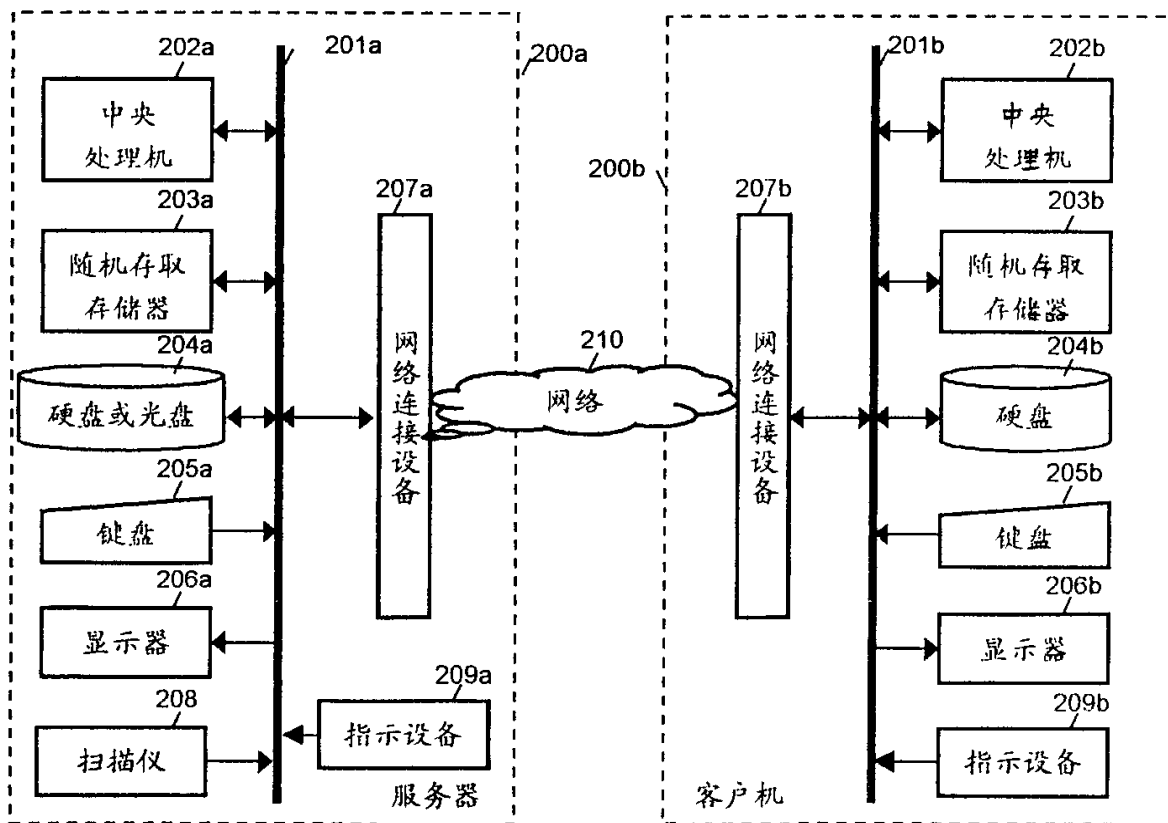


图 2

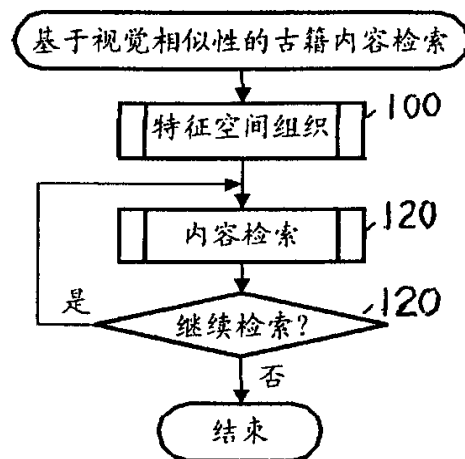


图 3

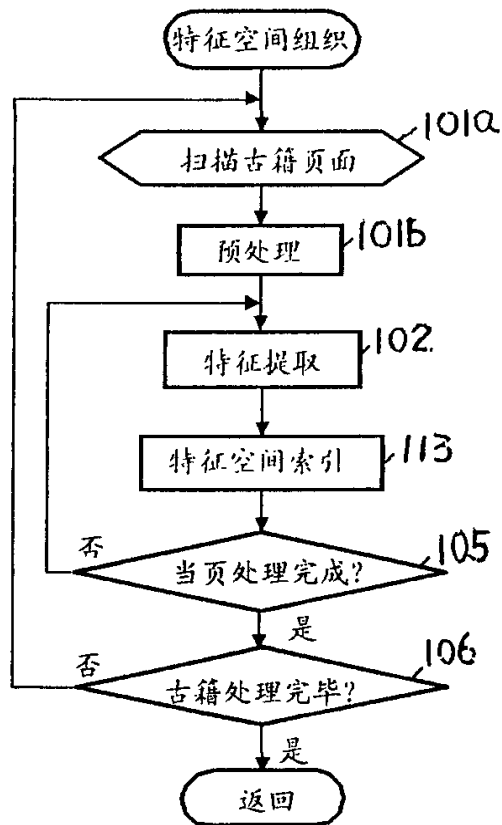


图 4

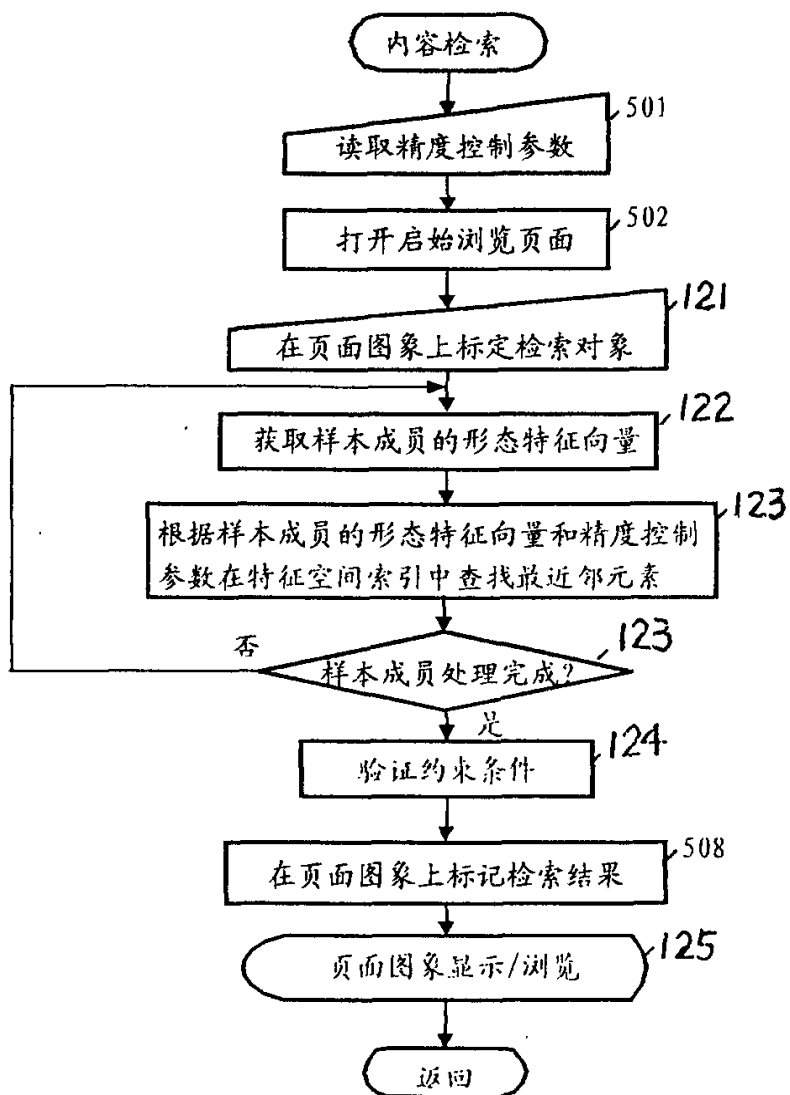


图 5

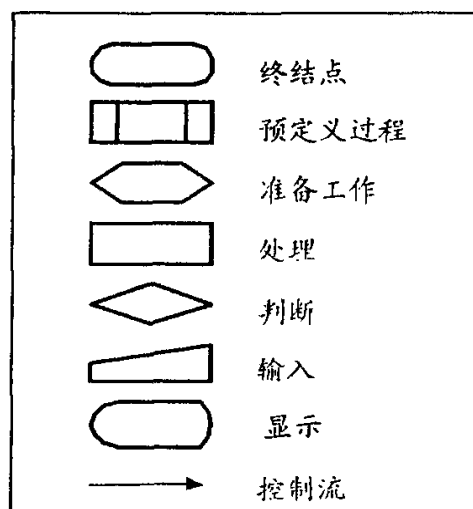


图 6

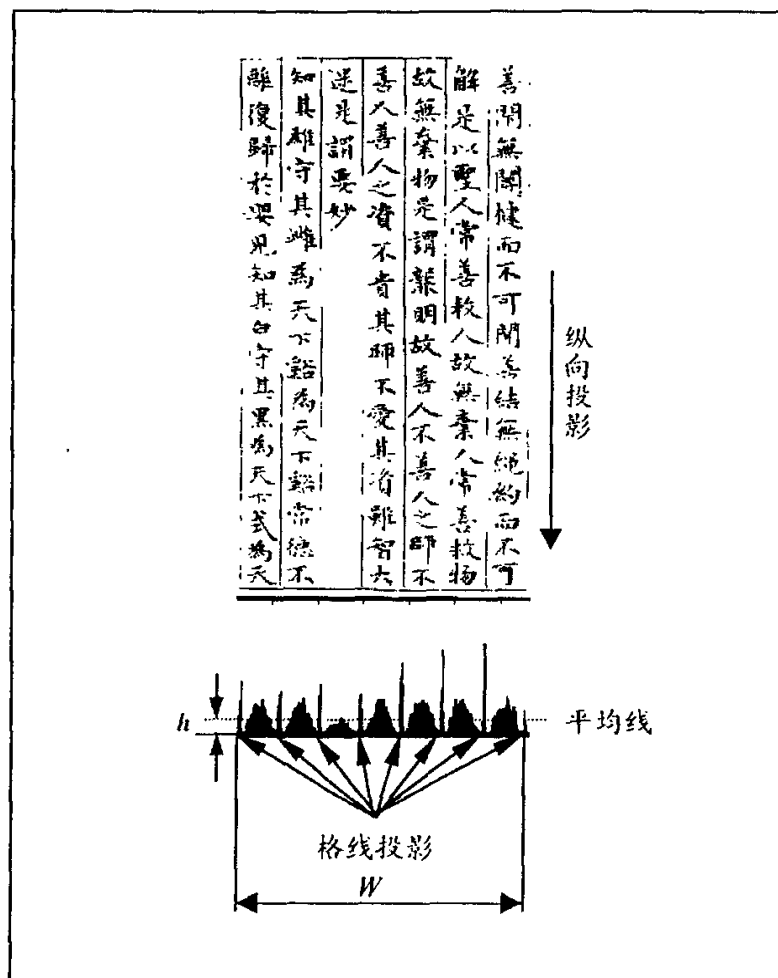


图 7

x_4	x_3	x_2
x_5	x_0	x_1
x_6	x_7	x_8

图 8

善閑無關鍵而不可
解是以聖人常善以
故無棄物是謂嚴
善以善人之資不貴
迷是謂要妙
知其雄守其雌為
離復歸於嬰兒知其

图 9(a)

善閑無關鍵而不可
解是以聖人常善以
故無棄物是謂嚴
善以善人之資不貴
迷是謂要妙
知其雄守其雌為一
離復歸於嬰兒知其

图 9(b)

故無棄物是謂嚴

丟弃

(a)

故無棄物是謂嚴

(b)

图 10



图 11



图 12



图 13

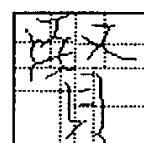
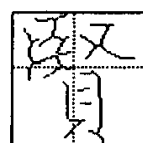


图 14

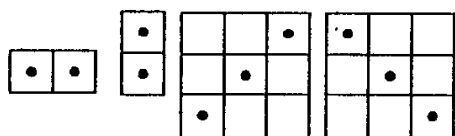


图 15

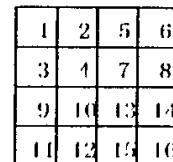
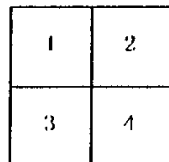


图 16

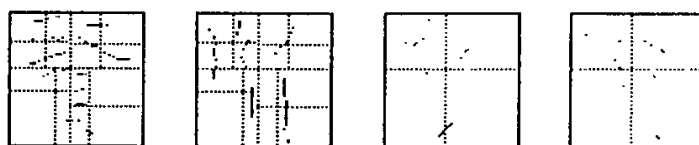


图 17

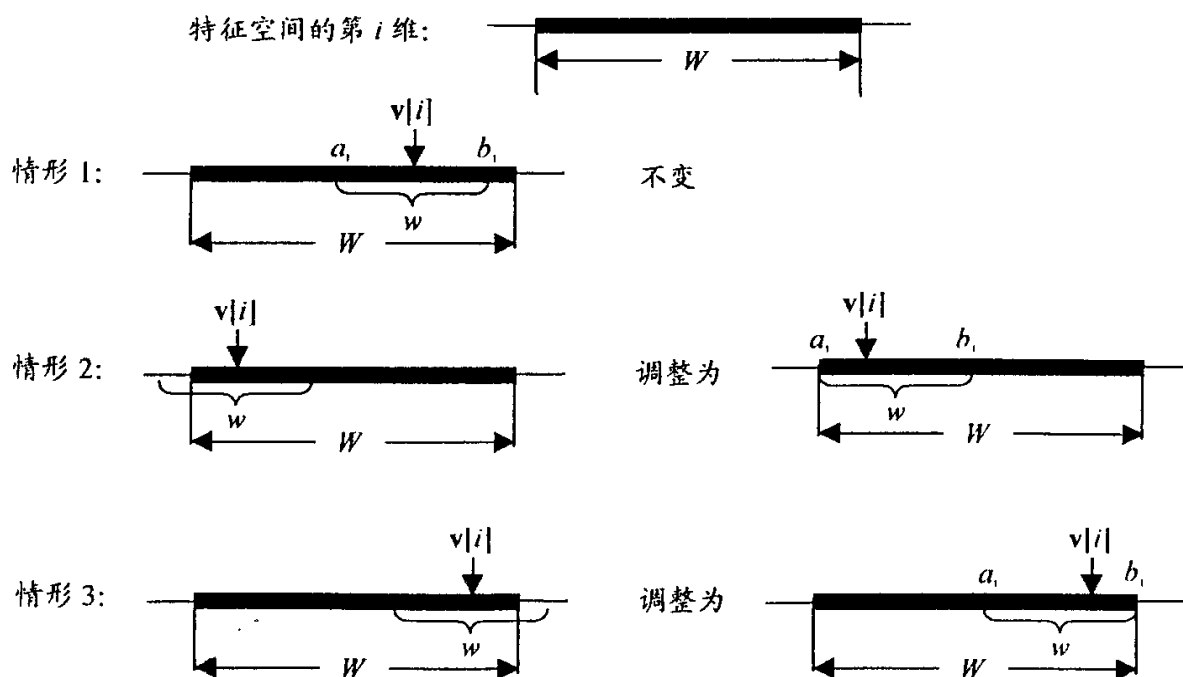


图 18