一种利用WEB方式进行OCR图文识别检索方法和系统

申请号：200910076155.2
申请日：2009-01-09

申请(专利权)人	江阴明伦科技有限公司
地址	214433江苏省江阴市滨江西路2号一号楼1209室
发明(设计)人	凌辉　黄惠良
主分类号	G06F17/30(2006.01)I
分类号	G06F17/30(2006.01)I　G06K9/20(2006.01)I
公开(公告)号	101464903A
公开(公告)日	2009-06-24
专利代理机构	北京路浩知识产权代理有限公司
代理人	张国良

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)
G06K 9/20 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200910076155.2

[43] 公开日 2009 年 6 月 24 日

[11] 公开号 CN 101464903A

[22] 申请日 2009.1.9

[21] 申请号 200910076155.2

[71] 申请人 江阴明伦科技有限公司

地址 214433 江苏省江阴市滨江西路 2 号一号楼 1209 室

[72] 发明人 凌 辉 黄惠良

[74] 专利代理机构 北京路浩知识产权代理有限公司

代理人 张国良

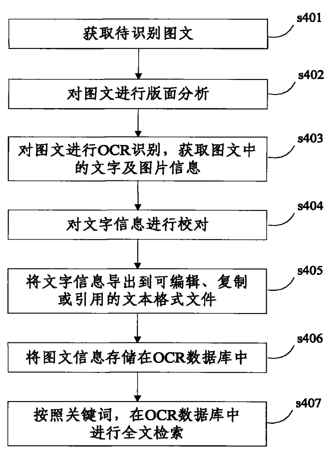
权利要求书 2 页 说明书 7 页 附图 4 页

[54] 发明名称

一种利用 web 方式进行 OCR 图文识别检索方法和系统

[57] 摘要

本发明公开了一种利用 web 方式进行 OCR 图文识别检索方法，所述方法包括以下步骤：获取待识别图文中的文字及图片信息；将所述文字及图片信息存储在 OCR 数据库中；按照关键词，在所述 OCR 数据库中进行全文检索。本发明还公开了一种 OCR 图文识别检索系统，所述系统包括图文信息获取单元、OCR 数据库和检索单元。本发明利用 OCR 图文识别技术，将其高效识别，导出可编辑的文本格式，再利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效地检索出所需要的信息资源。



1、一种利用 web 方式进行 OCR 图文识别检索方法，其特征在于，所述方法包括以下步骤:

A. 获取待识别图文中的文字信息;

B. 将所述文字及图片信息存储在 OCR 数据库中;

C. 按照关键词，在所述 OCR 数据库中进行全文检索。

2、如权利要求 1 所述的 OCR 图文识别检索方法，其特征在于，步骤 A 具体包括:

A1. 获取待识别图文;

A2. 对所述图文进行版面分析;

A3. 对所述图文进行 OCR 识别，获取所述图文中的文字及图片信息。

3、如权利要求 1 所述的利用 web 方式进行 OCR 图文识别检索方法，其特征在于，在步骤 A 之后，还包括:

D. 对所述文字信息进行校对。

4、如权利要求 3 所述的 OCR 图文识别检索方法，其特征在于，步骤 D 具体包括:

D1. 对所述文字信息进行横向校对;

D2. 对所述文字信息进行纵向校对。

5、如权利要求 1 所述的 OCR 图文识别检索方法，其特征在于，在步骤 B 之前，还包括:

E. 将所述文字及图片信息导出到可编辑、复制或引用的文本格式文件。

6、一种 OCR 图文识别检索系统，其特征在于，所述系统包括:

图文信息获取单元，用于获取待识别图文中的文字及图片信息;

OCR 数据库，用于存储所述文字信息;

检索单元，用于按照关键词，在所述 OCR 数据库中进行全文检

索。

7、如权利要求6所述的OCR图文识别检索系统，其特征在于，所述文字信息获取单元包括:

图文获取子单元，用于获取待识别图文;

版面分析子单元，用于对所述图文进行版面分析;

图文识别子单元，用于对所述图文进行OCR识别，获取所述图文中的文字及图片信息。

8、如权利要求7所述的OCR图文识别检索系统，其特征在于，所述图文获取子单元为具备拍摄或扫描功能的设备。

9、如权利要求8所述的OCR图文识别检索系统，其特征在于，所述图文获取子单元为扫描仪、数码相机、一体化机或拍照手机。

10、如权利要求6所述的OCR图文识别检索系统，其特征在于，所述系统还包括校对单元，用于对所述文字信息进行横向校对和纵向校对。

一种利用web方式进行OCR图文识别检索方法和系统

技术领域

本发明涉及图文识别技术领域，特别是涉及一种OCR（Optical Character Recognition，光学字符识别）图文识别检索方法和系统。

背景技术

检索是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术，即从信息集合中找出所需要的信息的过程。

由于对图像文件中的文字不能进行很好地识别，所以对不可随意编辑的图文格式的检索存在很大的困难，这使管理机构面对不同内容的图像格式，显得那么无所适从，不得不花费大量人力、物力，用人工方式重新整理、录入、归类，然后才能统一成某种文本格式再检索。

发明内容

本发明实施例要解决的问题是提供一种利用 web 方式进行 OCR 图文识别检索方法和系统，以克服现有技术中很难对不可随意编辑的图文格式进行检索的缺陷。

为达到上述目的，本发明实施例的技术方案提供一种利用 web 方式进行 OCR 图文识别检索方法，所述方法包括以下步骤：A. 获取待识别图文中的文字及图片信息；B. 将所述文字及图片信息存储在 OCR 数据库中；C. 按照关键词，在所述 OCR 数据库中进行全文检索。

其中，步骤 A 具体包括：A1. 获取待识别图文；A2. 对所述图文进行版面分析；A3. 对所述图文进行 OCR 识别，获取所述图文中的文字及图片信息。

其中，在步骤 A 之后，还包括：D. 对所述文字信息进行校对。

其中，步骤 D 具体包括：D1. 对所述文字信息进行横向校对；D2. 对所述文字信息进行纵向校对。

其中，在步骤 B 之前，还包括：E. 将所述文字及图片信息导出到可编辑、复制或引用的文本格式文件。

本发明实施例的技术方案还提供一种 OCR 图文识别检索系统，所述系统包括：图文信息获取单元，用于获取待识别图文中的文字及图片信息；OCR 数据库，用于存储所述文字及图片信息；检索单元，用于按照关键词，在所述 OCR 数据库中进行全文检索。

其中，所述文字信息获取单元包括：图文获取子单元，用于获取待识别图文；版面分析子单元，用于对所述图文进行版面分析；图文识别子单元，用于对所述图文进行 OCR 识别，获取所述图文中的文字信息。

其中，所述图文获取子单元为具备拍摄或扫描功能的设备。

其中，所述图文获取子单元为扫描仪、数码相机、一体化机或拍照手机。

其中，所述系统还包括校对单元，用于对所述文字信息进行横向校对和纵向校对。

与现有技术相比，本发明的技术方案具有如下优点：

本发明实施例利用 OCR 图文识别技术，将其高效识别，导出可编辑的文本格式，再利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效地检索出所需要的信息资源。

附图说明

图1是本发明实施例的一种利用web方式进行OCR图文识别检索方法的流程图；

图2是本发明实施例的另一种利用web方式进行OCR图文识别检索方法的流程图；

图3是本发明实施例的另一种利用web方式进行OCR图文识别检

索方法的流程图

图4是本发明实施例的另一种利用web方式进行OCR图文识别检索方法的流程图

图5是本发明实施例的一种利用web方式进行OCR图文识别检索系统的结构图。

具体实施方式

下面结合附图和实施例，对本发明的具体实施方式作进一步详细描述。以下实施例用于说明本发明，但不用来限制本发明的范围。

实施例一

本发明实施例的一种利用web方式进行OCR图文识别检索方法如图1所示，包括以下步骤:

步骤 s101，获取待识别图文。本实施例通过扫描仪、数码相机、一体化机、拍照手机等任何具备拍摄、扫描功能的设备获取待识别图文。

步骤 s102，对所述图文进行版面分析。

步骤s103，对所述图文进行OCR识别，获取所述图文中的文字及图片信息。

步骤 s104，将所述文字及图片信息存储在 OCR 数据库中。

步骤s105，按照关键词，在所述OCR数据库中进行全文检索。本实施例利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效的检索出所需要的信息资源。

实施例二

本发明实施例的一种利用web方式进行OCR图文识别检索方法如图2所示，包括以下步骤:

步骤 s201，获取待识别图文。本实施例通过扫描仪、数码相机、一体化机、拍照手机等任何具备拍摄、扫描功能的设备获取待识别图文。

步骤 s202，对所述图文进行版面分析。

步骤s203，对所述图文进行OCR识别，获取所述图文中的文字及图片信息。

步骤s204，对所述文字信息进行校对。本实施例对复杂版面进行自动分析，智能分析各种混排格式的文本，针对识别文件实行横向和纵向全面校队，无需过多人工干预。

步骤 s205，将所述文字及图片信息存储在 OCR 数据库中。

步骤s206，按照关键词，在所述OCR数据库中进行全文检索。本实施例利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效的检索出所需要的信息资源。

实施例三

本发明实施例的一种利用web方式进行OCR图文识别检索方法如图3所示，包括以下步骤:

步骤 s301，获取待识别图文。本实施例通过扫描仪、数码相机、一体化机、拍照手机等任何具备拍摄、扫描功能的设备获取待识别图文。

步骤 s302，对所述图文进行版面分析。

步骤s303，对所述图文进行OCR识别，获取所述图文中的文字及图片信息。

步骤s304，将所述文字及图片信息导出到可编辑、复制或引用的文本格式文件。本实施例中，所述文本格式文件包括word、rtf等多种可编辑、复制和引用的文本格式文件。

步骤 s305，将所述文字信息存储在 OCR 数据库中。

步骤s306，按照关键词，在所述OCR数据库中进行全文检索。本实施例利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效的检索出所需要的信息资源。

实施例四

本发明实施例的一种利用web方式进行OCR图文识别检索方法如图4所示，包括以下步骤:

步骤 s401，获取待识别图文。本实施例通过扫描仪、数码相机、一体化机、拍照手机等任何具备拍摄、扫描功能的设备获取待识别图文。

步骤 s402，对所述图文进行版面分析。

步骤s403，对所述图文进行OCR识别，获取所述图文中的文字及图片信息。

步骤s404，对所述文字信息进行校对。本实施例对复杂版面进行自动分析，智能分析各种混排格式的文本，针对识别文件实行横向和纵向全面校队，无需过多人工干预。

步骤s405，将所述文字及图片信息导出到可编辑、复制或引用的文本格式文件。本实施例中，所述文本格式文件包括word、rtf等多种可编辑、复制和引用的文本格式文件。

步骤 s406，将所述文字及图片信息存储在 OCR 数据库中。

步骤s407，按照关键词，在所述OCR数据库中进行全文检索。本实施例利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效的检索出所需要的信息资源。

本发明实施例的一种利用 web 方式进行 OCR 图文识别检索系统如图 5 所示，包括图文信息获取单元、校对单元、OCR 数据库和检索单元。其中，校对单元分别与图文信息获取单元和 OCR 数据库连接，检索单元与 OCR 数据库连接。

图文信息获取单元用于获取待识别图文中的文字及图片信息；校对单元用于对所述文字信息进行横向校对和纵向校对；OCR 数据库用于存储所述文字及图片信息；检索单元用于按照关键词，在所述 OCR 数据库中进行全文检索。

图文信息获取单元包括图文获取子单元、版面分析子单元和图文

识别子单元，其中版面分析子单元分别与图文获取子单元和图文识别子单元连接。

图文获取子单元为具备拍摄或扫描功能的设备，用于获取待识别图文，可以是扫描仪、数码相机、一体化机或拍照手机等；版面分析子单元用于对所述图文进行版面分析；图文识别子单元用于对所述图文进行 OCR 识别，获取所述图文中的文字及图片信息。

本发明将不可随意编辑的图文格式资料，依托 OCR 研发技术的优势，将其随意导出到指定的 word、rtf 等多种可编辑、复制和引用的文本格式文件，经处理后可将图像文字存储于数据库中，便于大量文档的存储、管理、共享、传输和检索。本发明识别准确率高，鲁棒性强，无缝整合了版面分析、图像识别、智能识别和全文检索的全过程。本发明可以通过扫描仪、数码相机、一体化机、拍照手机等任何具备拍摄、扫描功能的设备，随时随地的对图像文件中的图文进行 OCR 识别，现有的 OCR 产品都是软硬件结合在一起的，而本发明摆脱了硬件的束缚，实现了单一软件和多种硬件的随意结合，充分利用现有的设备，完成繁琐的录入、整理及后期的文档共享及检索工作。

本发明利用 OCR 图文识别技术，将其高效识别，导出可编辑的 word、rtf 等文本格式，再利用全文检索技术，通过输入嵌入在图片资料里的文字，即可方便高效的检索出所需要的信息资源，从而能够快捷、高效、精准的完成对图像格式的智能识别，充分满足了管理人员、办公人员等不同需求的录入工作，为其节省了大量的时间，提高了效率。

本发明对于复杂版面可以进行自动分析，智能分析各种混排格式的文本，针对识别文件实行横向和竖向全面校队，无需过多人工干预。而且，本发明可以进行版面还原，精确保留了原版面格式，准确恢复文本原貌。本发明具有强大的公文处理能力，能够准确再现公文原貌。本发明实现了单一软件和多种硬件的随意结合，充分利用现有的设

备，完成繁琐的录入、整理工作。本发明利用全文检索技术，输入嵌入在图片资料里的文字信息，即可快捷、高效的查找到所需要的图文资料

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明技术原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

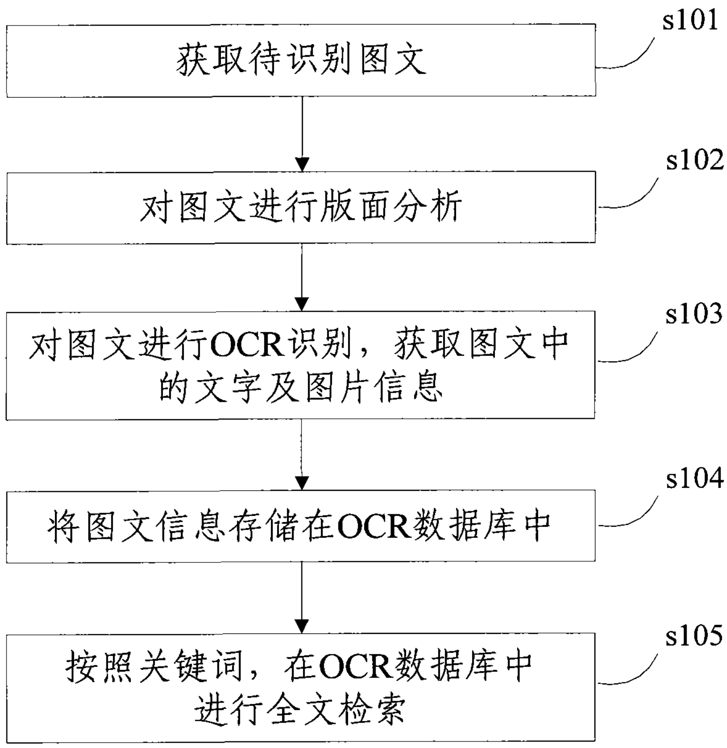

获取待识别图文
s101
对图文进行版面分析
s102
对图文进行OCR识别，获取图文中的文字及图片信息
s103
将图文信息存储在OCR数据库中
s104
按照关键词，在OCR数据库中进行全文检索
s105

图 1

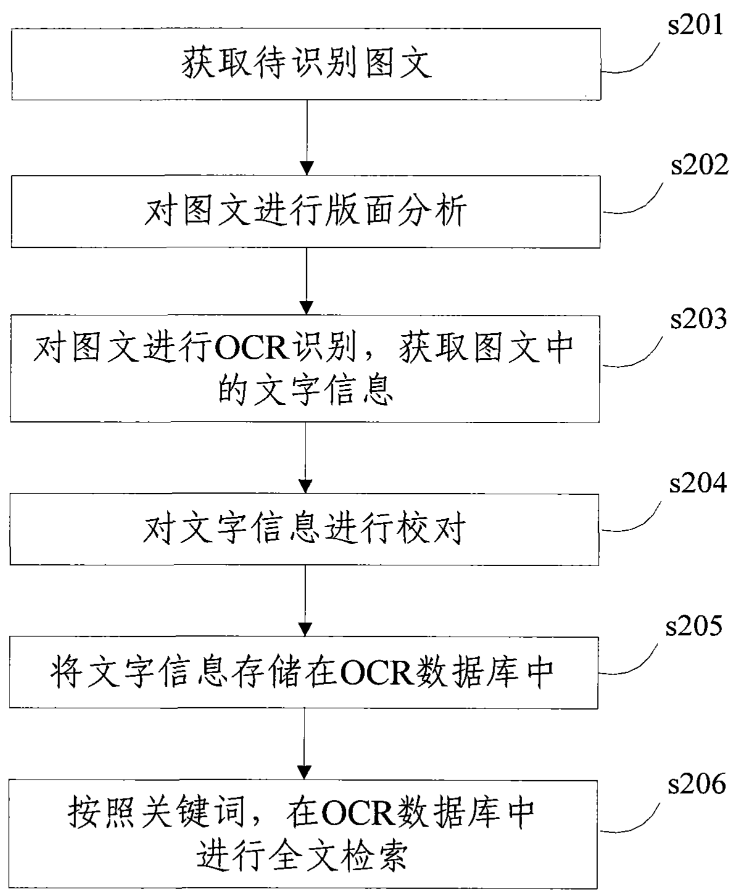

s201
获取待识别图文
s202
对图文进行版面分析
s203
对图文进行OCR识别，获取图文中的文字信息
s204
对文字信息进行校对
s205
将文字信息存储在OCR数据库中
s206
按照关键词，在OCR数据库中进行全文检索

图 2

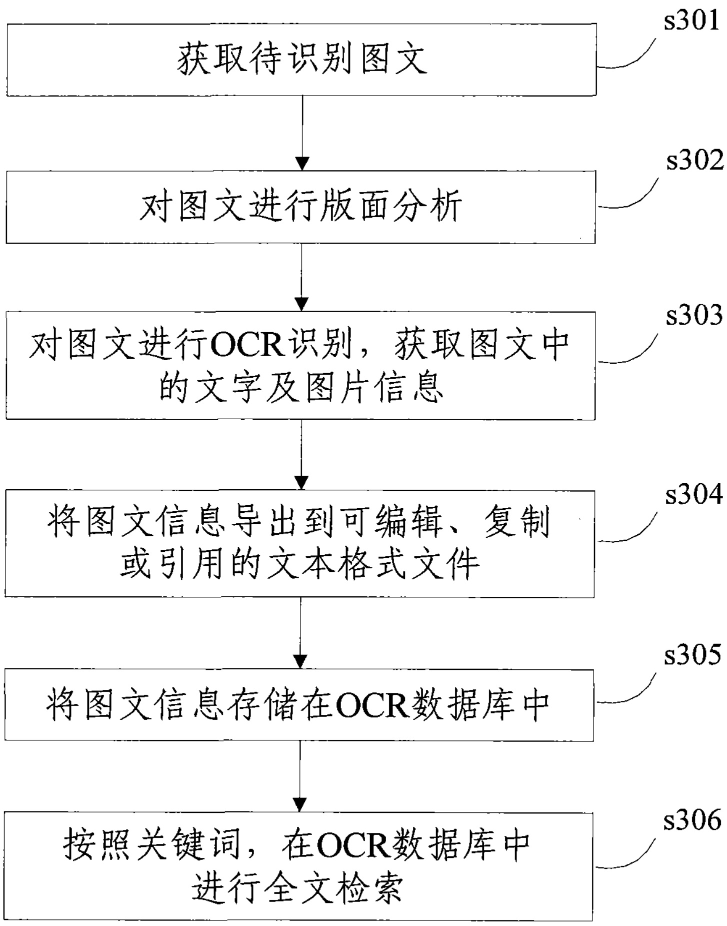

获取待识别图文
s301
对图文进行版面分析
s302
对图文进行OCR识别，获取图文中的文字及图片信息
s303
将图文信息导出到可编辑、复制或引用的文本格式文件
s304
将图文信息存储在OCR数据库中
s305
按照关键词，在OCR数据库中进行全文检索
s306

图 3

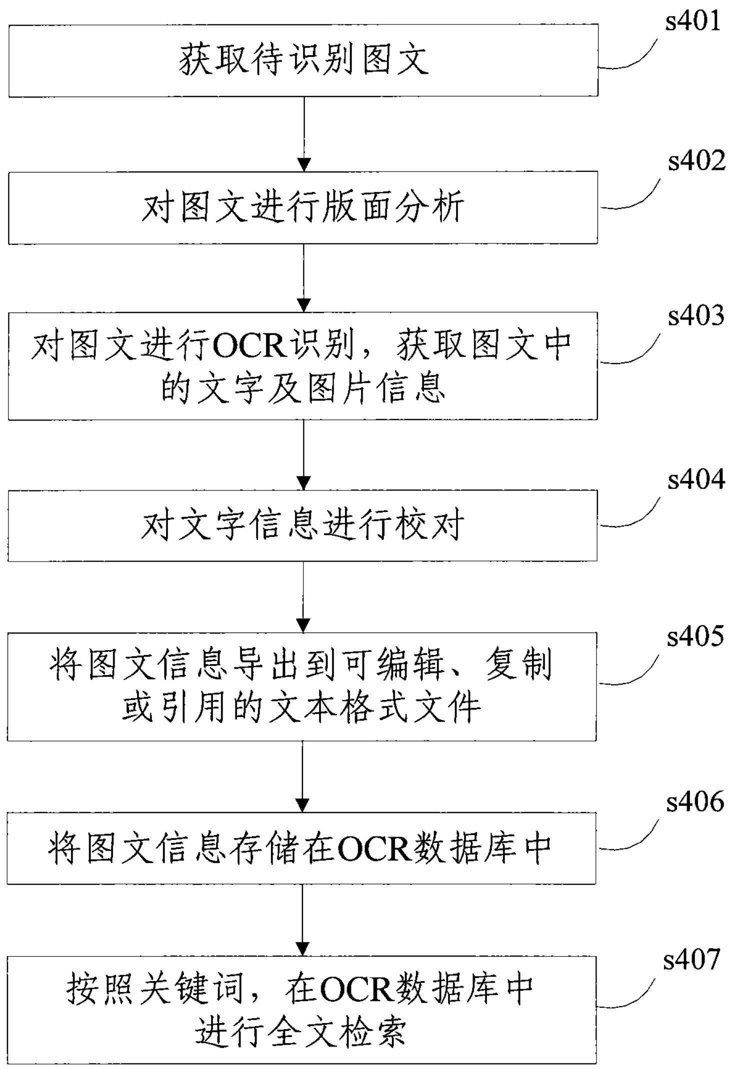

获取待识别图文
s401
对图文进行版面分析
s402
对图文进行OCR识别，获取图文中的文字及图片信息
s403
对文字信息进行校对
s404
将图文信息导出到可编辑、复制或引用的文本格式文件
s405
将图文信息存储在OCR数据库中
s406
按照关键词，在OCR数据库中进行全文检索
s407

图 4

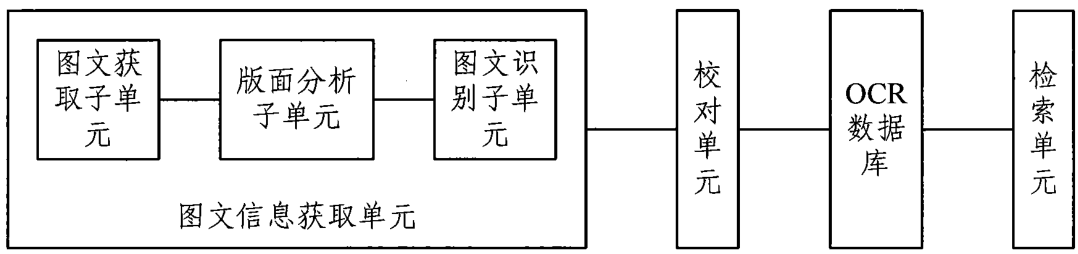

图文获取子单元
版面分析子单元
图文识别子单元
图文信息获取单元
校对单元
OCR数据库
检索单元

图 5