(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102467653 A
(43)申请公布日 2012.05.23

(21)申请号 201010530660.2

(22)申请日 2010.10.29

(71)申请人 方正国际软件(北京)有限公司
地址 100080 北京市海淀区北四环西路 52 号中芯大厦 19 层
申请人 方正国际软件有限公司

(72)发明人 吴建宇

(74)专利代理机构 北京天悦专利代理事务所（普通合伙）11311
代理人 田明 任晓航

(51)Int.Cl.
G06K 9/00(2006.01)
G06F 17/30(2006.01)

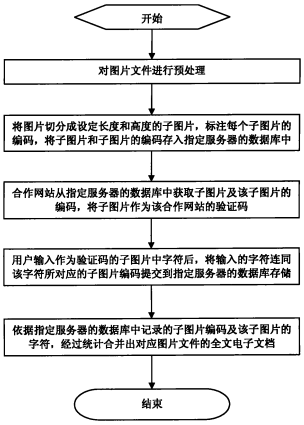
权利要求书 2 页 说明书 4 页 附图 2 页

(54)发明名称

一种图文识别方法及系统

(57)摘要

本发明涉及一种图文识别方法及系统，属于图像识别技术领域。本发明首先对图片文件进行预处理后将图片切分成设定长度和高度的子图片，并标注每个子图片的编码；合作网站从指定服务器的数据库中获取子图片及该子图片的编码，将子图片作为用户登录该合作网站的验证码；在合作网站用户输入验证码子图片中的字符后，将输入的字符连同该字符所对应的子图片编码提交到指定服务器的数据库存储；最后依据数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应扫描图片文件的全文电子文档。本发明通过利用网民的用户身份验证功能，经过统计完成对图片中文字识别，特别适用于古文、手书文字识别。



CN 102467653 A

1. 一种图文识别方法，包括以下步骤：

(1) 对图片文件进行预处理，删除图片中非字符内容；

(2) 将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入指定服务器的数据库中；

(3) 合作网站获取指定服务器的数据库中的子图片及该子图片的编码，将子图片作为该合作网站的验证码；

(4) 合作网站用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到指定服务器的数据库存储；

(5) 依据指定服务器的数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应图片文件的全文电子文档。

2. 如权利要求 1 所述的图文识别方法，其特征在于：步骤 (2) 中所述切分的顺序与文档中的字符排列顺序相同。

3. 如权利要求 2 所述的图文识别方法，其特征在于：步骤 (2) 中所述切分的方法如下：按照字符行 / 列及行 / 列方向，先将图片切分成若干字符行 / 列，再将每个字符行 / 列按照设定长度和高度切分成若干子图片。

4. 如权利要求 3 所述的图文识别方法，其特征在于：步骤 (2) 中所述子图片的编码包括该子图片所在字符行 / 列数、及其在行 / 列中的序号。

5. 如权利要求 4 所述的图文识别方法，其特征在于：所述子图片编码中还包含该子图片的归属信息。

6. 如权利要求 5 所述的图文识别方法，其特征在于：步骤 (2) 中将子图片编码作为该子图片的文件名。

7. 如权利要求 1 至 6 之一所述的图文识别方法，其特征在于：步骤 (2) 中所述设定长度为图片中 2 至 6 个连续字符的平均长度，所述设定高度大于字符的高度。

8. 如权利要求 7 所述的图文识别方法，其特征在于：如果某行 / 列长度不能被整切分，则最后一个子图片长度可以小于设定长度。

9. 如权利要求 1 至 6 之一所述的图文识别方法，其特征在于：步骤 (3) 中所述合作网站获取指定服务器的数据库中的子图片及该子图片的编码方法如下：

合作网站主动向指定服务器请求子图片及该子图片的编码；或者指定服务器向合作网站分发子图片及该子图片的编码，合作网站被动接受。

10. 如权利要求 9 所述的图文识别方法，其特征在于：合作网站获取指定服务器子图片的优先顺序如下：

优先获取数据库中识别内容相似度低的子图片，在同等情况下，获取识别次数少的子图片，再同等条件下，在该范围内随机分发；

所述相似度是指对于同一子图片，多次输入文字内容中相同的次数与输入的总次数的比值。

11. 如权利要求 1 至 6 之一所述的图文识别方法，其特征在于：步骤 (4) 中还包括统计各合作网站用户输入的字符数量的步骤，依据字符数量向合作网站付费。

12. 如权利要求 1 至 6 之一所述的图文识别方法，其特征在于：步骤 (5) 中，合并对应图片文件的全文电子文档时，对于每个子图片，以用户输入相同次数最多的字符作为该子

图片的字符。

13. 一种图文识别系统，包括：

预处理装置 (11)，用于对图片文件进行预处理，删除图片中非字符内容；

切分装置 (12)，用于将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入服务器 (17) 的数据库中；

合作网站 (13)，用于获取服务器 (17) 的数据库中的子图片及该子图片的编码，将子图片作为该合作网站的验证码；

存储装置 (14)，用于在合作网站 (13) 用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到服务器 (17) 的数据库存储；

合成装置 (15)，用于依据服务器 (17) 的数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应图片文件的全文电子文档。

14. 如权利要求 13 所述的图文识别系统，其特征在于：所述系统还包括用于统计各合作网站用户输入的字符数量，并根据字符数量计算费用的费用清分装置 (16)。

一种图文识别方法及系统

技术领域

[0001] 本发明属于图像识别技术领域，具体涉及一种图片中字符的识别方法及系统。

背景技术

[0002] 信息化使文化存储、传播等方式产生了变革性的变化，但是目前还存在大量的纸质材料，简单的文件扫描生成图片不适用于内容检索等，因此，目前大量的纸质的书籍、档案需要进行OCR文字识别，OCR(Optical CharacterRecognition，光学字符识别）技术从扫描件中识别出字符及版面信息等内容，从而将纸质图书转化成了数字内容。OCR技术对于规范印刷字体的文字识别效果好，但是对于大量古籍文献以及手写内容，采用OCR技术识别的准确率很低，无法满足数字化处理的要求。如果采用人工方法对OCR识别后的字符进行校对，则工作量太大，成本太高。

[0003] 另外，由于现有纸质资料的文字识别，人工识别也存在一定比例的错识率，或者是有争议的，特别是古籍中的缺笔、变体等，往往需要结合多次识别以及前后文分析才能有较准确的识别。这种工作方式也决定了其巨大的工作量。

[0004] 目前验证码是互联网常用的一种防止对某一个特定注册用户用特定程序暴力破解方式进行不断的登陆尝试，实际上使用验证码是现在很多网站通行的方式。所谓验证码，就是将一串随机产生的数字或文字，生成一幅图片，图片里加上一些干扰像素（防止OCR)，由用户肉眼识别其中的验证码信息，输入表单提交网站验证，验证成功后才能使用某项功能。

[0005] 本发明正是在这种背景下，利用该反向思维，希望通过技术手段，结合目前广大的互联网用户验证码需求，通过广大网民零散的、海量的验证，实现对OCR识别率低的部分进行人工识别补充，从而解决目前古籍、手书的纸质文字的数字化。

发明内容

[0006] 针对现有技术中存在的缺陷，本发明要解决的技术问题是提供一种准确率和效率高的图文识别方法与系统。

[0007] 为解决上述技术问题，本发明采用的技术方案如下：

[0008] 一种图文识别方法，包括以下步骤：

[0009] (1) 对图片文件进行预处理，删除图片中非字符内容；

[0010] (2) 将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入指定服务器的数据库中；

[0011] (3) 合作网站获取指定服务器的数据库中的子图片及该子图片的编码，将子图片作为该合作网站的验证码；

[0012] (4) 合作网站用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到指定服务器的数据库存储；

[0013] (5) 依据指定服务器的数据库中记录的子图片编码及该子图片的字符，经过统计

合并出对应图片文件的全文电子文档。

[0014] 如上所述的图文识别方法，步骤 (2) 中所述切分的顺序与文档中的字符排列顺序相同。

[0015] 如上所述的图文识别方法，步骤 (2) 中所述切分的方法如下：按照字符行 / 列及行 / 列方向，先将图片切分成若干字符行 / 列，再将每个字符行 / 列按照设定长度和高度切分成若干子图片。

[0016] 如上所述的图文识别方法，步骤 (2) 中所述子图片的编码包括该子图片所在字符行 / 列数、及其在行 / 列中的序号。子图片编码中还包含该子图片的归属信息。

[0017] 如上所述的图文识别方法，步骤 (2) 中将子图片编码作为该子图片的文件名。

[0018] 如上所述的图文识别方法，步骤 (2) 中所述设定长度为图片中 2 至 6 个连续字符的平均长度，所述设定高度大于字符的高度。如果某行 / 列长度不能被整切分，则最后一个子图片长度可以小于设定长度。

[0019] 如上所述的图文识别方法，步骤 (3) 中所述合作网站获取指定服务器的数据库中的子图片及该子图片的编码方法如下：

[0020] 合作网站主动向指定服务器请求子图片及该子图片的编码；或者指定服务器向合作网站分发子图片及该子图片的编码，合作网站被动接受。

[0021] 合作网站获取指定服务器子图片的优先顺序如下：

[0022] 优先获取数据库中识别内容相似度低的子图片，在同等情况下，获取识别次数少的子图片，再同等条件下，在该范围内随机分发；

[0023] 所述相似度是指对于同一子图片，多次输入文字内容中相同的次数与输入的总次数的比值。

[0024] 如上所述的图文识别方法，步骤 (4) 中还包括统计各合作网站用户输入的字符数量的步骤，依据字符数量向合作网站付费。

[0025] 如上所述的图文识别方法，步骤 (5) 中，合并对应图片文件的全文电子文档时，对于每个子图片，以用户输入相同次数最多的字符作为该子图片的字符。

[0026] 一种图文识别系统，包括：

[0027] 预处理装置，用于对图片文件进行预处理，删除图片中非字符内容；

[0028] 切分装置，用于将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入服务器的数据库中；

[0029] 合作网站，用于获取服务器的数据库中的子图片及该子图片的编码，将子图片作为用户登录该合作网站的验证码；

[0030] 存储装置，用于在合作网站用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到服务器的数据库存储；

[0031] 合成装置，用于依据服务器的数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应图片文件的全文电子文档。

[0032] 如上所述的图文识别系统，还包括用于统计各合作网站用户输入的字符数量，并根据字符数量计算费用的费用清分装置。

[0033] 本发明所述的方法及系统，由于将图片中字符的大量人工识别工作进行了分散，分散给了千千万万的网站用户，不仅大大提高了识别效率，而且也大大提高了识别的准确

性。

附图说明

[0034] 图 1 是具体实施方式中图文识别系统的结构框图；

[0035] 图 2 是具体实施方式中图文识别方法的流程图。

具体实施方式

[0036] 下面结合具体实施方式和附图对本发明进行详细描述。

[0037] 图 1 示出了本实施方式中图文识别系统的结构。如图 1 所示，该系统包括预处理装置 11，与预处理装置 11 连接的切分装置 12，与切分装置 12 连接的服务器 17，与服务器 17 连接的合作网站群 13、存储装置 14 和合成装置 15，以及与合作网站群 13 连接的费用清分装置 16。其中，合作网站群 13 可以包括多个合作网站。

[0038] 预处理装置 11 用于对图片文件进行预处理，删除图片中的非字符内容。

[0039] 切分装置 12 用于将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入服务器 17 的数据库中。

[0040] 合作网站 13 用于获取服务器 17 的数据库中的子图片及该子图片的编码，将子图片作为该合作网站的验证码。

[0041] 存储装置 14 用于在合作网站 13 用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到服务器 17 的数据库存储。

[0042] 合成装置 15 用于依据服务器 17 的数据库中记录的子图片编码及该子图片的字符，合并出对应图片文件的全文电子文档。

[0043] 费用清分装置 16 用于统计各合作网站用户输入的字符数量，并根据字符数量计算费用。

[0044] 图 2 示出了采用图 1 所示系统识别图文的方法流程。如图 2 所示，该方法包括以下步骤：

[0045] (1) 预处理装置 11 对图片文件进行预处理，删除图片中非字符内容。

[0046] 图片文件可以是纸质资料的扫描件，采用现有的抠图工具将图片中的非字符内容删除，如图像、表格等内容。

[0047] (2) 切分装置 12 将图片切分成设定长度和高度的子图片，并标注每个子图片的编码，将子图片连同该子图片的编码存入指定服务器的数据库中。

[0048] 图片切分的顺序与文档中的字符排列顺序相同。

[0049] 如果字符排列顺序为从左到右、从上到下，则按照字符行及行方向，先将图片切分成若干字符行，再将每个字符行按照设定长度和高度切分成若干子图片。

[0050] 如果字符排列顺序为从上到下、从右到左，则按照字符列及列方向，先将图片切分成若干字符列，再将每个字符列按照设定长度和高度切分成若干子图片。

[0051] 较佳地，所述设定长度为图片中 2 至 6 个连续字符的平均长度，设定高度稍大于字符的高度。如果某行或某列长度不能被整切分，则最后一个子图片长度可以小于设定的长度。

[0052] 所述子图片的编码中包含该子图片的归属信息，该子图片所在字符行 / 列数、及

其在行 / 列中的序号。例如，假设某子图片为某书籍（编码为 001）中第 1 页第 2 行第 4 个子图片，则该子图片的编码可以为 :0010010204。第 1 ～ 3 位数字表示该子图片所属书籍的编号，第 4 ～ 6 位表示所属页码，第 7 ～ 8 位表示所属行数，低 9 ～ 10 位表示所在行的子图片序号。优选地，将子图片的编码作为子图片的名称存储，这样可以很方便的存储子图片的编码。

[0053] (3) 合作网站获取指定服务器的数据库中的子图片及该子图片的编码，将子图片作为该合作网站的验证码。

[0054] 在用户登录网站、发帖、身份认证等场合，通常需要输入验证码。将子图片作为验证码，在需要输入验证码时，用户输入子图片中的字符。这样可以将图片中的字符识别工作分散给千千万万的网络用户，从而增加图片中字符的识别效率。

[0055] 合作网站获取指定服务器的数据库中的子图片及子图片编码的方法可以采用如下两种方式之一 :①合作网站主动向指定服务器请求子图片及该子图片的编码 ;②指定服务器向合作网站分发子图片及该子图片的编码，合作网站被动接收。

[0056] 合作网站不论主动还是被动从指定服务器的数据库中获得子图片和子图片编码，其获得的优先顺序为 :优先获取数据库中识别内容相似度低的子图片，在同等情况下，获取识别次数少的子图片，再同等条件下，在该范围内随机分发。所述相似度是指对于同一子图片，多次输入文字内容中相同的次数与输入的总次数的比值。例如，假设某子图片字符被用户输入过 100 次，其中，50 个用户输入的字符相同，则该子图片的相似度为 50%。相似度低的子图片表明该子图片中的字符难识别，可以通过增加用户识别次数的方式提高识别的准确率。

[0057] 在合作网站用户输入子图片中字符后，费用清分装置 16 统计该合作网站用户输入的字符数量，依据字符数量向合作网站付费。

[0058] (4) 存储装置 14 在合作网站用户输入作为验证码的子图片中的字符后，将输入的字符连同该字符所对应的子图片编码提交到指定服务器的数据库存储。

[0059] (5) 合成装置 15 依据指定服务器的数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应图片文件的全文电子文档。

[0060] 在合并时，对于每个子图片，以用户输入相同次数最多的字符作为该子图片的字符。例如，对于含有“个人”字符的子图片，如果有 100 个用户输入了该子图片中的字符，90 个用户输入“个人”，10 个用户输入了“人人”，则将“个人”作为该子图片包含的字符。这样，可以提高字符识别的准确性。

[0061] 本发明将纸质文献扫描后生成的图片文件中的字符内容进行切分，将切分后的子图片作为合作网站的验证码，从而在网站用户输入验证码时，实现了子图片中字符的识别。数以千万计的互联网用户可以在上网的同时进行文献字符数字化的转换处理，转换后的字符根据子图片编码进行拼接合成。这种方法大大的提高了文献数字化的处理效率，而且可以保证文献中字符识别的准确性，为数字化图书馆的建设提供了强有力的技术保障。

[0062] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其同等技术的范围之内，则本发明也意图包含这些改动和变型在内。

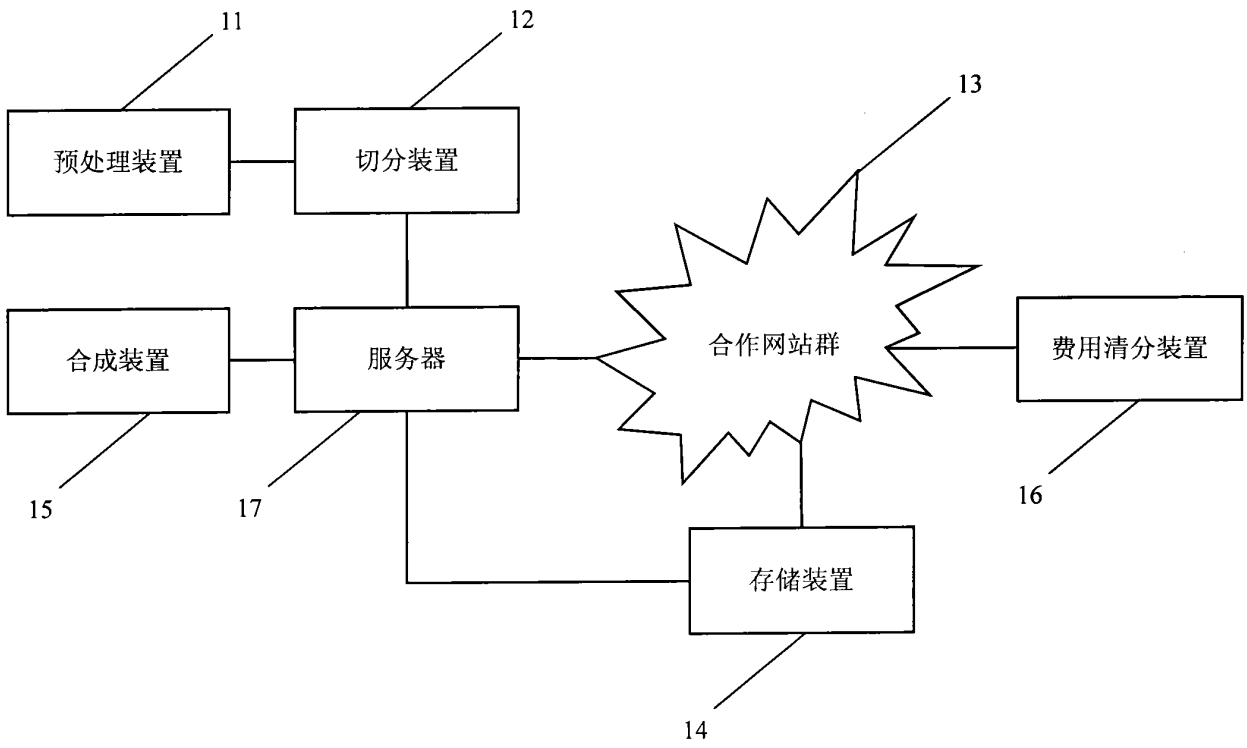


图 1

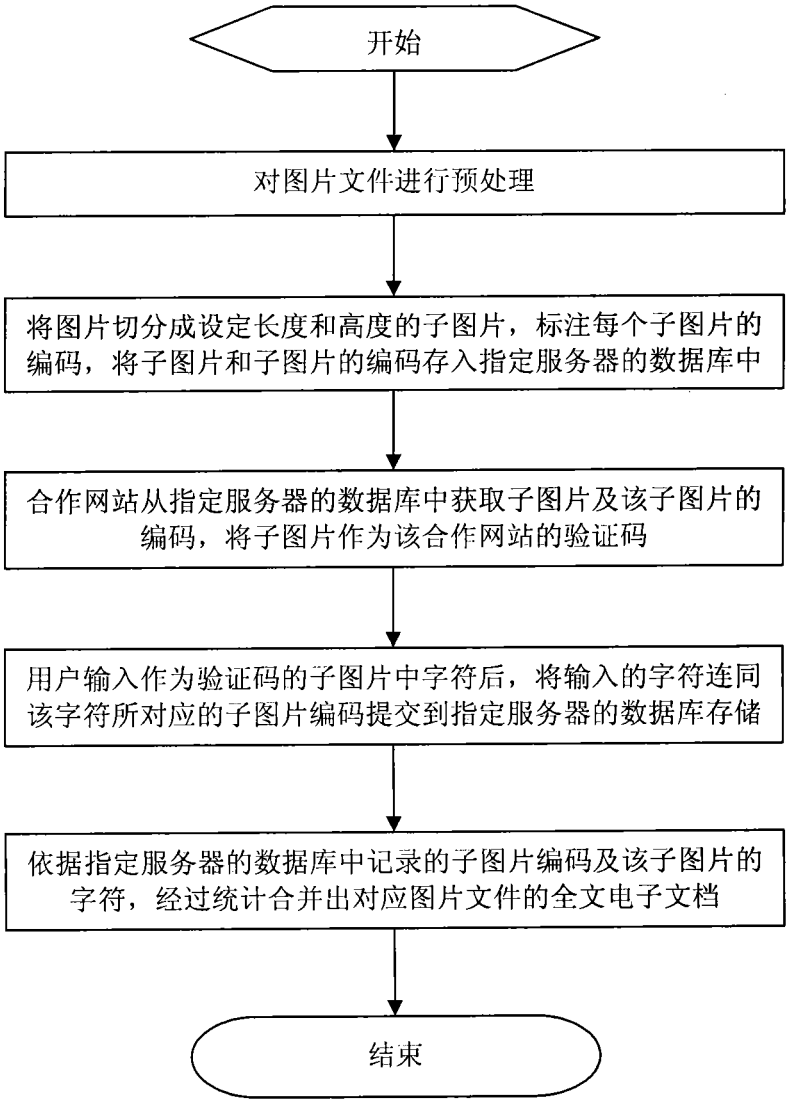

开始
对图片文件进行预处理
将图片切分成设定长度和高度的子图片，标注每个子图片的编码，将子图片和子图片的编码存入指定服务器的数据库中
合作网站从指定服务器的数据库中获取子图片及该子图片的编码，将子图片作为该合作网站的验证码
用户输入作为验证码的子图片中字符后，将输入的字符连同该字符所对应的子图片编码提交到指定服务器的数据库存储
依据指定服务器的数据库中记录的子图片编码及该子图片的字符，经过统计合并出对应图片文件的全文电子文档
结束

图 2