[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06K 9/03
G06K 9/62

[12] 发明专利申请公开说明书

[21] 申请号 01144254.9

[43] 公开日 2003 年 6 月 25 日

[11] 公开号 CN 1426017A

[22] 申请日 2001.12.14 [21] 申请号 01144254.9
[71] 申请人 全景软体股份有限公司
地址 台湾省新竹市
[72] 发明人 赵善隆 郑绍余 杨靖宇

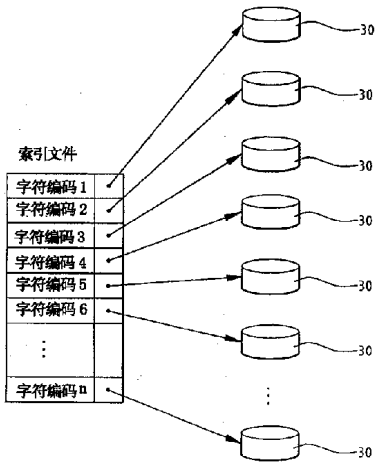
[74] 专利代理机构 隆天国际专利商标代理有限公司
代理人 陈 红 潘培坤

权利要求书 2 页 说明书 5 页 附图 6 页

[54] 发明名称 一种校对多个电子文件的方法及其系统

[57] 摘要

本发明提供一种校对 OCR 系统产生的多个电子文件的方法，该 OCR 系统将实际文件扫描成文件图像，再将其中的每个字符分割成一字符图像，并对该字符图像进行识别而得到一字符编码，最后将对应于该字符编码的字符输出至一文本文件，该方法步骤为：为每个电子文件编一文件编号；为其中每个字符编一位置编号；建立多个临时文件，将同一字符编码对应的字符收在一个临时文件中，每个临时文件含多个字符数据，包括和该字符关联的文件编号、位置编号及字符图像；在校对界面中显示临时文件内所有字符图像，用户可在编辑区内输入修正字符；按每个字符数据关联的文件编号、位置编号，以用户修正的字符取代关联的电子文件内该位置编号的字符。



ISSN1008-4274

1、一种校对多个电子文件的方法，其中该每一个电子文件是通过一光学字符识别系统（OCR）对一实际的文件进行识别所产生的一文本文件，在该识别过程中，该光学字符识别系统将该实际的文件扫描成一文件图像，且分别对该文件图像内的每一个字符进行分割成一字符图像，且分别对该每一个字符图像进行识别而得到一字符编码，最后该光学字符识别系统依据该字符编码，将对应于该字符编码的字符输出至该文本文件，其特征在于，所述的方法包含下列步骤：

（A）对该多个电子文件进行编号，使得该每一个电子文件具有一文件编号；

（B）对该经过步骤（A）编号的每一个电子文件内的每一个字符进行编号，使得该每一个字符具有一位置编号；

（C）依据经识别后而得到的该每一个字符编码以建立多个临时文件，以将对应于同一字符编码的所有字符收集在一临时文件中，其中该每一个临时文件包含多个字符数据，该每一个字符数据包含该每一个字符所关联的文件编号、位置编号、及字符图像；

（D）依据该每一个临时文件以产生一集字校对界面，其中该集字校对界面显示一临时文件内所有字符数据的字符图像，且分别提供一编辑区给该每一个字符图像以供使用者输入一修正字符；

（E）依据该每一个字符数据所关联的文件编号、位置编号，将该每一个字符图像所属的编辑区内的该修正字符分别储存至所关联的电子文件，以取代位于该关联的电子文件内的该位置编号的字符。

2、权利要求1所述的方法，其特征在于所述的步骤（C）进一步包含：建立一索引文件，该索引文件由此该字符编码为搜寻键来连结该多个临时文件，以提供使用者通过该索引文件来搜寻该多个临时文件。

3、权利要求 1 所述的方法，其特征在于所述的字符编码是一中文码。

4、一种校对多个电子文件的系统，其特征在于所述的每一个电子文件是通过一光学字符识别系统（OCR）对一实际的文件进行识别所产生的一文本文件，该系统包括：

• 临时文件模块，所述的临时文件模块由多个临时文件组成，其中该每一个临时文件包含具同一字符编码的多个字符数据，该每一个字符数据包含该每一个字符所关联的文件编号、位置编号、及字符图像；

• 索引文件模块，所述的索引文件模块由一索引文件组成，其中该索引文件由此一字符编码为搜寻键来连结该多个临时文件，以提供使用者通过该索引文件来搜寻该多个临时文件；

• 一集字校对界面，其中该集字校对界面显示一临时文件内所有字符数据的字符图像，且分别提供一编辑区给该每一个字符图像以供使用者输入一修正字符；

其中该文本文件用于依据该每一个字符数据所关联的文件编号、位置编号，将该每一个字符图像所属的编辑区内的该修正字符分别储存至所关联的电子文件，以取代位于该关联的电子文件内的该位置编号的字符。

5、如权利要求 4 所述的系统，其特征在于所述的字符编码是一中文编码。

一种校对多个电子文件的方法及其系统

技术领域

本发明是关于由光学字符识别系统（Optical Character Recognition，OCR）针对既有已存的实际的文件作字符识别，并将其转换成一电子文件，该电子文件存有若干识别错误的字符，而必须对其进行校对更正的动作，特别是有关于对大量电子文件进行校正的方法。

背景技术

现有OCR会因某些因素而经常有识别错误的情形发生，例如：输入文件的表面不干净，或由于扫描仪的解析度太低或本身的失真所造成扫描后的字体存在污点或不完整的现象；另外，对于大量中字符的数据输入，现有OCR系统的识别错误率更将随字数增加而提高。

对于现有OCR系统的识别错误率高的缺点，目前的解决方法是由人工逐字校对的方式来对识别错误的字符进行修改，这种校对方式无疑造成校对人员的负担，而且费时费力，尤其在大量数据的电子化的过程中，更显得没有效率。

图1A是传统的针对大量通过光学字符识别系统识别后的电子文件进行错误校正的系统结构示意图。实际的文件10通过扫描仪12扫描成多个文件图像14，且每一文件图像14是一图像文件，随后文件图像14传送至光学识别服务器16进行字符识别作业，光学识别服务器16将输出多个电子文件至多个客户端电脑18，以使校对人员进行校正作业，其中该每一电子文件是一文本文件。

图1B是图1A的客户端电脑所显示的校正画面。文件图像14通过文件

图像区11来显示，该输出的电子文件则通过字符编辑区13来显示，图1B表示传统的校正方法是由校对人员参照文件图像区11，以人工逐字校对的方式来对字符编辑区13内的字符进行修改，最后再重新储存该文本文件，来实现校正的目的。然而，这种校对方式无疑造成校对人员的负担，而且费时费力，尤其在大量数据的电子化的过程中，更显得没有效率。

本发明目的，是解决传统以人工逐字校对方式来针对由OCR系统识别后的电子文件进行修正，所造成的缺点，如：费时费力、效率不高，以及无法应付大量文件的校对作业。

为实现上述目的，本发明提供一种校对多个电子文件的方法，其中该每一个电子文件是通过一光学字符识别系统对一实际的文件进行识别所产生的一文本文件，在该识别过程中，该光学字符识别系统将该实际的文件扫描成一文件图像，且分别对该文件图像内的每一个字符进行分割成一字符图像，且分别对该每一个字符图像进行识别而得到一字符编码，最后该光学字符识别系统依据该字符编码，将对应于该字符编码的字符输出至该文本文件，该方法包含下列步骤：（A）对该多个电子文件进行编号，使得该每一个电子文件具有一文件编号；（B）对该经步骤（A） 编号的每一个电子文件内的每一个字符进行编号，使得该每一个字符具有一位置编号；（C）依据经识别后而得到的该每一个字符编码以建立多个临时文件，以将对应于同一字符编码的所有字符收集在一临时文件中，其中该每一个临时文件包含多个字符数据，该每一个字符数据包含该每一个字符所关联的文件编号、位置编号、及字符图像；（D）依据该每一个临时文件以产生一集字校对界面，其中该集字校对界面显示一临时文件内所有字符数据的字符图像，且分别提供一编辑区给该每一个字符图像以供使用者输入一修正字符；（E）依据该每一个字符数据所关联的文件编号、位置编号，将该每一个字符图像所属的编辑区内的该修正字符分别储存至所关联的电子文件，以取代位于该关联的电子文件内的该位置编号的字符。

为使熟悉该项技术的人士了解本发明的目的、特征及功效，兹通过下述具体实施例，并配合附图，对本发明详加说明如后。

附图说明

图1A是传统的针对大量通过光学字符识别系统识别后的电子文件进行错误校正的系统结构示意图。

图1B是图1A的客户端电脑所显示的校正画面；

图2是本发明方法的流程图；

图3显示依据本发明方法所使用的索引文件；

图4A显示图3的临时文件的内部示意图；

图4B显示图4A的字符数据的数据结构；

图4C是图4B的一具体实施例；

图5显示依据本发明的方法，一客户端电脑执行字符校正的显示器画面。

具体实施方式

依据本发明，由此撰写软件的手段利用光学识别服务器16在进行字符识别的过程中所产生的信息。在该识别过程中，光学识别服务器16分别对文件图像14内的每一个字符进行分割成一字符图像，且分别对该每一个字符图像进行识别而得到一字符编码，其中该字符编码是一般中文码，如：BIG-5、GB等等，最后该光学识别服务器16依据该字符编码，将对应于该字符编码的字符输出至所属的文本文件。图2即本发明方法的流程图，步骤20是对该多个电子文件进行编号，使得该每一个电子文件具有一文件编号；步骤22是对该经步骤20编号的每一个电子文件内的每一个字符进行编号，使得该每一个字符具有一位置编号；步骤24是依据经识别后而得到的该每一个字符编码以建立多个临时文件，以将对应于同一字符编码的所有

字符收集在一临时文件中，其中该每一个临时文件包含多个字符数据，该每一个字符数据包含该每一个字符所关联的文件编号、位置编号、及字符图像；步骤26是依据该每一个临时文件以产生一集字校对界面，其中该集字校对界面显示一临时文件内所有字符数据的字符图像，且分别提供一编辑区给该每一个字符图像以供使用者输入一修正字符；步骤28是依据该每一个字符数据所关联的文件编号、位置编号，将该每一个字符图像所屠的编辑区内的该修正字符分别储存至所关联的电子文件，以取代位于该关联的电子文件内的该位置编号的字符。其中步骤24进一步包含：建立一索引文件，该索引文件由此该字符编码为搜寻键（Search key）来连结该多个临时文件，以提供使用者通过该索引文件来搜寻该多个临时文件。

图3显示依据本发明方法所使用的索引文件。该索引文件由此字符编码为搜寻键，且每一字符编码连结一临时文件30，每一个临时文件30收集对应于同一字符编码的所有字符，其中该字符是来自于所有编号的电子文件内具有相同字符编码的字符，使用者可以通过该索引文件来搜寻到该多个临时文件。

图4A显示图3的临时文件的内部示意图。临时文件30内储存有多个具同一字符编码的字符数据40，其中字符数据40的数据结构显示在图4B，字符数据40包含每一个字符所关联的文件编号、位置编号、及字符图像，其中该字符图像是一图像文件。图4C是图4B的一具体实施例，说明一中字符的字符图像图像文件为“躋”，亦即该中字符是“躋”，且该中字符是位于编号为5的电子文件内的第20个位置编号的地方。

图5显示依据本发明的方法，一客户端电脑执行字符校正的显示器画面。图5是依据中字符“擠”的临时文件所产生的集字校对界面，该集字校对界面显示该临时文件内所有字符数据的字符图像图像文件50，已知该临时文件内所有字符数据具有相同的字符编码，亦即中字符“擠”的BIG-5

码，但由集字校对界面可观察出若干；字符图像相异于大部分的字符图像，此原因是光学识别服务器16识别错误所致，如图中的中字符“躋”被识别成中字符“擠”，而被收集到中字符“擠”的临时文件，为修正这种错误，本发明的集字校对界面提供一编辑区51给每一个字符图像以供使用者输入一修正字符，于是在中字符“躋”的字符图像50的下方编辑区51内，输入正确字符（即“躋”），最后依据中字符“躋”所在的文件编号及位置编号，将该编辑区51内的字符储存，至中字符“躋”所关联的电子文件上，如此即可将识别错误的字符“擠”修正为正确的字符“躋”。

本发明的特点是将识别错误的字符突出显示拥有众多相同字形的显示器画面上，使得校对人员一眼即可看出该识别错误的字符，此是使用视觉落差的效果以快速发现识别错误的字符，依据本发明的方法所带来的好处有三：其一，校对人员可快速找到识别错误的字符并修正之；其二，校对时间不会随着欲校对的数据量大幅增加而加倍；其三，校对人员不须经过特别训练而能轻松操作。

虽然本发明以一较佳实施例揭露如上，然而并非用以限定本发明，任何熟悉此技术者，在不脱离本发明的精神和范围内，当可作各种的更动与润饰，因此本发明的保护范围当视后附的权利要求范围所界定者为准。

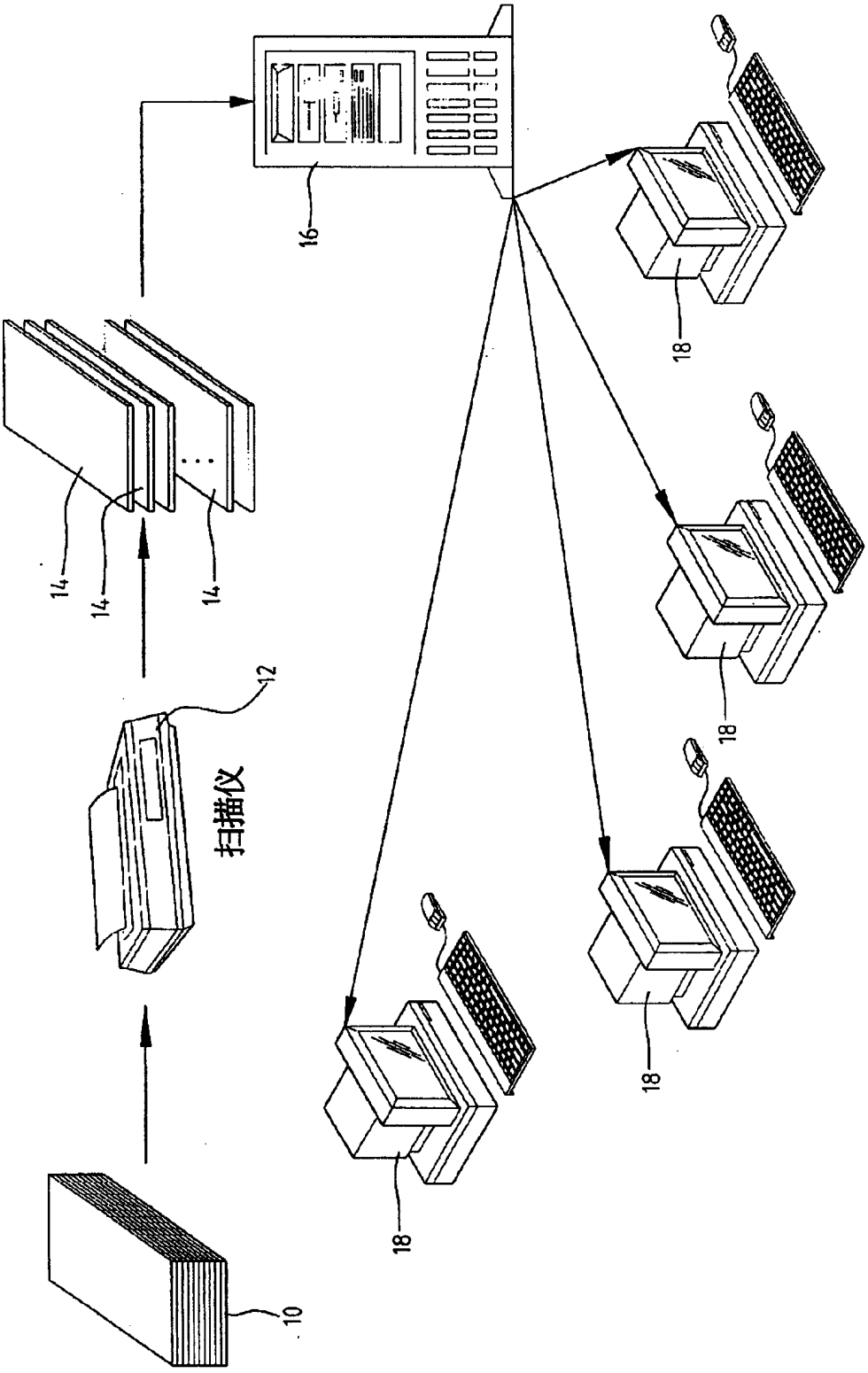


图1A

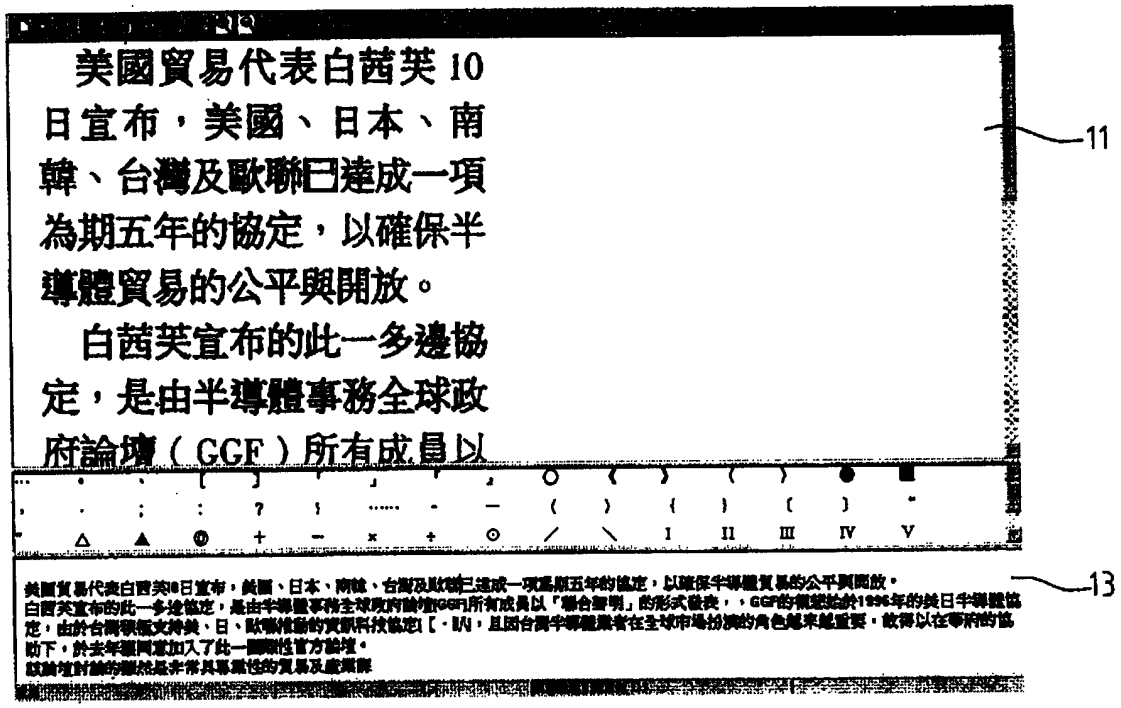

美國貿易代表白茜芙10日宣布，美國、日本、南韓、台灣及歐聯已達成一項為期五年的協定，以確保半導體貿易的公平與開放。
白茜芙宣布的此一多邊協定，是由半導體事務全球政府論壇（GGF）所有成員以
11
13

图1B

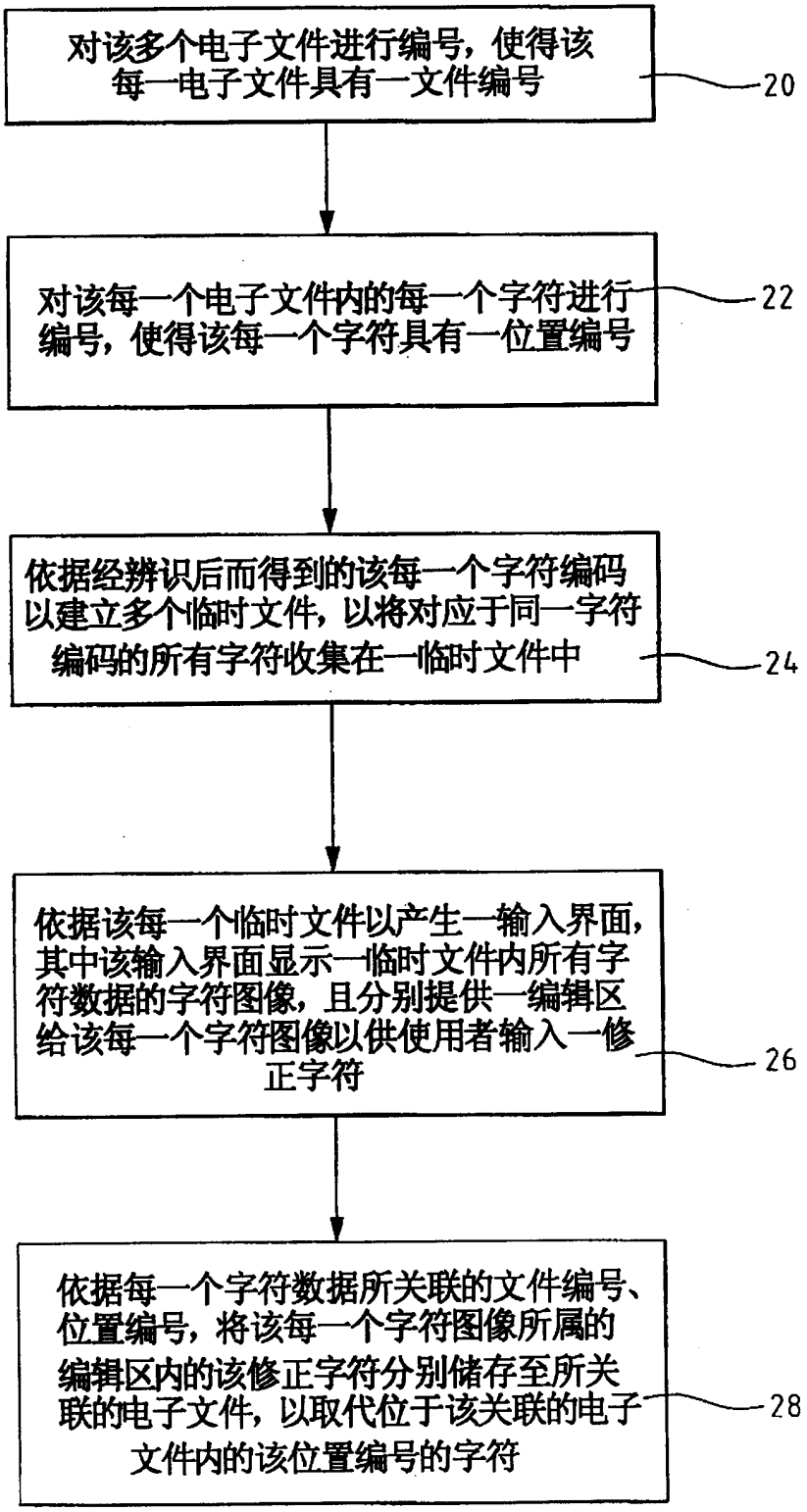

对该多个电子文件进行编号，使得该每一电子文件具有一文件编号
20
对该每一个电子文件内的每一个字符进行编号，使得该每一个字符具有一位置编号
22
依据经辨识后而得到的该每一个字符编码以建立多个临时文件，以将对应于同一字符编码的所有字符收集在一临时文件中
24
依据该每一个临时文件以产生一输入界面，其中该输入界面显示一临时文件内所有字符数据的字符图像，且分别提供一编辑区给该每一个字符图像以供使用者输入一修正字符
26
依据每一个字符数据所关联的文件编号、位置编号，将该每一个字符图像所属的编辑区内的该修正字符分别储存至所关联的电子文件，以取代位于该关联的电子文件内的该位置编号的字符
28

图 2

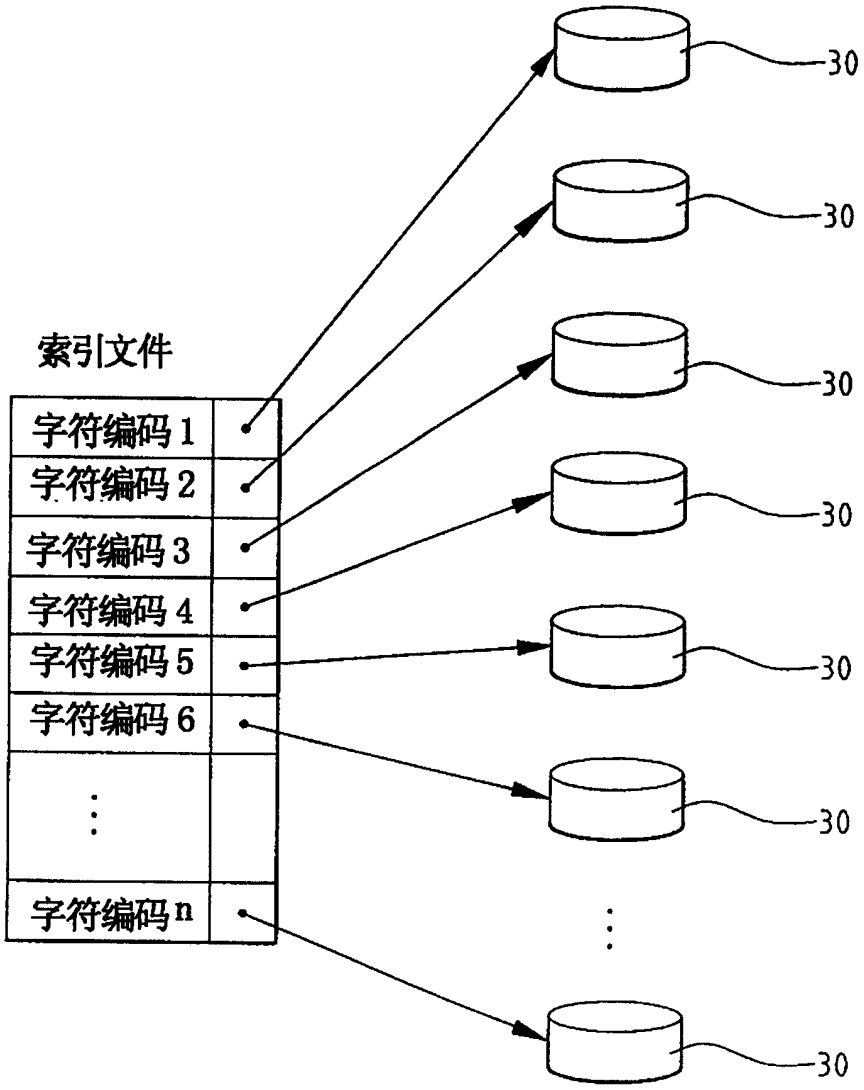

索引文件
字符编码 1
字符编码 2
字符编码 3
字符编码 4
字符编码 5
字符编码 6
字符编码 n
30
30
30
30
30
30
30

图 3

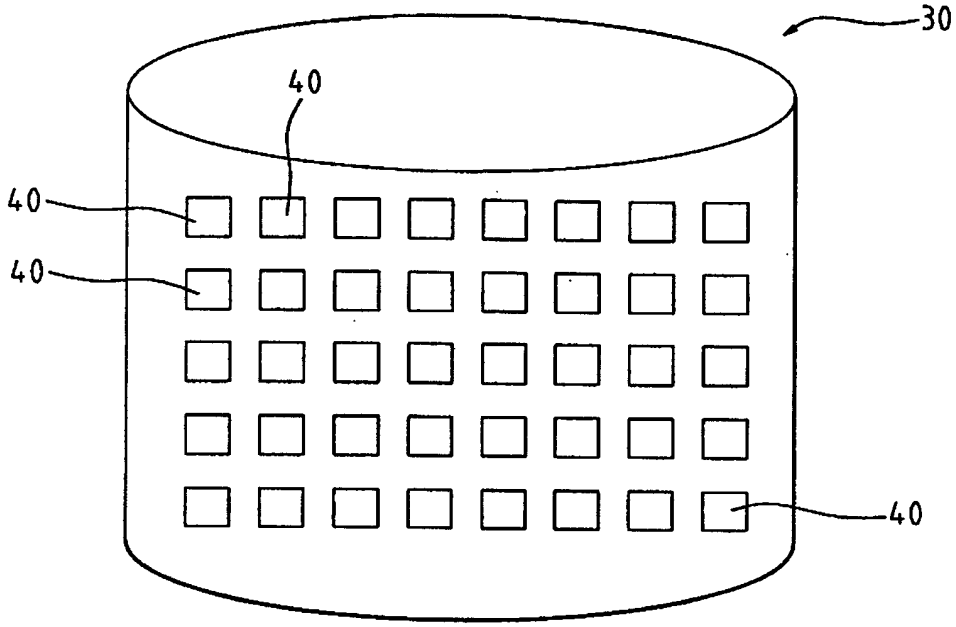


图 4A

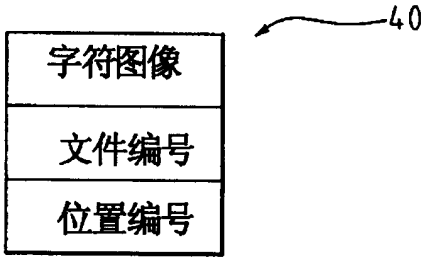


图 4B

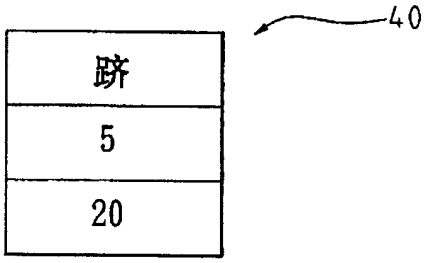


图 4C

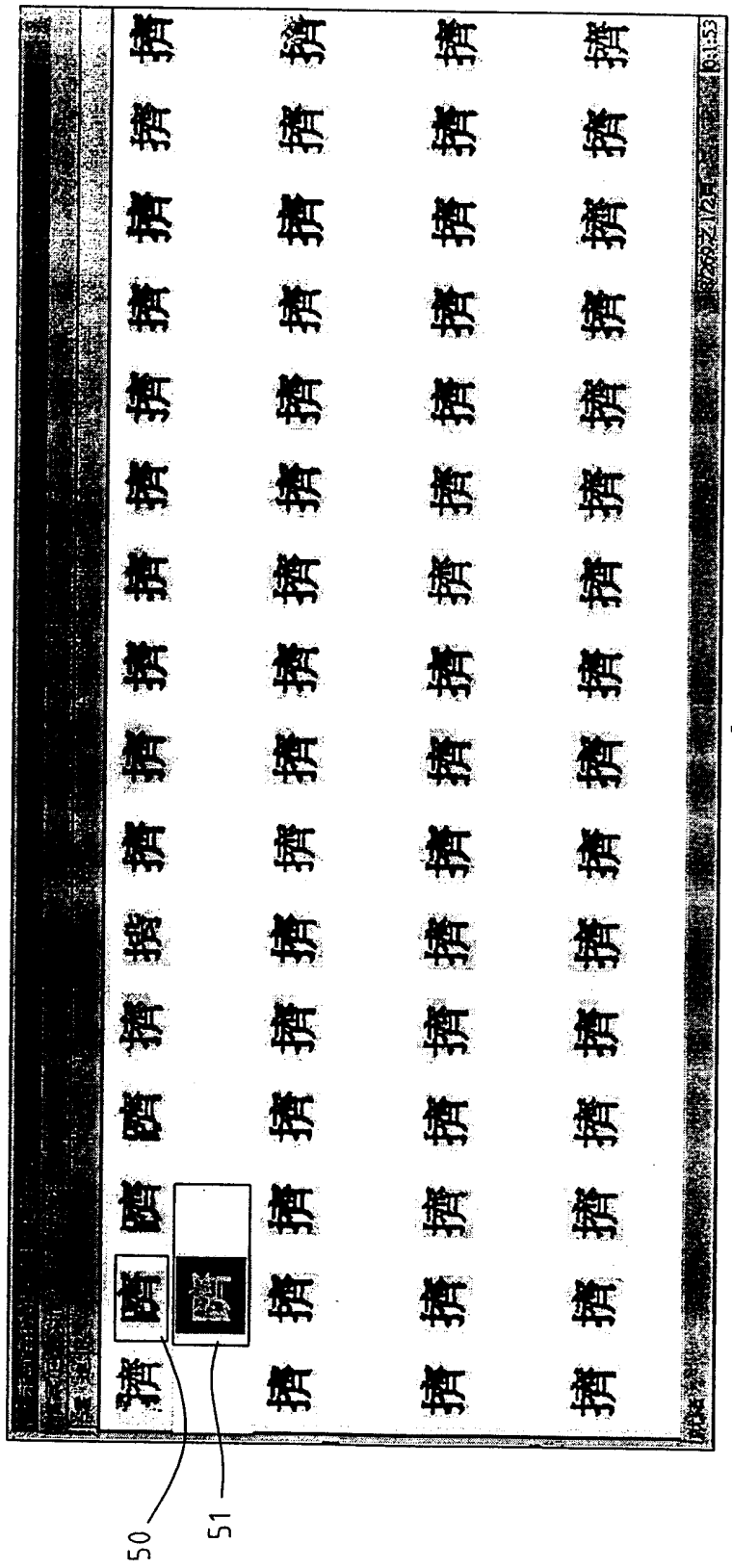


图5