



(12) 发明专利

(10) 授权公告号 CN 101458699 B

(45) 授权公告日 2015. 11. 25

(21) 申请号 200810183281. 3

(22) 申请日 2008. 12. 12

(30) 优先权数据

2007-321283 2007. 12. 12 JP

(73) 专利权人 佳能株式会社

地址 日本东京都大田区下丸子 3 丁目 30 番
2 号

(72) 发明人 榎本诚 金津知俊 中西惠子

(74) 专利代理机构 北京魏启学律师事务所
11398

代理人 魏启学

(51) Int. Cl.

G06F 17/30(2006. 01)

G06K 9/20(2006. 01)

(56) 对比文件

CN 1752992 A, 2006. 03. 29, 全文.

JP 特开 2005-275863 A, 2005. 10. 06, 说明书摘要, 说明书第【0013】段—第【0036】段, 说明书附图 1—4.

审查员 徐淑娴

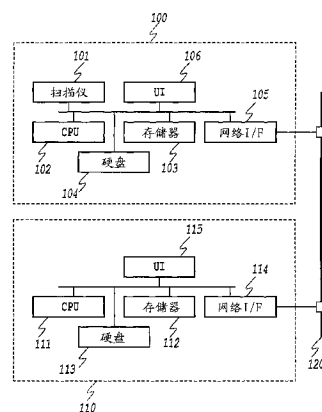
权利要求书2页 说明书17页 附图18页

(54) 发明名称

图像处理装置和图像处理方法

(57) 摘要

本发明提供一种图像处理装置和图像处理方法。在页面图像和字符的绘制描述的电子文档中, 希望尽管绘制字符所需的字体数据被保持在电子文档中, 但是可以将电子文档的大小最小化。另外, 希望在突出显示搜索时确保可见度。生成了一种存储了文档图像、多个字符代码以及多种字体数据的电子文档, 其中, 多个字符代码是对文档图像执行字符识别处理而获得的。在绘制与多个所述字符代码相对应的字符时, 选择所述多种字体数据中的一种字体数据, 并且对于多个所述字符代码共同使用所选择的字体数据。字体数据是具有简单形状的字体数据, 则是更令人满意的。



1. 一种图像处理装置,包括:

字符识别单元,用于通过对文档图像中的多个字符图像执行字符识别处理,来获得代表多个不同的字符的多个不同的字符代码;以及

生成单元,用于生成存储了所述文档图像、代表多个不同的字符的所述多个不同的字符代码、以及多种字符形状数据的电子文档,其中,所述多个不同的字符代码是由所述字符识别单元获得的,

其中,在通过关键字搜索而搜索到多个不同的字符代码时,从存储在所述电子文档中的所述多种字符形状数据中仅选择一种字符形状数据,并且对于存储在所述电子文档中的所述多个不同的字符代码全部共同使用所述一种字符形状数据,以突出显示由搜索到的多个不同的字符代码代表的多个字符。

2. 根据权利要求1所述的图像处理装置,其特征在于,在通过关键字搜索而搜索到多个不同的字符代码时,由用户的指示或用于显示的应用程序从所述多种字符形状数据仅选择所述一种字符形状数据。

3. 根据权利要求1所述的图像处理装置,其特征在于,所述多种字符形状数据中的任一种字符形状数据是具有简单形状的字符形状数据。

4. 根据权利要求1所述的图像处理装置,其特征在于,存储在所述电子文档中的所述多种字符形状数据包括具有矩形形状的字符形状数据。

5. 根据权利要求1所述的图像处理装置,其特征在于,存储在所述电子文档中的所述多种字符形状数据包括如下的字符形状数据:该字符形状数据被绘制在相应的字符图像的下端部处,使得在搜索时,如对所述相应的字符图像添加下划线那样来指定该相应的字符图像。

6. 根据权利要求1所述的图像处理装置,其特征在于,存储在所述电子文档中的所述多种字符形状数据包括至少具有波浪线、虚线、三角形、圆形和四边形之一的字符形状数据。

7. 根据权利要求1所述的图像处理装置,其特征在于,所述电子文档是以XML格式或XPS格式描述的电子文档。

8. 根据权利要求1所述的图像处理装置,其特征在于,还包括:

压缩单元,用于压缩所述文档图像,

其中,要存储在所述电子文档中的所述文档图像是由所述压缩单元进行了压缩处理的文档图像。

9. 根据权利要求8所述的图像处理装置,其特征在于,所述压缩单元对包括在所述文档图像中的区域进行分析,并且适应性地进行压缩。

10. 根据权利要求1所述的图像处理装置,其特征在于,还包括:

搜索单元,用于对所生成的所述电子文档搜索所输入的关键字,并且使用所述多种字符形状数据中的任一种字符形状数据来突出显示与所述关键字相一致的部分。

11. 根据权利要求2所述的图像处理装置,其特征在于,还包括:

搜索单元,用于对所生成的所述电子文档搜索所输入的关键字,并且使用由用户的指示或所述应用程序从所述多种字符形状数据之中选择出来的字符形状数据来突出显示与所述关键字相一致的部分。

12. 一种图像处理装置,包括:

搜索单元,用于利用输入的关键字来对存储了文档图像、代表多个不同的字符的多个不同的字符代码、以及多种字符形状数据的电子文档进行搜索,其中,所述多个不同的字符代码是通过对所述文档图像执行字符识别处理而获得的,以及在通过关键字搜索而搜索到多个不同的字符代码时由存储在所述电子文档中的所述多个不同的字符代码全部共同使用所述多种字符形状数据中的一种字符形状数据,以突出显示由搜索到的多个不同的字符代码代表的多个字符;

切换单元,用于接收来自用户的、关于所述多种字符形状数据的字符形状数据切换指示;以及

电子文档显示单元,用于使用所述切换单元切换的一种字符形状数据,来突出显示作为所述搜索单元的搜索结果的、与所述关键字相一致的部分。

13. 一种图像处理装置,包括:

搜索单元,用于利用输入的关键字来对存储了文档图像、代表多个不同的字符的多个不同的字符代码、多种字符形状数据、以及属性数据的电子文档进行搜索,其中,所述多个不同的字符代码是通过对所述文档图像执行字符识别处理而获得的,所述属性数据在选择用于绘制的字符形状数据时用作判断标准,以及在通过关键字搜索而搜索到多个不同的字符代码时由存储在所述电子文档中的所述多个不同的字符代码全部共同使用所述多种字符形状数据中的一种字符形状数据,以突出显示由搜索到的多个不同的字符代码代表的多个字符;以及

电子文档显示单元,用于使用基于所述属性数据从所述多种字符形状数据之中选择出来的一种字符形状数据,来突出显示作为所述搜索单元的搜索结果的、与所述关键字相一致的部分。

14. 一种图像处理方法,包括:

通过对文档图像中的多个字符图像执行字符识别处理,来获得代表多个不同的字符的多个不同的字符代码;以及

生成存储了文档图像、代表多个不同的字符的所述多个不同的字符代码、以及多种字符形状数据的电子文档,其中,所述多个不同的字符代码是通过所述字符识别处理而获得的,

其中,在通过关键字搜索而搜索到多个不同的字符代码时,从存储在所述电子文档中的所述多种字符形状数据中仅选择一种字符形状数据,并且对于存储在所述电子文档中的所述多个不同的字符代码全部共同使用所述一种字符形状数据,以突出显示由搜索到的多个不同的字符代码代表的多个字符。

图像处理装置和图像处理方法

技术领域

[0001] 本发明涉及一种将纸文档转换成可以进行电子搜索的数据的技术。

背景技术

[0002] 近年来,扫描仪或硬盘等大容量存储装置允许用计算机对一直以来存储为纸文档的文档进行处理的简单操作,以及将文档存储为电子文档的简单操作。尤其是,不仅扫描纸文档以将其转换成图像数据,一般还执行了通过字符识别技术来读取写在纸文档上的字符信息以将其存储为图像的附加信息。对于已经以这种方式存储的电子文档,用户通过将包括在输入的原始文档中的字符串作为搜索关键字,可以高速地从大量存储的文档组中检索出所期望的文档。

[0003] 另外,提出了如下的技术,在用户使用搜索关键字对涉及了这样的字符信息的电子文档进行搜索时,突出显示 (highlight) 在文档图像上描述了该搜索关键字的部分,从而使用户能够识别出该部分 (例如,日本特开 2000-322417)。在这种方式下,由于是在突出显示的状态来显示与该搜索关键字相对应的字符部分,所以即使在文档中出现相同关键字的多个描述点的情况下,通过切换页面图像,用户可以高效率地识别出关键字的描述点。

[0004] 然而,还有一种将字符识别处理的结果作为透明文本 (transparent text) (被指定为绘制颜色的透明色的字符代码) 嵌入到图像文件中,并以 PDF (便携文档格式) 格式来存储的技术。当显示已经以这种方法创建的 PDF 文件时,将透明文本绘制在文档图像中的字符图像上。因此,当进行关键字搜索时,搜索透明文本。然而,用户不能看到透明文本自身,从而看上去好像搜索图像。在这种方式下,基于利用能够绘制图像和字符的页面描述语言来描述的格式的文件,可以绘制能够利用搜索关键字搜索的图像。

[0005] 为了使用 PDF 或 SVG 等页面描述语言来绘制在电子文档中字符,需要每个字符的字符形状信息,即字体数据。然而,由于一般字体数据的大小很大,出于减小电子文档的大小的目的,一般不将字体数据存储存在电子文档中,并且在电子文档中做出对字体种类的指定。这样,当利用应用程序 (application) 来绘制时,可以使用安装在个人计算机中的字体来完成绘制。

[0006] 另一方面,存在一些期望将字体数据存储存在电子文档中的情况。例如,在使用其它个人计算机打开了已经通过使用文档创建应用程序而创建的电子文档的情况下,当在该电子文档中使用的字体数据没有被安装在该个人计算机时,不能准确地打开该电子文档。换句话说,如果字体数据自身存储在电子文档中,即使在使用未安装有指定的字体数据的个人计算机或应用程序来再生电子文档的情况下,也可准确地再生该电子文档。

[0007] 另外,依赖于应用程序,在一些情况下,将用于绘制字符的字体数据存储存在电子文档中是优选的必要条件。例如,对于意图长期存储的文件,在经过长的时间段后,由于 OS 的改变,作为默认安装的字体可能改变。因此,预计需要存储字体数据存储的形式。

[0008] 此外,依赖于格式的形式,还存在将字体数据存储存在电子文档中是必要条件的格式。例如,在 XPS (XML Paper Specification, XML 文件规格) 格式中,在存储了文本数据的

情况下,还需将字体数据一同存储。

[0009] 然而,当将字体数据存储于电子文档中时,电子文档自身的大小增大。在文件大小增大的情况下,存在当通过网络发送电子文档时需要长的时间,或者存储时需要大的存储容量的问题。

[0010] 以这种方法,在通过使用存储于电子文档中的字体数据来绘制文件格式的电子文档中,期望防止文件大小的增大。尤其是,在将扫描图像、字符识别处理的结果的文本数据以及用于文本绘制的字体数据一同存储于电子文档中的情况下,期望防止文件大小的增大。当由于格式的限制或系统的限制而必须将字体数据存储于电子文档中时,文件大小的增加可能是难以解决的。

[0011] 另外,突出显示搜索结果时,依赖于显示文档的查看器(viewer)的特性,存在对搜索结果的不同的突出显示方法。即,依赖于对搜索结果突出显示的性能,有可能难以看到图像上的字符图像。

[0012] 在这样的情况下,在将纸文档转换为可以进行电子搜索的电子文档的处理中,需要如下的功能。即,期望在将电子文档的大小最小化的同时确保搜索的突出显示时的可见度,即使将要使用的字体数据保持在电子文档中。

发明内容

[0013] 本发明的一个目的是在能将电子文档的大小最小化的同时确保搜索的突出显示时的可见度,即使将要使用的字体数据保持在电子文档中。

[0014] 根据本发明的第一个方面,提供一种图像处理装置,包括:字符识别单元,用于通过对文档图像中的多个字符图像执行字符识别处理,来获得与每个字符图像相对应的字符代码;以及生成单元,用于生成存储了所述文档图像、多个所述字符代码以及多种字体数据的电子文档,其中多个所述字符代码是由所述字符识别单元获得的,其中,在绘制与多个所述字符代码相对应的字符时,选择所述多种字体数据中的一种字体数据,并且对于多个所述字符代码共同使用所选择的字体数据。

[0015] 根据本发明的第二个方面,提供一种图像处理装置,包括:搜索单元,用于利用输入的关键字来对存储了文档图像、多个字符代码以及多种字体数据的电子文档进行搜索,其中,所述多个字符代码是通过对所述文档图像执行字符识别处理而获得的,以及所述多种字体数据在绘制与所述多个字符代码相对应的字符时由所述多个字符代码共同使用;切换单元,用于接收来自用户的、关于所述多种字体数据的字体数据切换指示;以及电子文档显示单元,用于使用所述切换单元切换的字体数据,来突出显示作为所述搜索单元的搜索结果、与所述关键字相一致的部分。

[0016] 根据本发明的第三个方面,提供一种图像处理装置,包括:搜索单元,用于利用输入的关键字来对存储了文档图像、多个字符代码、多种字体数据以及属性数据的电子文档进行搜索,其中,所述多个字符代码是通过对所述文档图像执行字符识别处理而获得的,所述多种字体数据在绘制与所述多个字符代码相对应的字符时由所述多个字符代码共同使用,以及所述属性数据在选择用于绘制的字体数据时用作判断标准;以及电子文档显示单元,用于使用基于用作判断标准的所述属性数据从所述多种字体数据之中选择出来的字体数据,来突出显示作为所述搜索单元的搜索结果、与所述关键字相一致的部分。

[0017] 根据本发明的第四个方面,提供一种图像处理方法,包括:通过对文档图像中的多个字符图像执行字符识别处理来获得与每个字符图像相对应的字符代码;以及生成存储了文档图像、多个所述字符代码以及多种字体数据的电子文档,其中所述多个所述字符代码是通过所述字符识别处理而获得的,其中,在绘制与多个所述字符代码相对应的字符时,选择所述多种字体数据中的一种字体数据,并且对于多个所述字符代码共同使用所选择的字体数据。

[0018] 根据本发明,纸文档被转换成包括被扫描的页面图像的绘制描述和(用透明色)绘制已经从该页面图像提取出的字符的描述的电子文档。在这种情况下,在该电子文档的内部部分,存储了由简单的字符形状形成的多种字体数据。另外,在每种字体数据中,使得由多个字符种类(多个不同的字符代码)共同使用一种字符形状。由此,尽管要使用的字体数据被保持在该电子文档中,也只需要少量的字体数据(字符形状数据),所以能够将该电子文档的文件大小(数据容量)最小化。此外,存储了一些字体数据,并且可以在多个字符形状之中进行切换并进行绘制。从而,即使在使用不同的应用程序的突出显示搜索的情况下,也可以实现高的可见度或操作性的显示

[0019] 根据以下(参照附图)对典型实施例的说明,本发明的其它特征将变得显而易见。

附图说明

- [0020] 图 1 是示出根据本发明的实施例 1 和 2 的图像处理装置的配置例子的框图;
- [0021] 图 2 是示出根据本发明的实施例 1 和 2 的电子文档生成处理的例子的流程图;
- [0022] 图 3 是示出根据本发明的实施例 1 和 2 的电子文档搜索和浏览处理的例子的流程图;
- [0023] 图 4 是示出在图 2 中的步骤 S208 中执行的电子文档数据生成处理的例子的流程图;
- [0024] 图 5 是示出在图 3 中的步骤 S307 中执行的页面的绘制处理的例子的流程图;
- [0025] 图 6 是根据本发明的实施例 1 生成的电子文档的一个例子;
- [0026] 图 7 是在本发明的实施例 1 和 2 中处理的页面图像的一个例子;
- [0027] 图 8 是示出在本发明的实施例 1 和 2 中执行的区域分析处理的例子的图;
- [0028] 图 9 是在本发明的实施例 1 和 2 中生成的区域数据的一个例子;
- [0029] 图 10 是示出在本发明的实施例 1 和 2 中执行的字符识别处理的例子的图;
- [0030] 图 11A 和图 11B 是在本发明的实施例 1 和 2 中生成的字符代码串数据的例子;
- [0031] 图 12 是在本发明的实施例 1 和 2 中使用的字符代码串表的一个例子;
- [0032] 图 13 是在本发明的实施例 1 和 2 中执行的、将搜索点突出显示的页面显示的例子;
- [0033] 图 14 是在本发明的实施例 1 中执行的、将搜索点突出显示的页面显示的例子;
- [0034] 图 15 是根据本发明的实施例 2 生成的电子文档的一个例子;
- [0035] 图 16 是在本发明的实施例 1 中显示的 UI 的例子;
- [0036] 图 17 是当突出显示了搜索部分时,字符图像变为不可见的页面显示的例子;
- [0037] 图 18 是要处理的页面图像的例子;
- [0038] 图 19 是将搜索结果突出显示的页面显示的例子;以及

[0039] 图 20 是将搜索结果突出显示的页面显示的例子。

具体实施方式

[0040] 下面将参照附图说明根据本发明的优选实施例。然而,应理解,这些实施例中说明的组件是出于说明性的目的,本发明的范围不限于这些组件。

[0041] 实施例 1

[0042] 下面将参照附图说明根据本发明的典型实施例。

[0043] 图 1 是示出可以实现本发明的图像处理装置的配置的框图的一个例子。

[0044] 附图标记 100 表示由如下各装置构成的、将纸文档转换成电子文档的图像处理装置。

[0045] 附图标记 101 表示将已经被读取的纸文档的版面 (pagespace) 信息转换成图像的数据的扫描仪。附图标记 102 表示执行例如分析图像数据并将该图像数据转换成可搜索的电子文档的程序的 CPU。存储器 103 和硬盘 104 是用于存储利用上述程序操作进行的向电子文档转换的结果或进程 (progress) 中的数据的数据的存储装置。

[0046] 附图标记 105 表示向装置外部输出已经根据上述程序生成的数据的网络 I/F。附图标记 106 表示用于接收来自用户的指示的接口,该接口由输入键或触摸板等输入装置以及液晶等显示装置构成。

[0047] 附图标记 110 表示搜索并浏览已经在图像处理装置 100 处创建的电子文档的图像处理装置,该图像处理装置由如下装置构成。

[0048] CPU 111 执行解释电子文档的数据并创建用于搜索或浏览的显示数据的程序,或者执行控制搜索操作的程序。存储器 112 和硬盘 113 是用于存储电子文档数据或者存储根据上述程序的操作而创建的显示数据或进程中的数据的数据的存储装置。附图标记 114 表示用于将已经在装置外部创建了的电子文档传送到装置中的网络 I/F。附图标记 115 表示用于接收来自用户的指示的接口,该接口由输入键或触摸板等输入装置以及液晶等显示装置构成。

[0049] 附图标记 120 表示在图像处理装置 100 和图像处理装置 110 之间提供电子的连接的网络。

[0050] 现在,使用图 2 和 3 的流程图说明根据实施例 1 的处理的例子。

[0051] 图 2 是示出从已经通过例如在图像处理装置 100 处扫描纸文档而获得的图像数据来创建可搜索的电子文档,并且将该电子文档发送到图像处理装置 110 的处理的例子的流程图。

[0052] 首先,在步骤 S201 中,根据来自用户的指示操作,确定要创建的电子文档的发送目的地和发送方法。经由用户接口 106 执行来自用户的指示。另外,从电子邮件或使用 FTP 的文件传送等可选方式 (alternative) 之中选择发送方法。

[0053] 当用户设置纸文档并按下开始键时,在步骤 S202 中,扫描仪 101 通过使用已知的光电转换技术来扫描纸文档,以将纸文档转换成页面图像数据 (文档图像)。在手动或使用自动文档给送器输入多个文档页面的情况下,将已经被扫描的纸文档的每一个页面转换成一个页面图像数据,并按输入顺序将其存储在存储器 103 中。

[0054] 图 7 示出页面图像的例子。在图 7 的页面图像 701 中,出现了照片 704 以及由“A、

I、U(每个均为日文字符)”的字符串 702 和“KA、KI、KU(每个均为日文字符)”的字符串 703 组成的字符图像。顺便提及,出于说明的目的,尽管用黑色矩形来简单地示出照片 704,照片 704 实际上是自然图像。此外,在图 7 的例子中,尽管仅示出了字符串 702 和 703 以及照片 704 的例子,但是还可以是图形等其它区域。在图 7 中,示出了文档图像的例子,其中作为字符串 702、703 说明了平假名字符(日文字符),然而在本实施例中,字符不限于平假名字符,也可以使用字母和汉字字符(中文字符)等其它字符。图 18 是包括字母字符串 1802、1803 和照片 1804 的文档图像 1801 的例子。即,如果执行适合于每种语言的字符识别处理,则可以类似地对使用其它语言描述的文档执行将在后面说明的处理,而限于日文文档。

[0055] 至于页面图像数据的形式,例如,当纸文档为彩色时,是利用 RGB 的各 8 比特来表现其灰度的彩色图像;当纸文档为黑白时,是利用 8 比特来表现亮度的灰度级(gray scale) 图像或者利用 1 比特来表现黑白的二值图像。

[0056] 在步骤 S203 中,选择存储在存储器 103 中的未处理的页面图像数据作为处理对象图像。顺便提及,在有多个页面的图像的情况下,按输入顺序选择一个页面的图像作为处理对象。

[0057] 在步骤 S204 中,从已经被选择的图像中,对具有文本区域、图形区域、照片区域和表区域等不同特性的各个区域进行区域识别的区域分析处理,并且生成关于每个已经被识别出的区域的区域数据并将其存储在存储器 103 中。顺便提及,在该区域数据中,包括了:在关于相应区域的边界矩形(bounding rectangle)的左上位置的图像中像素的 x 和 y 方向上的坐标值 x 和 y;代表该边界矩形的大小(宽度和高度)的像素宽度和高度的数值 Width 和 Height;以及文本和照片等区域类型。

[0058] 在上述的区域分析处理中,采用了已知的技术(该技术还被称为区域识别处理、区域确定处理和区域提取处理)。例如,在采用了日本特开平 6-68301 所公开的技术的情况下,可以从二值文档图像数据中,提取沿横向或纵向排成一行的类似大小的黑像素块的范围作为文本区域。

[0059] 这里,在图 8 和 9 中,示出了对在图 7 中示出的页面图像 701 进行区域分析处理的例子。图 8 中的附图标记 801 表示被确定为文本区域的区域,并且附图标记 802 表示被确定为照片区域的区域。图 9 是通过该区域分析处理而获得的区域数据的例子。同样当对包括字母字符串的页面图像(例如,图 18 中的图像 1801) 执行区域分析处理时,将获得与图 7 中的区域分析结果类似的区域分析结果。

[0060] 接着,在步骤 S205 中,通过对已经被区域分析处理所识别的各文本区域中的字符图像执行字符识别处理,来获得各文本区域的字符代码串的数据并将其存储在存储器 103 中。这里,字符代码串的数据中,包括了:作为对区域中的每个字符图像的识别结果的字符代码信息;以及每个字符图像的边界矩形信息(边界矩形的左上处的坐标 x 和 y 及其宽度(Width) 和高度(Height) 信息)。

[0061] 这里,将简要说明上述字符识别处理的例子。顺便提及,字符图像的字符识别处理可以采用已有的技术。

[0062] 首先,在文档图像不是二值图像的情况下,对文本区域的内部部分执行二值化以获得在文本区域中的二值图像。对于该已经被二值化的各文本区域的内部部分,对每个垂直行或水平行中的黑像素的数量进行计数以创建直方图。基于长度和宽度的直方图,令周

期性的直方图的方向为行方向,并且令直方图的黑像素的数量不小于预定阈值的部分成为形成字符行的部分,从而获得行图像(line image)带(strip)。随后,对每个行图像在垂直于行方向的方向上获得直方图,并且基于直方图的结果切割出每个字符的图像。该已经被切割出的范围是一个字符的边界矩形信息。顺便提及,尽管这里使用对黑像素的数量进行计数的直方图来做出判断,还可以使用示出每一行中有无黑像素的投影来做出字符区域的判断。

[0063] 接着,从每个字符图像的边界矩形中的图像,检索例如边界成分以获得特征矢量,将该特征矢量与已经预先登记的字符识别字典中的特征矢量进行比较,以获得相似度。然后,将具有最高相似度的字符种类(字符的种类)的代码作为该矩形中的字符图像的字符代码。以这种方式,对于在文本区域中出现的所有字符的边界矩形,可以获得分配了字符代码的数据。然后,将已经从每个文本区域获得的字符代码集(set)作为字符代码串。

[0064] 另外,关于英文句子的字符区域,还判断在字符之间是否存在单词之间的空格。例如,做出关于单词之间的距离是否很长的判断,或者通过字符图像的字符识别结果的字符串和单词字典之间的匹配来做出关于是否为单词之间的间隙的判断,由此可以判断是否存在单词之间的空格。在判断为存在单词之间的空格的情况下,将该空格的字符代码插入到字符代码串中。

[0065] 顺便提及,上述的说明是一个例子,并且通过采用利用了其它已知技术的处理方法,可以获得字符代码串。

[0066] 在图 10 以及图 11A 和 11B 中,示出了对图 8 中示出的文本区域 801 进行字符识别处理的例子。

[0067] 从图 10 中的文本区域 1000,切割出第一字符行 1001 和 1002,并且从字符行 1001 的内部进一步切割出三个字符 1011、1012 和 1013。然后,识别出每个字符,并且作为结果,获得对应于每个字符的字符代码,来生成如图 11A 的表 1101 所示的字符代码串数据。以相同的方法,对已经从字符行 1002 的内部切割出的三个字符 1021、1022 和 1023 进行字符识别处理,由此生成图 11B 中的字符代码串数据 1102。图 10 以及图 11A 和 11B 示出对日文文本区域执行字符识别处理的情况,然而,本发明不限于日文。例如,还可以通过对包括其它语言(例如,英语字母)的文本区域类似地执行字符识别处理来获得字符代码。即,当英文字母等字符被识别时,替代日文字符代码将字母的字符代码输入到图 11A 和 11B 中的字符代码串数据中。

[0068] 随后,在步骤 S206 中,使将要处理的页面图像数据、区域数据和字符代码串数据相关联,并将它们存储在存储器 103 或硬盘 104 中。

[0069] 在步骤 S207 中,判断是否存在尚未处理的图像数据。当存在未处理的图像数据时,操作返回到步骤 S203,并进行下一页面图像数据的处理。当没有未处理的图像数据时,操作进入到步骤 S208。

[0070] 在步骤 S208 中,按页面顺序(处理的顺序)合成存储在存储器 103 或硬盘 104 中的所有页面的数据,以生成由多个页面形成的可搜索的电子文档。

[0071] 在上述步骤 S208 中生成的电子文档的数据是能够保持绘制信息和内容信息两者的数据,其中绘制信息用于在显示器等上电子地显示每个页面图像或使用打印机打印每个页面图像,以及内容信息用于使得能够利用搜索关键字进行搜索。作为满足这样的数据保

持条件的数据格式,存在 PDF 格式或 SVG 格式等多种已知的例子。在本实施例中,作为要在这时生成的电子文档的格式,假定要指定嵌入字体数据。顺便提及,作为具有嵌入字体数据的必要条件的格式形式,例如 XPS。尽管在下文中,在假定使用 XML 表示的页面描述格式的规格 (specification) 情况下做出了描述,但本发明不限于该格式。当然也可以使用现有的嵌入字符形状的 XPS 或 PDF/A 等电子文档格式。

[0072] 图 6 是在输入了由两页的页面图像形成的文档的情况下,基于在本发明的说明中所使用的页面描述格式的规格而生成的电子文档的页面描述的例子。顺便提及,尽管这里,如图 6 所示,页面描述格式的例子是在一个文件中做出了全部描述的一个例子,但是并不限于该例子。例如,可优选为这样的格式(例如 XPS):在该格式中字体数据的部分在另一文件中,从主体的文件引用字体数据文件,并且通过例如 ZIP 压缩使这些文件合并在一个电子文档。

[0073] 以下,将使用图 4 的流程图说明在步骤 S208 中执行的电子文档数据生成处理的例子。

[0074] 首先,在步骤 S401 中生成表示电子文档的开始描述。

[0075] 在根据本发明的说明的页面数据描述格式规格中,要素 (element) <Document> 代表电子文档的开始标识 (tag),并且夹在该 <Document> 和作为该文档的结束标识的 </Document> 之间的范围中的 XML 描述是与包括在该文档中的每个页面有关的描述数据。在图 6 的例子中,附图标记 601 表示电子文档的开始标识,并且附图标记 613 表示结束标识。

[0076] 在步骤 S402 中,从尚未描述的页面之中指定第一页面的数据,并使其成为要处理的对象。

[0077] 在步骤 S403 中,生成代表要处理的页面数据的开始的标识以对其进行描述。在本例子中,要素标识 <page> 描述页面数据的开始,并且夹在该 <page> 和作为该页面的结束标识的 </page> 之间的范围中的 XML 描述为该页面中的绘制数据和内容数据。另外,在 <page> 标识中,使用表示该页面的像素的宽度和高度的属性 Width 和 Height 以及表示分辨率的属性 Dpi 来描述页面的物理大小,并且还使用表示页面数量的属性 Number 来描述页面数量。

[0078] 在图 6 的描述例子中,在 <page> 要素的开始标识 602 中,描述了该页面的宽度 Width = “1680”、高度 Height = “2376”、分辨率 Dpi = “200” 和页面数量 Number = “1”。此外,从标识 603 到结束标识 606 (603 到 606) 描述了这一页面的数据。

[0079] 在步骤 S404 中,生成形成页面的数据中代表图像的绘制数据的标识描述 (图像绘制描述) (图像绘制描述生成)。

[0080] 在本发明说明的页面数据描述格式规格中,一个 <Image> 要素代表一个图像的绘制数据。另外,在属性 Data 中描述图像数据的内容,并且利用属性 X、Y、Width 和 Height 的坐标信息来描述在页面中绘制出的该图像的位置。在页面中存在多个图像的情况下,这意味着按出现的顺序来重写各个图像数据。在属性 Data 中描述的是已经被已知方法压缩了的图像数据,例如,在彩色或灰度级图像数据的情况下,图像数据为已经被 JPEG 压缩了的代码串;而在二值图像数据的情况下,图像数据为已经被 MMR 压缩了的代码串。

[0081] 在图 6 的描述例子中,在标识 603 中,描述了 X = “0”、Y = “0”、Width = “1680” 和 Height = “2376”,从而将已经在图 2 的步骤 S203 中选择出的文档的第一页面上的扫描图像绘制到整个页面上。另外,描述了令转换成 JPEG 压缩图像的代码串的文本后的字符串

为属性数据值的 <Image> 要素 603 (顺便提及, 在图 6 中, 为了简单地示出图, 省略 Data 属性的字符串的一部分)。

[0082] 在步骤 S405 中, 生成形成页面的数据中代表字符的绘制数据的描述 (字符绘制描述) (字符绘制描述生成)。

[0083] 在本发明说明的页面数据描述格式规格中, 一个 <Text> 要素描述垂直的或水平的每一行字符的绘制数据。如下是在 <Text> 要素中描述的属性数据。

[0084] ●属性 Direction, 表示字符串的垂直书写 / 水平书写 (顺便提及, 在没有指定 Direction 的情况下, 默认为水平书写)

[0085] ●属性 X 和 Y, 指定字符开始位置的坐标

[0086] ●属性 FONT, 指定绘制字符时所应用的字体数据的 ID

[0087] ●属性 Size, 指定字体大小

[0088] ●属性 Color, 利用 R 成分值、G 成分值、B 成分值和代表透明度的阿尔法 (alpha) 通道值这四个值的集合来指定绘制时的字符颜色

[0089] ●属性 String, 指定字符串的内容 (字符代码串)

[0090] ●属性 Cwidth, 指定 String 中从每个字符到下一字符的给送宽度

[0091] ●属性 CGlyphId, 指定在绘制时使用的 String 中每个字符的字符形状数据, 即字形的 ID。

[0092] 这里, 形成 <Text> 要素的字符串是通过将已经在图 2 的步骤 S205 中生成的字符代码串的数据进一步分割成各个字符行而获得的字符串, 其中该字符行是垂直或水平延伸的字符的集合。顺便提及, 在还没有定义属性 Font 的情况下, 将对应于默认字体 ID 的字符形状用作所有字符共用的字符形状。

[0093] 在图 6 的描述例子中, 两个 <Text> 标识 604 和 605 涉及第一页面的字符绘制描述, 并且是分别对应于图 11 的字符代码串数据 1101 和 1102 的描述。

[0094] 例如, 在与图 11A 的数据 1101 中的水平书写的字符串“A、I、U (每个均为日文字符)”的三个字符相对应的 <Text> 要素描述 604 中, 将每个属性值按如下进行描述。

[0095] ●将属性 X 和 Y 指定为 X = “236”和 Y = “272”, 作为这三个字符的边界矩形的左上处的坐标。

[0096] ●没有定义字体数据 ID 的属性 Font。

[0097] ●通过估计行中字符的高度, 将字体大小的属性 Size 指定为“97”像素。

[0098] ●将属性 Direction 指定为水平书写“Horizontal”。

[0099] ●将绘制时字符颜色的属性 Color 指定为“0,0,0,255”, 意味着 R 成分值 = G、成分值 = B、成分值 = 0 和 alpha 通道 = 255 (即, 指定了透明色)。

[0100] ●将指定字符串内容 (与各个字符相对应的字符代码的序列) 的属性 String 指定为“0x2422,0x2424,0x2426”。在上述的例子中, 当识别出如图 7 所示的日文字符的字符时, 描述字符代码, 然而, 当识别出如图 18 所示的文档图像中的英文字符串的字符时, 当然也会描述英文字符代码。

[0101] ●将指定每个字符的给送宽度的属性 CWidth 指定为“104,96,59”, 对于左边的两个字符, 每个字符的给送宽度对应于与其右侧的下一字符的左端之间的坐标差, 对于最后一个字符, 其给送宽度对应于其字符宽度本身。

[0102] ●作为每个字符的字符形状数据的、指定字形的 ID 的属性 CGlyphId 通常被指定为与每个字符的字符形状数据相对应的字形的 ID。在本实施例中,然而,由于在扫描图像上绘制了透明色字符的字符形状,任何字符形状都在用户可见度之外。因此,在本实施例中,即使字符是不同的字符(字符种类),通过相同字形 ID 的指定,只需要少量的字符形状数据(字体数据)。因此,在图 6 的例子中,将属性 CGlyphId 描述为相同的属性值“0,0,0”。另外,由该字形 ID 所指定的字符形状可以是简单的形状(例如,矩形)。顺便提及,下面将说明字形形状的细节。

[0103] 顺便提及,上述的属性值是范例,并且可以描述为具有相同含义的其它值。例如,优选地根据像素高度和图像分辨率,不利用像素数量而利用点数量的值等来描述代表字体大小的属性 Size。另外,在本发明的说明中,要绘制的字符被指定为透明色,不能被用户看到,因此使得要绘制的字符串可以不正好重叠在相应的字符图像上。例如,可以在相应的字符图像的下端部处,绘制透明字符串。例如,在图 6 的例子 604 的情况下,假定 X = “236”, Y = “368”并且 Size = “10”,在字符图像的下端部处绘制低的高度的透明字符串。此时,要绘制的透明字符串的大小(高度)为小于字符图像的大小的预定大小(例如,10)。

[0104] 将在后面利用搜索关键字进行搜索时使用要绘制的透明字符串,并且突出显示(例如,以改变后的颜色来显示)与搜索关键字相一致的字符串。由于在与相应的字符图像的位置基本相对应的位置绘制透明字符串,尽管在搜索时使用透明字符串来做出搜索,在用户看来如同搜索了字符图像。因此,在使用在这样的搜索时突出显示字符的应用程序的情况下,即使在相应的字符图像的下端部处绘制出透明字符串,在搜索时,就像相应的字符图像被加以强调那样来突出显示并指定该透明字符串,而不会产生问题。顺便提及,透明字符串的绘制位置不限于下端部,也可以描述字符串使得透明字符串被绘制到字符图像的上半部分或下半部分的位置。

[0105] 现在,在步骤 S406 中,说明表示页面描述的结束的</Page>。

[0106] 在步骤 S407 中,判断是否存在未被描述的页面。在存在未被描述的页面的情况下,从步骤 S403 开始重复操作;在不存在未被描述的页面的情况下,操作进入到步骤 S408。在图 6 的描述例子中,还对第二页面执行步骤 S404 到 S406 的处理,并且将描述 607 到 610 的部分。

[0107] 在步骤 S408 中,生成包括该电子文档中绘制所使用的所有字形的字体数据的内容的描述(字符形状数据描述)(字符形状数据描述生成)。在本发明的页面数据描述格式规格中,将包括在字体数据中的每个字形描述为夹在和之间的范围中的<Glyph>要素。在要素中,包括表示该字体的种类的属性 ID。此外,在<Glyph>要素中,包括表示字形的种类的属性 ID 和表示对应于该 ID 的字形(字符形状)的属性 Path。在<Glyph>要素中的属性 Path,在令左下为原点的 1024×1024 的绘制矩形单位中,是使用直线或曲线函数来代表字形的描述。

[0108] 在图 6 的描述例子中,在要素 611 和 612 中,分别在要素 611 和 612 中定义了字体 ID = “Font 01”的字体和字体 ID = “Font 02”的字体。在各自的内容中,定义了具有 ID = “0”的字形的一个种类。代表要素 611 的“Font01”的字形的字符形状的 Path 属性“M0,0V-1024H 1024V 1024f”的含义如下。

[0109] “移动到原点(0,0),向上绘制 1024 单位的垂直线,向右绘制 1024 单位的水平

线,向下绘制 1024 单位的垂直线,并且从当前点向起始点绘制线以填充由这些线包围的范围。”

[0110] 即,这是代表填充 1024×1024 范围的方形字形的描述。

[0111] 另外,代表 要素 612 的“Font02”的字形的字符形状的 Path 属性“M0,0V-64H1024 V64f”的含义如下。

[0112] “移动到原点 (0,0),向上绘制 64 单位的垂直线,向右绘制 1024 单位的水平线,向下绘制 64 单位的垂直线,并且从当前点向起始点绘制线以填充由这些线包围的范围。”

[0113] 即,这是代表填充在绘制矩形单位的下部处 1024×64 区域的水平直线字形的描述。

[0114] 顺便提及,图 6 的 要素 611 和 612 的描述只是一个例子,并且优选定义垂直直线、波浪线、虚线、三角形、圆形或四边形等其它简单的字符形状。

[0115] 接下来,在步骤 S409 中,描述表示电子文档的结束的 </Document>,并且结束电子文档的生成。将已经生成的电子文档作为文件存储在图像处理装置 100 的存储器 103 或硬盘 104 中。在存储时,优选使用已知的文本压缩技术进行压缩。

[0116] 这里,返回图 2,进行描述。在步骤 S209 中,将已经在步骤 S208 中生成的电子文档通过已经在步骤 S201 中指定的发送方法传送到作为指定的发送目的地的图像处理装置 110。发送处理优选采用已知的技术,从而将省略描述。

[0117] 如上所述,图像处理装置 110 通过网络接口 114 接收已经被传送的电子文档,并将电子文档存储在硬盘 114 中。

[0118] 这里,用于在硬盘内部指定要存储的电子文档的信息优选是文件名等任何 ID 信息。在本发明的说明中,指派关于接收时间的字符串作为例子。尽管存在选择并自动给出唯一的编号或者用户已经预先输入了文件名作为用户在图像处理装置 100 生成时指定的信息或其它方法,由于这些方法是不同于本发明的基本特征的处理,所以省略详细的说明。

[0119] 现在,将根据图 3 的流程图说明使用图 1 的图像处理装置 110 来搜索并浏览电子文档的处理的例子。这里,尽管将说明使用图像处理装置 110 来进行搜索的例子,但并不限于该例子,并且可以配置为使用图像处理装置 100 来做出搜索。

[0120] 在步骤 S 301 中,用户使用 UI 115 从电子文档所保持 (hold) 的字体 ID 的列表中,选择在没有定义 Font 属性的情况下使用的默认字体 ID。图 16 示出了选择画面 UI 的一个例子,该选择画面显示要被搜索的电子文档所持有的字体的列表以及被选择的字体的预览图像 (图中,黑色四边形部分示出 Font01 的字形)。用户可以使用这样的 UI 来选择默认字体 ID。顺便提及,做出图 3 中步骤 S301 的字体 ID 指定处理的序列顺序 (sequential order) 是范例,而可以是在步骤 S307 之前的任意步骤。

[0121] 在步骤 S302 中,为了从存储在图像处理装置 110 中的电子文档组中搜索出所期望的电子文档的字符串,从 UI 115 输入用户认为包括在该电子文档的文本中的搜索关键字。这里已经被输入的该字符串的长度为 k。

[0122] 在步骤 S303 中,对于所有存储在图像处理装置 110 的硬盘 114 中的电子文档文件,判断是否存在未被搜索的电子文档文件。在存在未被搜索的电子文档文件的情况下,指定其中的一个电子文档文件。另外,在该电子文档文件被压缩后的情况下,解压该电子文档文件并进入步骤 S304。在不存在未被搜索的电子文档的情况下,操作进入到 S313,其中通

知用户已经结束对所有电子文档的搜索。

[0123] 在步骤 S304 中,令已经在 S303 中指定的电子文档中的文本数据为目标,为利用搜索关键字进行搜索做出准备。这里,排列文档中的文本(字符代码),初始化搜索开始位置 n ,即,设置 $n = 0$ 。

[0124] 这里,将在下面说明步骤 S304 中处理的例子的详情。

[0125] 在利用 XML 解析器(parser)来分析电子文档数据的阶段中,当<Text>要素出现时,获得在属性 String 中描述的字符代码串。另外,对于<Text>要素中 String 属性中描述的每个字符,将其字符代码以及在电子文档数据中描述了该字符代码值的位置的组添加给字符代码串表。这里,描述了该字符代码值的位置是表示从电子文档数据的开头算起到描述该字符代码的第一个字符的字符数量的值。

[0126] 这里,为了清楚,在图 12 中示出了从图 6 的电子文档生成的字符代码串表的例子。在该例子中,在图 6 中的<Text>要素 604 的属性 String 中,描述了三个字符代码“0x2422”,“0x2424”和“0x2426”。这里,分别从该电子文档的开头算起的第 1093 个字符位置、第 1100 个字符位置和第 1107 个字符位置开始描述这三个字符代码。同样,基于<Text>要素 605 和 609,对于剩下的 6 个字符代码的描述位置,也获得描述位置,并且生成如图 12 所示的字符代码串表。顺便提及,在图 12 中,此时,以从 0 开始的顺序给出字符串编号(No.)。

[0127] 接着,在步骤 S305 中,对于字符代码串表,令搜索开始位置 n 为原点,判断字符代码串表与搜索关键字的字符代码串是否相一致。在检测到与搜索关键字的字符代码串相一致的部分的情况下,令此时的变量 n 为相一致的字符串的开头位置,操作进入到步骤 S306。

[0128] 反之,在步骤 S305 中判断为不一致的情况下,操作进入到步骤 S310,在步骤 S310 中判断是否搜索了该字符代码串表中的所有字符。在判断为存储在字符代码串表中的所有字符代码串的搜索已经结束的情况下,操作进入到步骤 S312,在步骤 S312 中通知作为当前搜索对象的电子文档的搜索结束。另一方面,在判断为所有搜索还没有结束的情况下,操作进入到步骤 S311,在步骤 S311 中变量 n 增加 1,并且操作返回到步骤 S305,在步骤 S305 中,判断在下一个搜索开始位置 n ,字符代码串表与搜索关键字的字符代码串是否相一致。顺便提及,在步骤 S310 中,在令存储在字符代码串表中的字符代码的总数量为 N 的情况下,当 $n < (N-k)$ 时,可以判断为所有搜索还没有结束;并且当 $n \geq (N-k)$ 时,可以判断为所有搜索已经结束。

[0129] 对于图 12 的字符代码串表的例子,例如,在从开头起扫描关键字字符“KA, KI(每个均为日文字符)”的字符代码串“0x242b 和 0x242d”,并且检测到了相一致的部分的情况下,提取 $n = 3$ 作为最初的相一致的字符串的字符串编号。另外,在下面说明的 S307 中进一步继续搜索,并且检测出了下一个相一致的字符串的情况下,提取位置 $n = 6$ 。顺便提及,在步骤 S303 到 S305 中使用了令这些字符代码和描述位置成为一对的字符代码和表的处理只是一个例子,并且可以采用其它方法。

[0130] 接下来,在步骤 S306 中,指定了与相一致的字符串编号 n 相对应的字符串数据属于电子文档的那个页面。例如,假定在分析电子文档数据时,判断出在哪个<Page>要素中描述了<Text>要素,则可以从 Number 属性识别出页面编号。因此,从图 12 的字符代码串表获取与已经在步骤 S306 中所指定的位置 n 相对应的字符串的描述位置,并且依赖于该描述位置位于哪些<Page>要素之间,可以指定出该字符串所属的页面。顺便提及,在步骤 S304

中分析电子文档数据时,在判断出各 <Text> 要素被描述在哪个页面,并且该判断结果已经被预先存储在图 12 的字符代码串表的情况下,基于字符串编号可以容易地指定页面编号。顺便提及,在步骤 S305 中的相一致的字符串的检测方法或者在步骤 S306 中的页面编号的指定方法不限于上述的例子。

[0131] 在步骤 S307 中,基于在步骤 S306 中确定的页面中所包括的绘制描述,在 UI 115 上显示页面的绘制结果(电子文档显示)。此时,在绘制字符串编号(No.)在 n 到 $(n+k-1)$ 范围中的字符时,为了使用户容易地识别出对应于该字符的点,突出显示该字符并绘制出该字符。下面将说明与该搜索关键字相一致的部分的突出显示的绘制的详情。

[0132] 这里,将根据图 5 的流程图说明要在步骤 S307 中执行的页面绘制处理的详情。

[0133] 在步骤 S501 中,从 <Page> 要素中的 Width 和 Height 属性的值确定将作为绘制结果的页面图像的大小。

[0134] 在步骤 S502 中,获得能够存储页面图像的像素信息的量的存储器区域。

[0135] 在步骤 S503 中,以从开头起的顺序提取出 <Page> 要素的子要素中未被处理的一个要素,并且判断已经被提取出的该未被处理的要素的种类。在该未被处理的要素被判断为 <Image> 的情况下,操作进入到步骤 S504。另一方面,在该未被处理的要素被判断为 <Text> 的情况下,操作进入到步骤 S505。在已经处理了该 <Page> 要素中所有的子要素的情况下,操作进入到步骤 S517。

[0136] 在步骤 S504 中,对描述为 <Image> 要素的 Data 属性值的压缩图像进行解压缩。此外,缩放(scale)该解压缩后的图像以使得该图像被保留在由 X、Y、Width 和 Height 属性来代表的页面图像中的整个绘制矩形区域。另外,将该图像重写在已经在上述步骤 S502 中获得的页面图像存储器的区域中。然后,操作返回步骤 S503。

[0137] 另一方面,在步骤 S505 中,从描述在要被处理的 <Text> 要素中的每个属性获得字符开始位置(X、Y)、字体 ID(F)、字符大小(S)和字符颜色(C)。另外,获得在该 <Text> 要素中描述的字符的数量(N)。顺便提及,在没有如图 6 的例子的中所示出的定义了字体 ID 的情况下,使用已经在上述的 S301 中指定的默认字体 ID。

[0138] 在步骤 S506 中,获得用于字形图像生成的存储器区域。这里,将获得用于 1024×1024 像素的二值图像的存储器。

[0139] 在步骤 S507 中,将正在被处理的字符的计数器 i 初始化为 1。

[0140] 在步骤 S508 中,判断是否 $i > N$ 。当 $i \leq N$ 时,操作进入到步骤 S509;并且当 $i > N$ 时,结束该 <Text> 要素的处理以返回步骤 S503。

[0141] 在步骤 S509 中,从 <Text> 要素的属性 String 获得第 i 个字符的字符代码(P),并且从属性 CGlyphId 获得第 i 个字符的 GlyphId(Q)。

[0142] 在步骤 S510 中,从电子文档检测出字体 $Id = F$ 的 要素描述,并且还从该 要素描述的子要素中的字形 $Id = Q$ 的 <Glyph> 要素获得 Path 属性。

[0143] 在步骤 S511 中,根据已经在步骤 S510 中获得的 Path 属性值,对于已经在步骤 S510 中获得的字形图像生成存储器,生成字形二值图像。顺便提及,字形二值图像是例如通过令被绘制部分为 1,并且令不被绘制的部分为 0 来代表的图像。顺便提及,在本实施例中,将在后面用透明色绘制被绘制的部分 1。

[0144] 在步骤 S512 中,缩放字形的二值图像,以使图像成为基于字符大小的属性值(S)

的矩形大小。

[0145] 在步骤 S513 中,在从页面图像存储器的位置 (X, Y) 开始的矩形区域中,绘制了缩放后的字形二值图像的信息。利用如下的表达式来定义当在页面图像上绘制二值图像时每个像素的像素值。此时,依赖于二值图像中相应的像素的像素值是 1 还是 0,分别将在页面图像的目标区域中的像素值 (r, g, b) 改变为如下的 (r', g', b')。

[0146] ●在字形二值图像的像素值为 0 的像素的情况下:

[0147] $(r', g', b') = (r, g, b)$

[0148] ●在字形二值图像的像素值为 1 的像素的情况下:

[0149] $(r', g', b') = (F(r, Cr), F(g, Cg), F(b, Cb))$

[0150] 其中: $F(r, Cr) = (r \times A + Cr \times (255 - A)) / 255$, $F(g, Cg) = (g \times A + Cg \times (255 - A)) / 255$, $F(b, Cb) = (b \times A + Cb \times (255 - A)) / 255$ 。另外,A 是字符颜色 C 的 alpha 通道值,Cr、Cg 和 Cb 分别是字符颜色 C 的 RGB 值。顺便提及,在将 255 指定为 alpha 通道值的情况下,由于该字形二值图像是透明的,即使对于字形二值图像的像素值为 1 的像素,也有 $(r', g', b') = (r, g, b)$ 。

[0151] 在步骤 S514 中,使用例如图 12 的字符代码串表来判断正在被处理的第 i 个字符是否是字符串编号 (No.) 在 n 到 (n+k-1) 范围中的字符。更具体地,由于根据字符代码串表确定在 n 到 (n+k-1) 范围中的字符的描述开始位置,基于正在被处理的字符 i 的开始位置是否与任何一个相一致来做出判断。在字符 i 是 n 到 (n+k-1) 范围中的字符的情况下,操作进入到步骤 S515;在其它情况下,操作进入到步骤 S516。

[0152] 在步骤 S515 中,做出突出显示以表示正在被处理的字符在作为搜索字符串被检测的范围中。具体地,相应的字形二值图像的像素值为 0 的像素保持不变,对于相应的字形二值图像的像素值为 1 的像素,分别将像素值 (r, g, b) 改变为如下的 (r', g', b')。

[0153] $(r', g', b') = (G(r), G(g), G(b))$

[0154] 其中: $G(r) = 255 - r$, $G(g) = 255 - g$, 以及 $G(b) = 255 - b$ 。

[0155] 顺便提及,上述的突出显示是范例。例如,被突出显示的宽度不是字形二值图像的宽度,而使用指定每个字符的给送宽度的属性 Cwidth 的值,来无间隔地实心填充字符串。

[0156] 另一方面,在步骤 S516 中,使 X 加上 Cwidth 属性的描述为第 i 个字符的字符的给送宽度,还使 i 加 1 ($i = i + 1$),并且从步骤 S503 开始重复处理。在步骤 S503 的阶段,当不存在未处理的子要素时,操作进入到步骤 S517。

[0157] 在步骤 S517 中,将作为一个页面的绘制结果的页面图像存储器的内容,即作为 <Page> 要素中的 <Image> 和 <Text> 要素描述的绘制结果传送到 UI 115 的显示缓冲器并将其显示。

[0158] 接着,说明在基于图 6 的电子文档中第一页面上的绘制描述,执行图 5 的流程图示出的图 3 中步骤 S307 的处理的情况下的例子。

[0159] 通过在步骤 S501 中的处理,从图 6 中第一页面的 <Page> 要素 602 的属性值 Width = "1680", Height = "2376", 确定页面的图像大小为 1680 × 2376 像素。

[0160] 通过在步骤 S502 中的处理,例如,在利用 RGB 24 比特颜色来表现页面图像的情况下,获得 1680 × 2376 × 3 比特的存储器。

[0161] 通过在步骤 S504 中的处理,将在图 6 的 <Image> 要素 603 的 Data 属性值中描述

的压缩代码解压缩为图像数据,将该图像数据重写在整个页面图像存储器上。顺便提及,在该例子中,图像数据本来具有与页面的大小相同的大小的 1680×2376 像素,所以没有应用缩放处理。

[0162] 通过在步骤 S505 中的处理,从图 6 中的 $\langle \text{Text} \rangle$ 要素 604 获得 $X = "236"$ 、 $Y = "272"$ 、字符数量 $N = 3$ 、字符大小 = "97" 和字符颜色 = "0,0,0,255"。在 $\langle \text{Text} \rangle$ 要素 604 中,没有定义 Font 要素,从而指定在 S 301 中已经指定的默认的字符代码。这里,将在步骤 S301 中指定字体 ID = "Font01"。

[0163] 通过在步骤 S 509 中的处理,首先,获得 $\langle \text{Text} \rangle$ 要素 604 的 String 属性中的第一个字符代码 = 0x2422 和 CGlyphId = "0"。

[0164] 由于指定了默认的字体 ID = "Font01",基于在图 6 的 $\langle \text{Font} \rangle$ 要素 611 中的 Id = "0" 的 $\langle \text{Glyph} \rangle$ 要素中所描述的 Path 属性来创建要在步骤 S 511 中生成的字形二值图像。具体地,基于 Path 属性的描述,用 1 填充 1024×1024 像素的整个 Glyph 图像区域。

[0165] 顺便提及,由于图 6 的电子文档中描述的 $\langle \text{Text} \rangle$ 要素 604 和 605 中字符的 CGlyphId 都是 "0",结果,对于所有字符,步骤 S511 的处理结果是相同的。因此,将已经在步骤 S511 中生成的字形图像临时存储在存储器中,并且在绘制其它字符时,还将使用该临时存储的字形图像。

[0166] 在步骤 S512 中,利用字符大小 = "97" 将字形的字符图像缩放为 97×97 像素。

[0167] 在步骤 S513 中,将从页面图像的位置 (236, 272) 开始的 97×97 像素的矩形范围作为缩放后的字形的字符图像的绘制目标区域。在图 6 的例子中,然而,由于字符颜色 = "0,0,0,255",即 alpha 值 $A = 255$,即使字形二值图像的相应的像素值是 1,也总是 $(r', g', b') = (r, g, b)$ 。即,在步骤 S513 的处理前后,页面图像中该矩形区域中的像素值不发生改变。

[0168] 在步骤 S514 中,基于字符代码串表判断图 6 中的 $\langle \text{Text} \rangle$ 要素 604 中的第一个字符是否是已经在图 3 的步骤 S305 中获得的、在位置范围 n 到 $(n+k-1)$ 中的字符。

[0169] 这里,例如,从图 6 的电子文档生成图 12 的字符代码串表,并且假定已经在图 3 的步骤 S305 中判断为与关键字相一致的字符串的位置在 3 到 4 的范围中。此时,在图 6 中的 $\langle \text{Text} \rangle$ 要素 604 中第一个字符代码描述的开头位置是如图所示的第 1093 个。该数值与字符代码串表的 3 到 4 的范围中的字符的描述位置都不一致,从而通过步骤 S516 处理进入到下一个字符。

[0170] 之后,处理继续,并且在图 6 中的 $\langle \text{Text} \rangle$ 要素 605 中的第一个字符的处理中,在步骤 S514 中,该第一个字符的位置判断为与字符代码串表的 3 到 4 的范围中字符的开始位置相一致。在步骤 S515 中,执行突出显示处理。

[0171] 关于该字符,在步骤 S515 中,对于从页面图像存储器的位置 (236, 472) 开始的 92×92 的区域中相应的字形二值图像的像素值为 1 的像素,将该像素的像素值 (r, g, b) 分别改变为 $(G(r), G(g), G(b))$ 。

[0172] 在如上述所有 $\langle \text{Text} \rangle$ 要素被绘制出之后的页面图像如图 13 所示。即,与在步骤 S305 中已经被判断为相一致的范围中的字符相对应的区域成为每个矩形中亮度被反转的状态。反之,与剩下的字符相对应的区域成为与利用 $\langle \text{Image} \rangle$ 要素绘制的图像数据一样。在上述的例子中,说明了日文文档的情况,然而在英文文档的情况下也可以进行类似处理。例

如,当对图 18 中示出的图像执行图 4 中示出的电子文档数据生成处理时,生成用字母字符代码来描述图 6 中的电子文档的 604 和 605 的电子文档。当利用例如字符串“EF”来搜索电子文档数据时,执行图 5 中示出的绘制处理,并且如图 19 所示,反转显示与被搜索的字符串相对应的矩形区域中的图像。

[0173] 因此,由于突出显示了已经被搜索到的字符串,用户可以仅通过观看在步骤 S 307 中显示的页面上的图像,容易地判断出页面中搜索关键字位于何处。

[0174] 另一方面,依赖于查看器应用程序的种类,存在以不同的显示方法来突出显示的一些情况。即,依赖于图 5 中的步骤 S515 的对字符部分突出显示处理的方法,在一些情况下,可能不会做出合适的突出显示,从而用户不能识别出被描述的字符。例如,在突出显示处理中,假定存在这样的查看器,该查看器使得相应的字形二值图像的像素值为 0 的像素保持原样,并且对于字形二值图像的像素值为 1 的像素,使得像素值 (r, g, b) 分别改变成已经预先确定的颜色 (例如, (0, 0, 0))。在这种情况下,由于根据本实施例的 Font01 使用实心填充的四边形的字形,当突出显示时,将显示如图 17 中的页面图像。当处于如图 17 的状态时,不能看到字符图像,并且用户可见度变差。

[0175] 因此,在如上所述的本实施例中,假定了使用这样的查看器应用程序的情况,在步骤 S208 中生成电子文档时,将存储具有不同字形的多种字体数据。因此,在用户通过再次使用图 3 的步骤 S301 中的字体 ID 指定处理来指定其它字体 ID 的情况下,突出显示的方法可能被改变。

[0176] 图 14 中示出了在步骤 S301 中将“Font02”指定为默认字体 ID 并进行突出显示的情况下的页面图像显示的例子。当通过使用在图 6 的描述 612 中指定的字形二值图像来绘制出分别改变为 (0, 0, 0) 的像素值 (r, g, b) 的突出显示时,生成如图 14 的页面图像。即,由于在“Font02”中使用的字形是在绘制矩形单元下部处填充 1024×64 的区域的水平直线字形,在该部分处的像素值为 (0, 0, 0) 的情况下,将做出突出显示,像如图 14 中字符为被添加了下划线的字符一样。因此,用户可以容易地判断出已经被搜索出的字符串在页面中位于何处,还可以确保字符图像的可见度。应注意,尽管图 14 示出包括日文字符串的文档图像的例子,还可以类似地处理包括英文 (字母) 字符的文档图像。例如,当利用字符串“EF”来搜索文档图像时,像如图 20 所示的给字符串提供下划线那样来突出显示搜索到的字符串。

[0177] 这里,说明返回到图 3。在步骤 S308 中,由用户选择搜索和浏览处理是否结束,或者是否另外将其它搜索点作为对象继续搜索。在由用户选择结束搜索的情况下,图 3 的处理结束;并且在选择继续搜索的情况下,操作进入到步骤 S309。

[0178] 在步骤 S309 中, $n = n + k$, 并且操作返回到步骤 S305, 之后重复处理。

[0179] 如上所述,根据本发明的实施例 1,在将纸文档转换成电子文档时,做出描述,从而使得利用透明色在页面图像上绘制已经从页面提取出的字符,并且存储具有不时字符形状的多个字体。对于该电子文档,用户可以选择在突出显示中使用的默认字符形状,即,他/她可以做出字符形状的切换指示。因此,可以切换到最适合各个查看器应用程序或者其文档图像的字符形状,并且将该字符突出显示。可以在用户以高可见度确认与搜索关键字相一致的点被突出显示的页面显示的同时继续搜索。

[0180] 在根据本实施例的电子文档中,其内部部分存在对于一个字符的由简单的字符形

状形成的多种字体数据,当绘制文档中的透明字符时,可以选择上述简单的字符形状中的一种来绘制。另外,在每种字体数据中,可以对于多个字符种类共同使用一种字符形状。因此,尽管要使用的字体数据被保持在电子文档中,但只有少量的字符形状数据,从而可以将该电子文档的文件大小(数据量)最小化。另外,存储了多种字体数据,从而能在突出显示搜索时做到高可见度或操作性的显示。

[0181] 实施例 2

[0182] 现在,参照附图说明根据本发明的第二典型实施例(实施例 2)。

[0183] 图 15 是根据该实施例 2 生成的电子文档的例子。同上述的实施例 1 相同,图像处理装置 100 生成并发送电子文档,并且图像处理装置 110 接收、浏览并搜索电子文档。

[0184] 图 15 中的附图标记 1501 和 1513 表示代表电子文档的开始和结束的描述。附图标记 1502 和 1506 表示代表在第一页面上的绘制的开始和结束的描述。附图标记 1503 表示在第一页面上图像数据绘制的描述。附图标记 1504 和 1505 表示在第一页面上字符绘制的描述。另外,附图标记 1507 和 1510 表示代表在第二页面上的绘制的开始和结束的描述。附图标记 1508 表示在第二页面上的图像数据绘制的描述。附图标记 1509 表示在第二页面上的字符绘制的描述。附图标记 1511 和 1512 表示在该电子文档中使用的字体数据的描述。

[0185] 尽管在实施例 1 中,在图 3 的步骤 S301 中由用户选择默认的字体 ID,在实施例 2 中,生成电子文档从而使得进行浏览操作的应用程序(查看器)可以做出判断以选择字体 ID。在这种情况下,在图 4 的步骤 S408 中的字体数据描述中,将应用程序的用于判断的属性添加到要素。在本实施例中,可以从图 15 中要素 1511 和 1512 中的 Shape 属性,容易地判断出字体数据的形状(字体数据的特征)。在这种情况下,通过使用字体数据的该 Shape 属性作为判断参考,应用程序可选择适合于通过这一应用程序进行突出显示的显示字体 ID。顺便提及,上述附加的属性只是一个例子,除了字体数据的特征以外,还可以将进行浏览处理的应用程序的名称和种类作为属性来描述。在这种情况下,令该属性作为判断参考,例如,应用程序做出包括有该应用程序的名称的字体数据的判断,并且确定用于默认显示的字体 ID。

[0186] 根据实施例 2,在将纸文档转换成电子文档时,描述已经从页面提取出的字符以在页面图像上利用透明色来绘制该字符,描述了具有不同字符形状的多种字体,并且包括用于判断应用程序应该使用哪种字体的属性。对于该电子文档,应用程序(查看器等)可以判断上述属性并且自动选择字符形状,并且可以切换到适合于使用该应用程序来突出显示的字符形状以被显示出。此外,在用户确认与搜索关键字相一致的点被突出显示、并且还自动确保字符图像的可见度的页面显示时,他/她可以继续搜索。

[0187] 如上所述,描述根据实施例 2 的电子文档,从而使得在绘制出所有已经在文档中描述的透明字符时,应用程序(查看器等)判断属性,并且从多种字符形状中选择一种字符形状并进行绘制。另外,在每种字体数据中,可以对于多个字符种类共同使用一种字符形状。因此,尽管电子文档保持使用在该电子文档中的字体数据,但只有少量的字符形状数据,从而可以将该电子文档的文件大小(数据量)最小化。此外,由于存储了一些字体数据,可以在搜索的突出显示时做出最适合于该应用程序的显示。

[0188] 实施例 3

[0189] 另外,尽管在上述的实施例中,生成了在<Image>要素中描述对于扫描图像做出

例如 JPEG 压缩的整个图像以及在 <Text> 要素中描述透明文本的电子文档,但不限于这样的例子。

[0190] 例如,替代 <Image> 要素中对通过 JPEG 压缩整个扫描图像而获得的图像的描述,可以在字符区域或图形区域中存储通过对各颜色创建二值图像以进行 MMR 压缩而获得的图像,并且在其它区域存储经 JPEG 压缩后的图像。这样,分析包括在文档图像中的区域以适当地做出压缩处理的方法,可以采用在例如日本特开平 07-236062 或日本特开 2002-077633 中所描述的方法。通过将根据本发明的抑制透明文本绘制中所使用的字体数据的数据量的处理和该图像压缩处理结合,可以生成更好压缩的电子文档。

[0191] 此外,替代整个图像,优选地只将字符区域、图形区域、表区域和照片区域等的部分区域与位置数据一起存储。

[0192] 实施例 4

[0193] 此外,在上述的实施例中,如图 3 和 5 中所说明的,在搜索时,按从文档开头开始的顺序来搜索与关键字相一致的字符串,并且突出显示被最先检测出的字符串。另外,在配置成存在“搜索下一个”的指示的情况下,顺次地,搜索出下一个相一致的字符串并突出显示。这样,尽管在上述实施例中,按从开头开始的顺序来搜索与搜索关键字相一致的字符串,并且在每次找到搜索关键字时顺次突出显示,但是不限于这种情况。例如,优选配置成将包括在电子文档中的所有字符串与搜索关键字比较,指定出与该搜索关键字相一致的所有字符串,并且同时突出显示与该关键字相一致的所有字符串。

[0194] 至此,已经说明了根据本发明的各种典型实施例。

[0195] 还通过系统或装置的计算机(或 CPU 或 MPU)从记录了进行上述实施例中描述的流程图的操作的程序代码的记录介质读取并执行上述程序来实现本发明的目的。在这种情况下,从记录介质读出的程序代码自身使得计算机实现上述实施例的功能。因此,这些程序代码和存储并记录了程序代码的计算机可读记录介质也成为本发明的一部分。

[0196] 可以采用例如软盘(注册商标)、硬盘、光盘、磁光盘、CD-ROM、CD-R、磁带、非易失性存储器卡和 ROM 作为用于提供程序代码的记录介质。

[0197] 另外,通过由计算机执行已经被读出的程序来实现上述实施例的功能。另外,该程序的执行还包括基于程序的指示由在计算机上操作的 OS 等来执行部分或全部实际处理的情况。

[0198] 另外,还通过插入到计算机的扩展板或连接到计算机的扩展单元来实现上述实施例的功能。在这种情况下,首先,将已经从记录介质读出的程序代码写入到设置在插入到计算机的扩展板或连接到计算机扩展单元上的存储器中。之后,基于该程序代码的指示,设置在这样的扩展板或扩展单元上的 CPU 等执行部分或全部实际处理。同样,通过这样的扩展板或扩展单元的处理,实现了上述实施例的功能。

[0199] 尽管已经参考典型实施例说明了本发明,但应理解,本发明不限于所公开的典型实施例。所附权利要求书的范围符合最宽的解释,以涵盖所有的修改、等同的结构和功能。

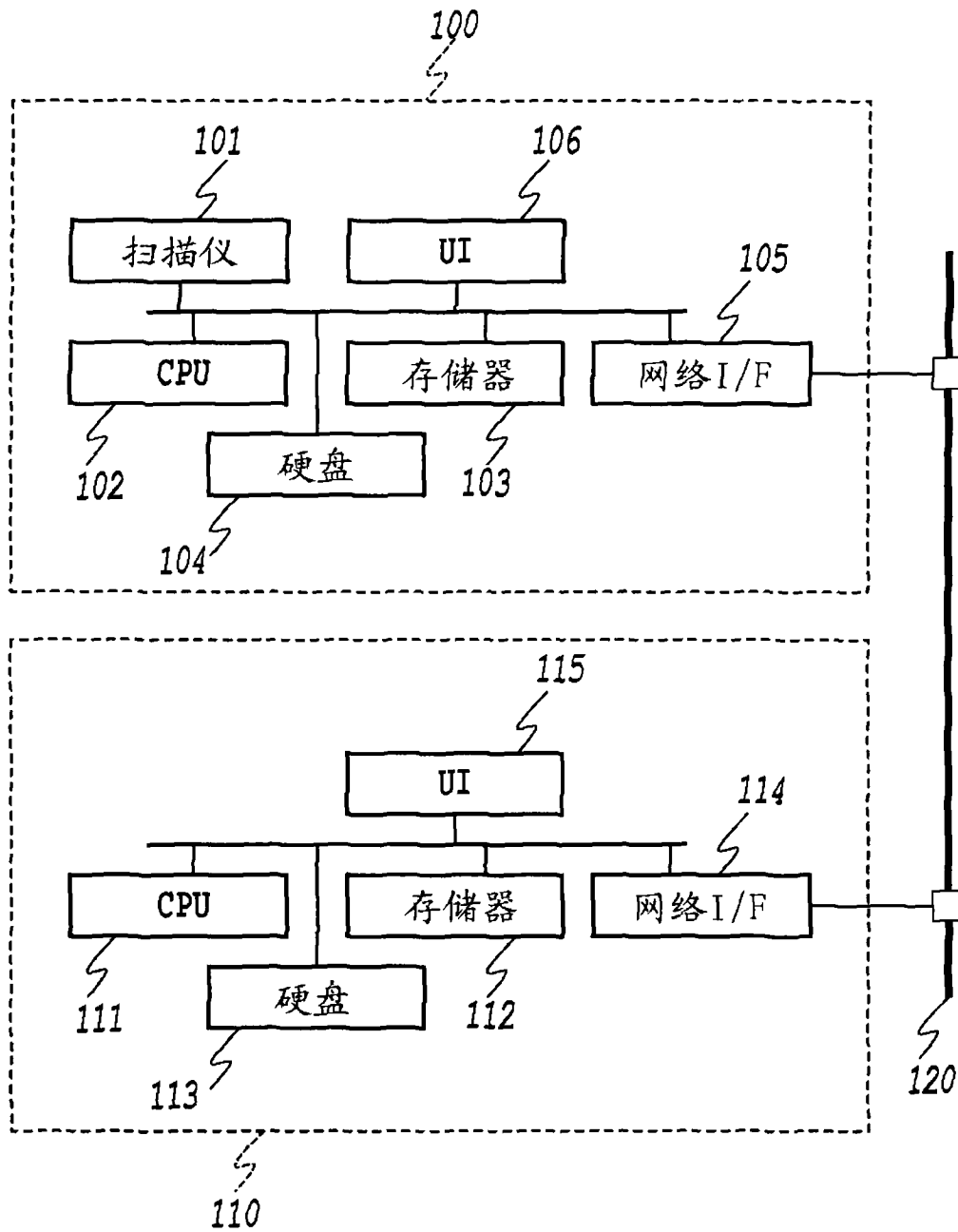


图 1

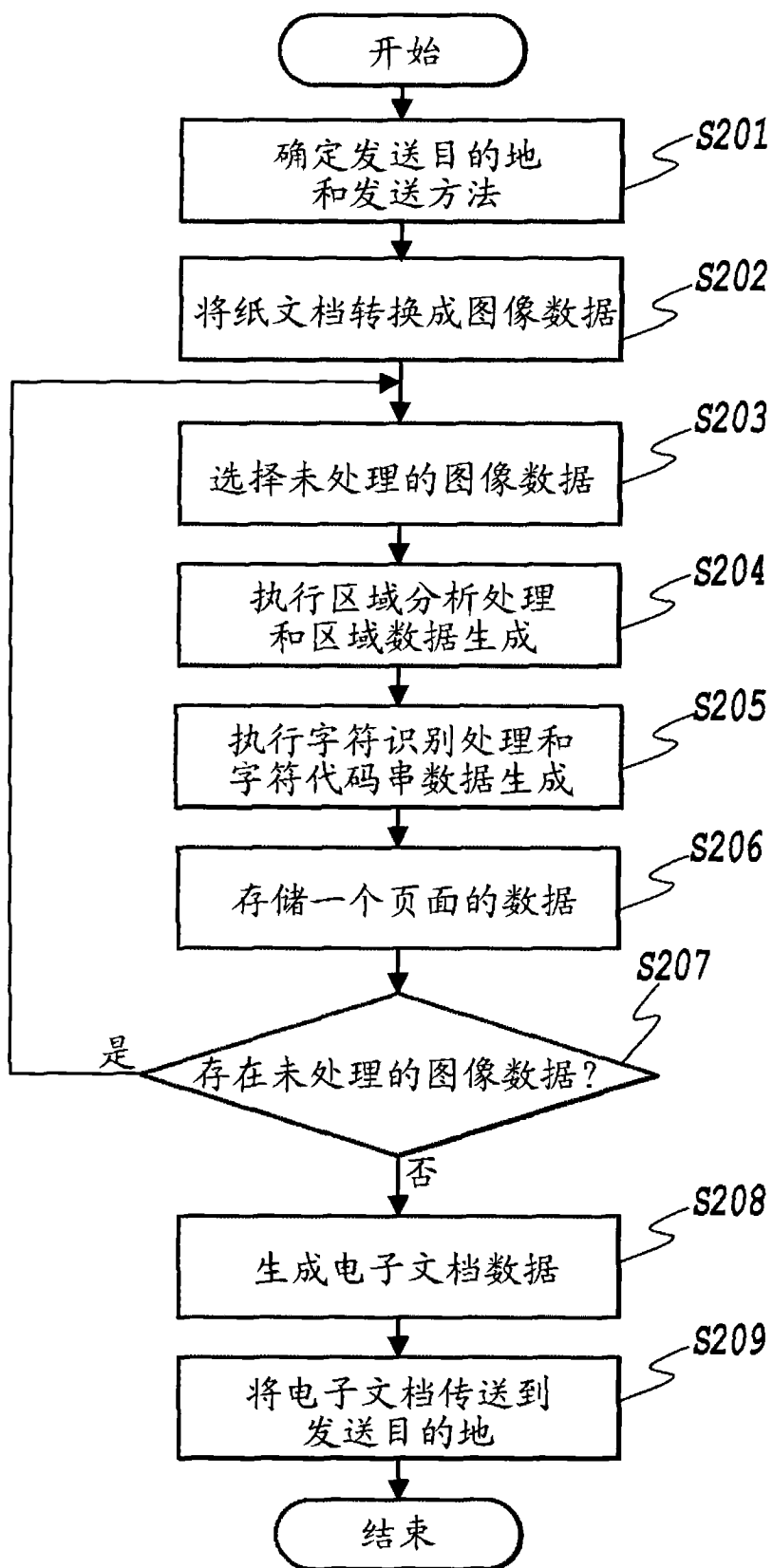


图 2

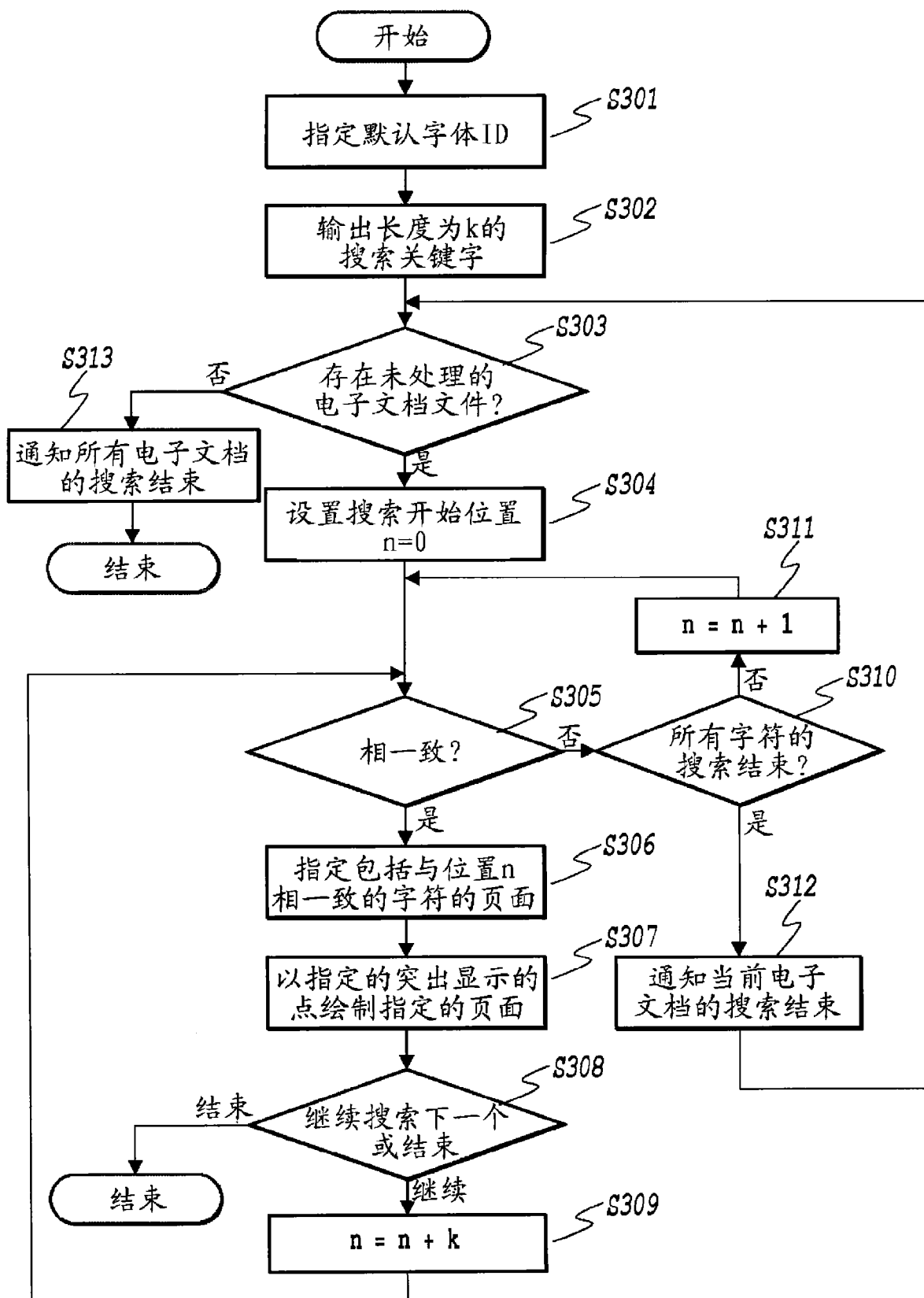


图 3

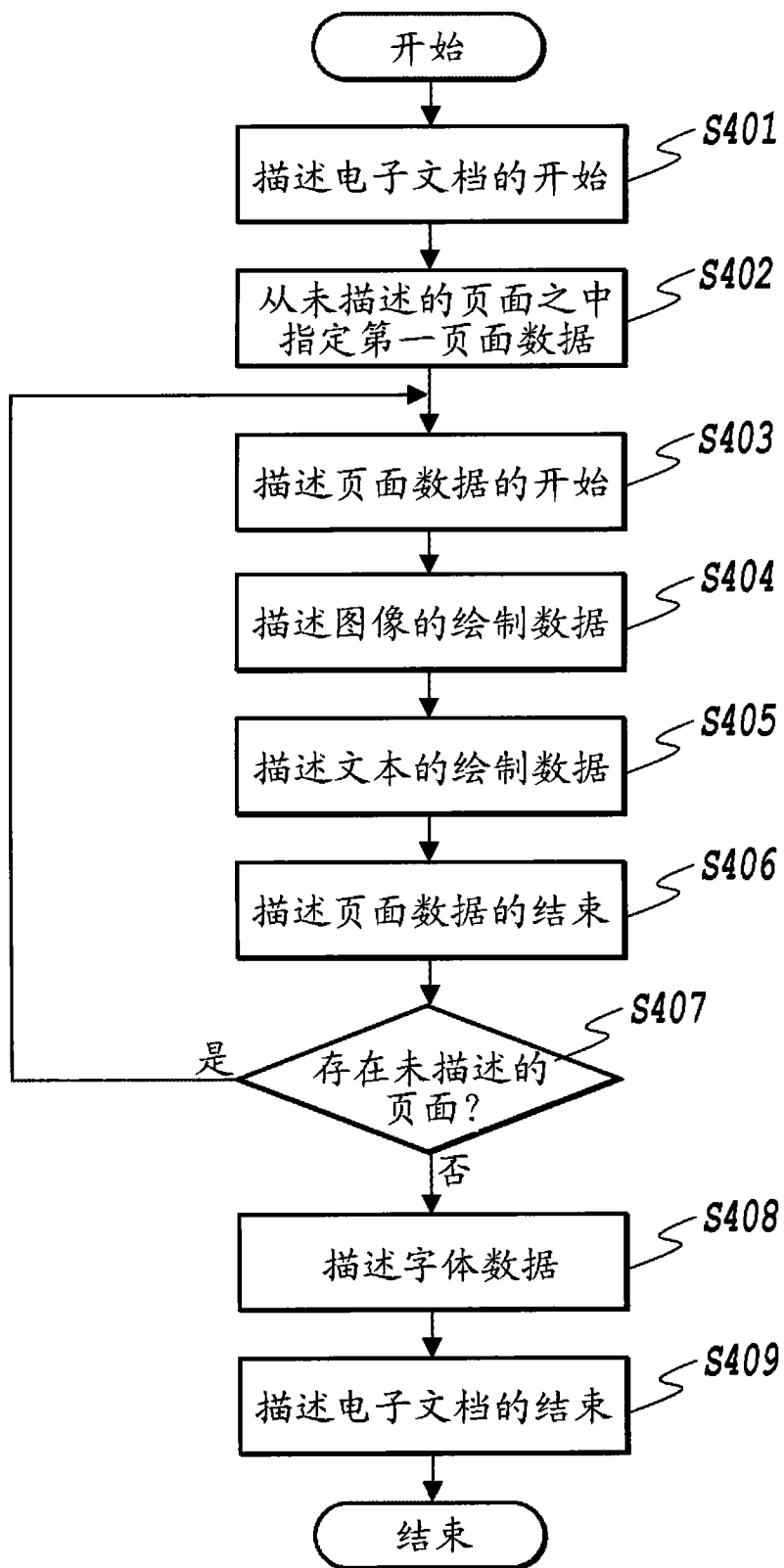


图 4

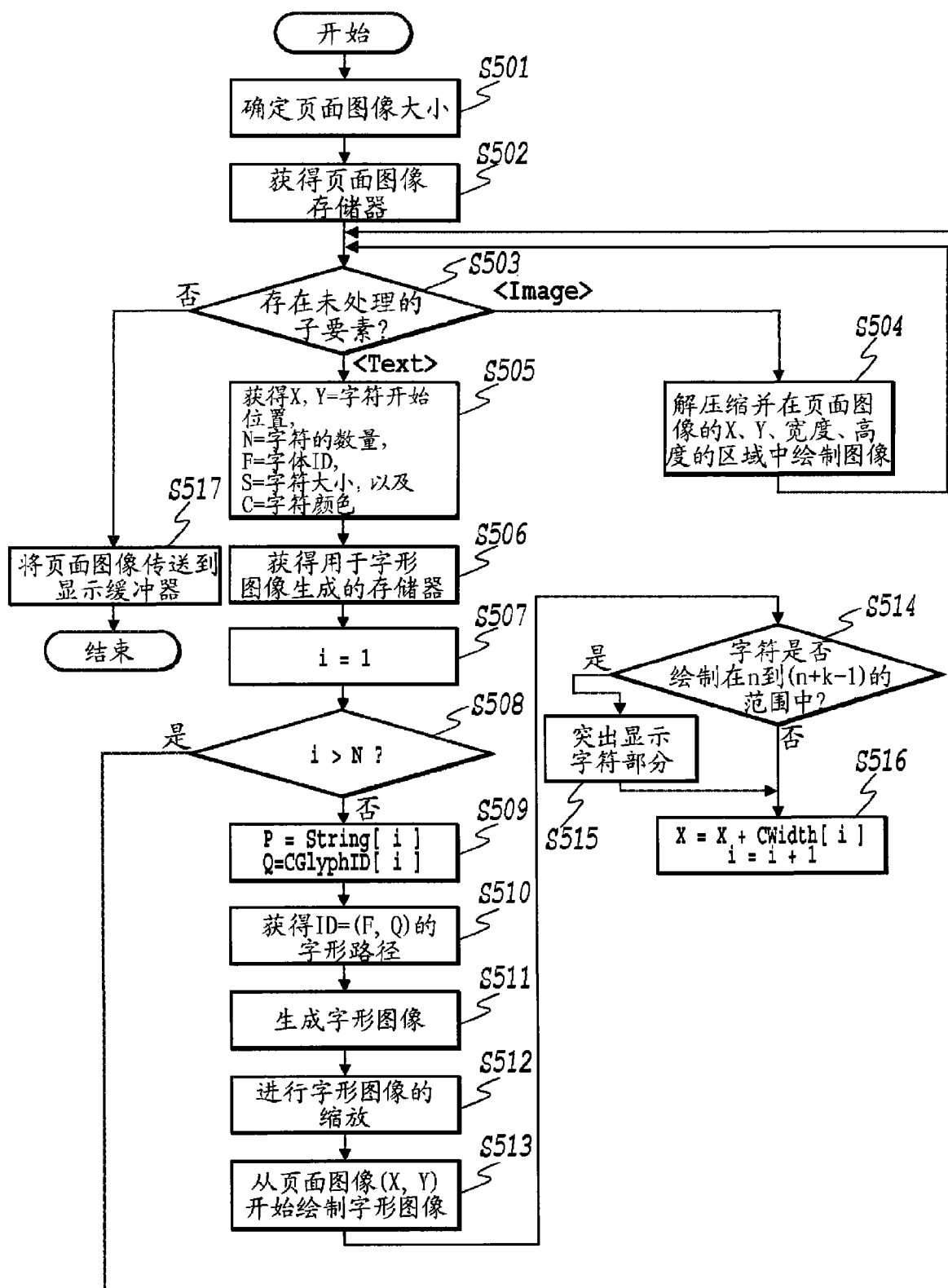


图5

```
601{ <?xml version="1.0"?>
    <Document>
602{   <Page Width="1680" Height="2376" Dpi="200" Number="1">
603{     <Image X="0" Y="0" Width="1680" Height="2376"
        Data="ACAf49sw89q34rs6dbc23cfasx....a3E52zAD"/>
604{     <Text X="236" Y="272" Size="97" Direction="Horizontal"
        Color="0,0,0,255" String="0x2422,0x2424,0x2426"
        CWidth="104,96,59", CGlyphId="0,0,0"/>
605{     <Text X="236" Y="472" Size="92" Direction="Horizontal"
        Color="0,0,0,255" String="0x242b,0x242d,0x242f"
        CWidth="113,101,47", CGlyphId="0,0,0"/>
606{   </Page>
607{   <Page Width="1680" Height="2376" Dpi="200" Number="2">
608{     <Image X="0" Y="0" Width="1680" Height="2376"
        Data="sa8OAN2qx7sa23Basd2x....aAqw8D2dDJ"/>
609{     <Text X="136" Y="872" Size="92" Direction="Horizontal"
        Color="0,0,0,255" String="0x242b,0x242d,0x242f"
        CWidth="113,101,47", CGlyphId="0,0,0"/>
610{   </Page>
611{   <Font Id="Font01">
        <Glyph Id="0" Path="M0,0 V-1024 H1024 V1024 f"/>
        </Font>
612{   <Font Id="Font02">
        <Glyph Id="0" Path="M0,0 V-64 H1024 V64 f"/>
        </Font>
613{ </Document>
```

图 6

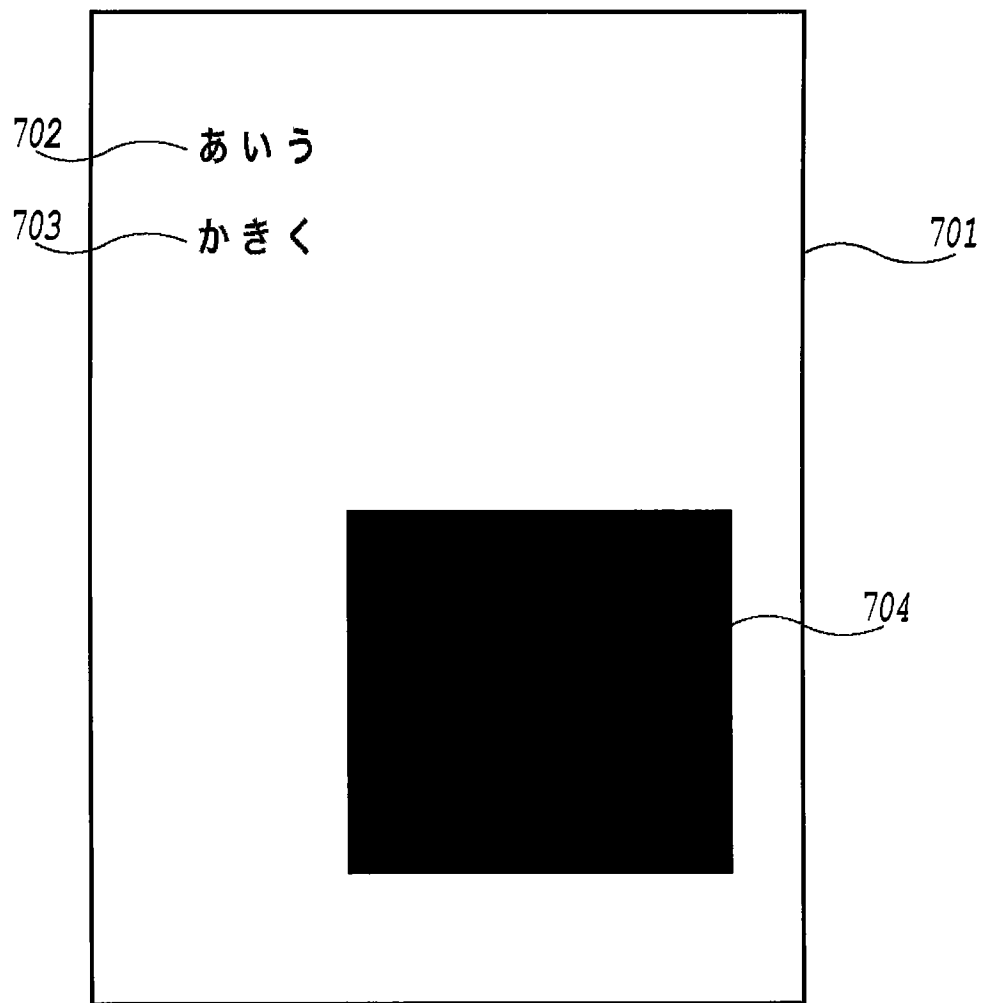


图 7

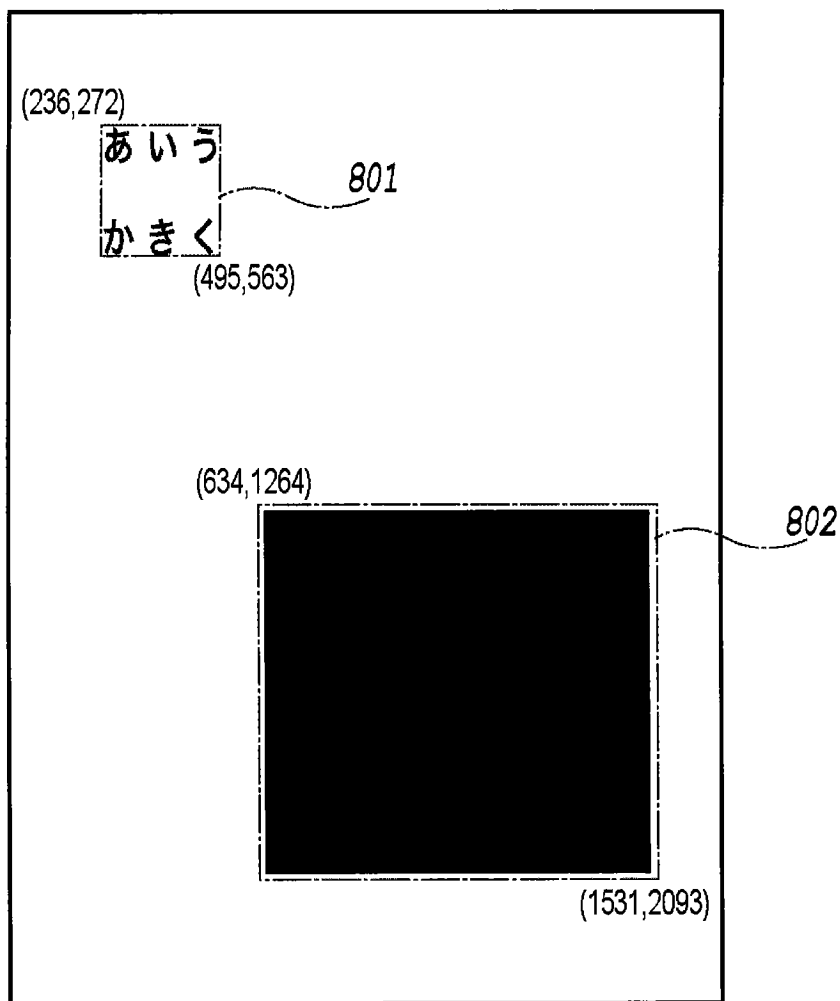


图 8

区域	x	y	宽度	高度	区域类型
1	236	272	260	292	文本
2	634	1264	898	830	照片

图 9

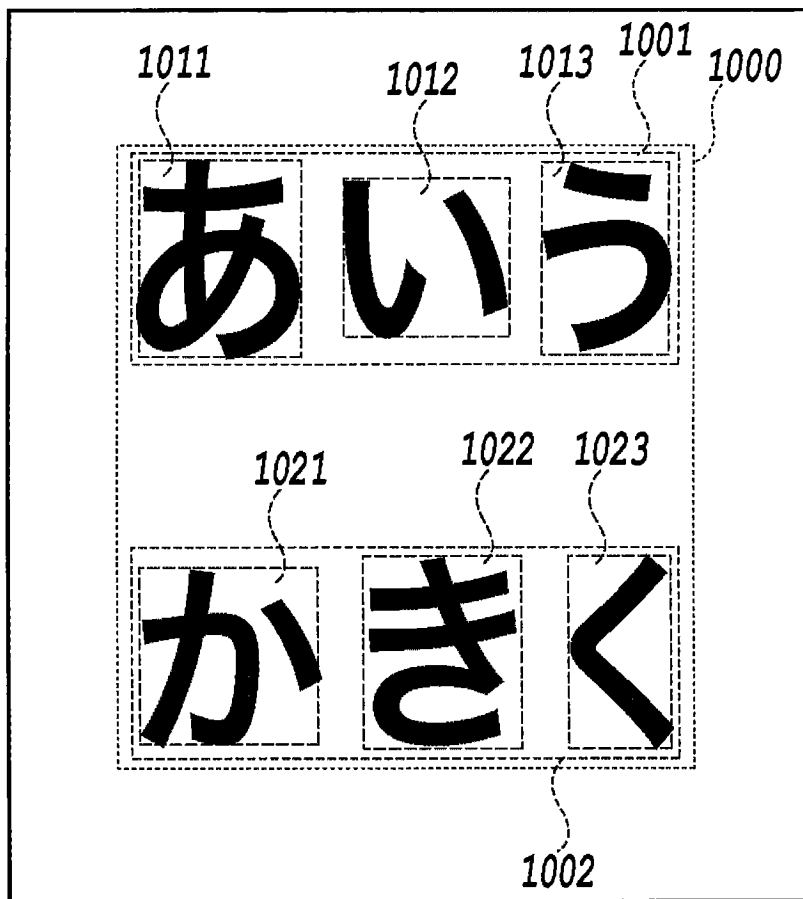


图 10

1101

字符	x	y	宽度	高度	代码
1	236	272	80	97	0x2422
2	339	248	79	75	0x2424
3	434	274	59	95	0x2426

图 11A

1102

字符	x	y	宽度	高度	代码
1	236	474	87	85	0x242b
2	348	472	75	92	0x242d
3	448	472	47	89	0x242f

图 11B

字符	代码	代码
0	0x2422	1093
1	0x2424	1100
2	0x2426	1107
3	0x242b	1250
4	0x242d	1257
5	0x242f	1264
6	0x242b	1601
7	0x242d	1608
8	0x242f	1615

图 12

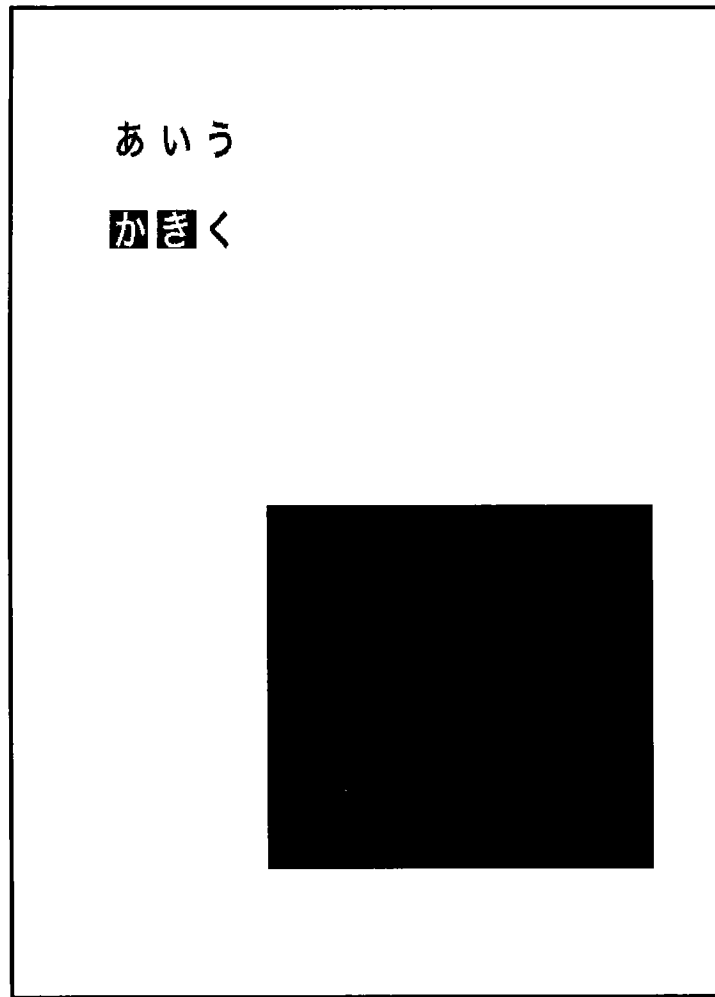


图 13

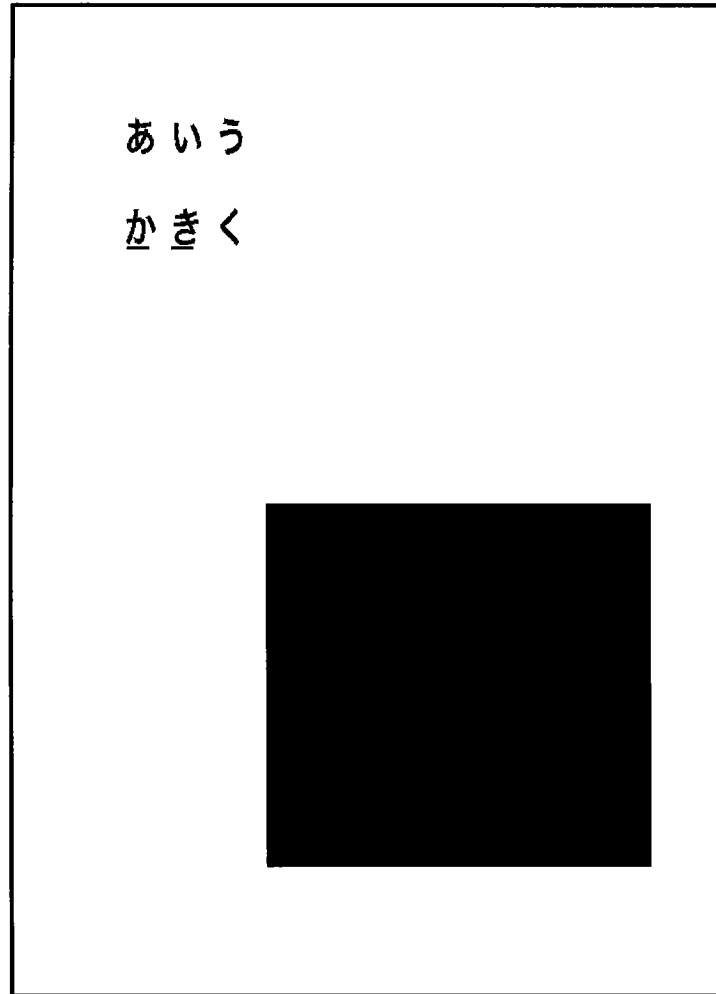


图 14

```

1501 { <?xml version="1.0"?>
1502 { <Document>
1503 { <Page Width="1680" Height="2376" Dpi="200" Number="1">
1504 { <Image X="0" Y="0" Width="1680" Height="2376"
      Data="ACAf49sw89q34rs6dbc23cfasx...a3E52zAD"/>
1505 { <Text X="236" Y="272" Size="97" Direction="Horizontal"
      Color="0,0,0,255" String="0x2422,0x2424,0x2426"
      CWidth="104,96,59", CGlyphId="0,0,0"/>
1506 { <Text X="236" Y="472" Size="92" Direction="Horizontal"
      Color="0,0,0,255" String="0x242b,0x242d,0x242f"
      CWidth="113,101,47", CGlyphId="0,0,0"/>
1507 { </Page>
1508 { <Page Width="1680" Height="2376" Dpi="200" Number="2">
1509 { <Image X="0" Y="0" Width="1680" Height="2376"
      Data="sa8OAN2qx7sa23Basd2x...aAqw8D2dDJ"/>
1510 { <Text X="136" Y="872" Size="92" Direction="Horizontal"
      Color="0,0,0,255" String="0x242b,0x242d,0x242f"
      CWidth="113,101,47", CGlyphId="0,0,0"/>
1511 { </Page>
1512 { <Font Id="Font01" Shape="BOX">
      <Glyph Id="0" Path="M0,0 V-1024 H1024 V1024 f"/>
      </Font>
1513 { <Font Id="Font02" Shape="UNDER_LINE">
      <Glyph Id="0" Path="M0,0 V-64 H1024 V64 f"/>
      </Font>
      </Document>

```

图 15

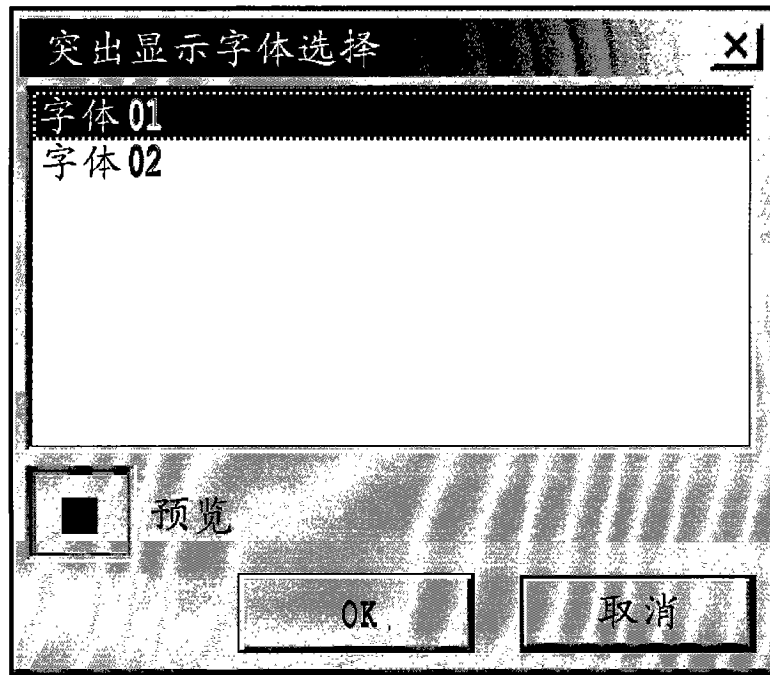


图 16

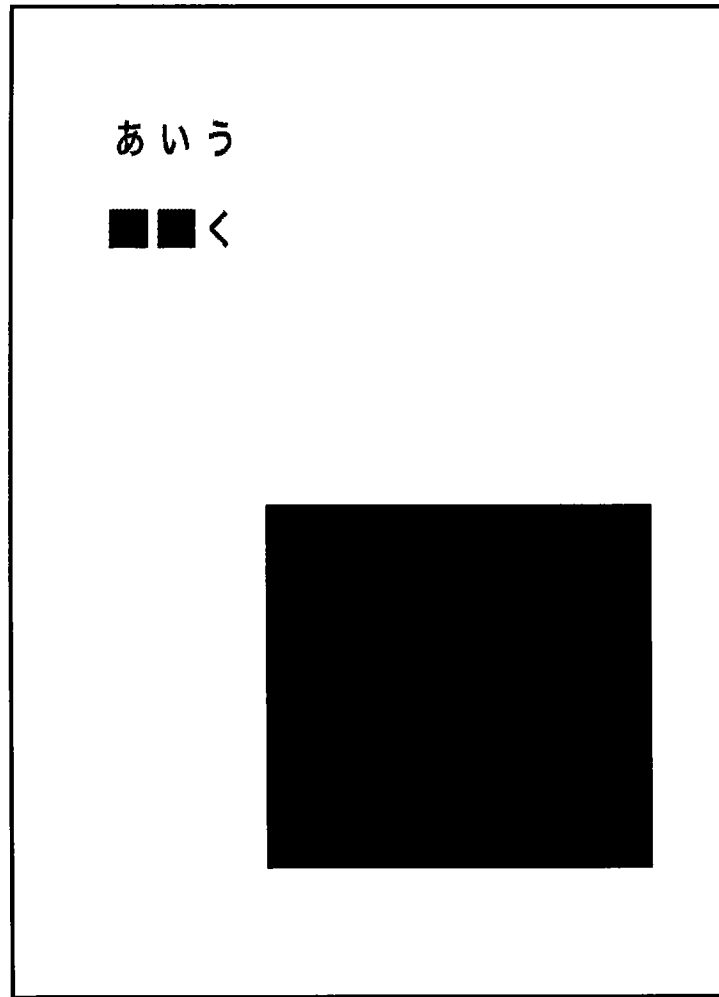


图 17

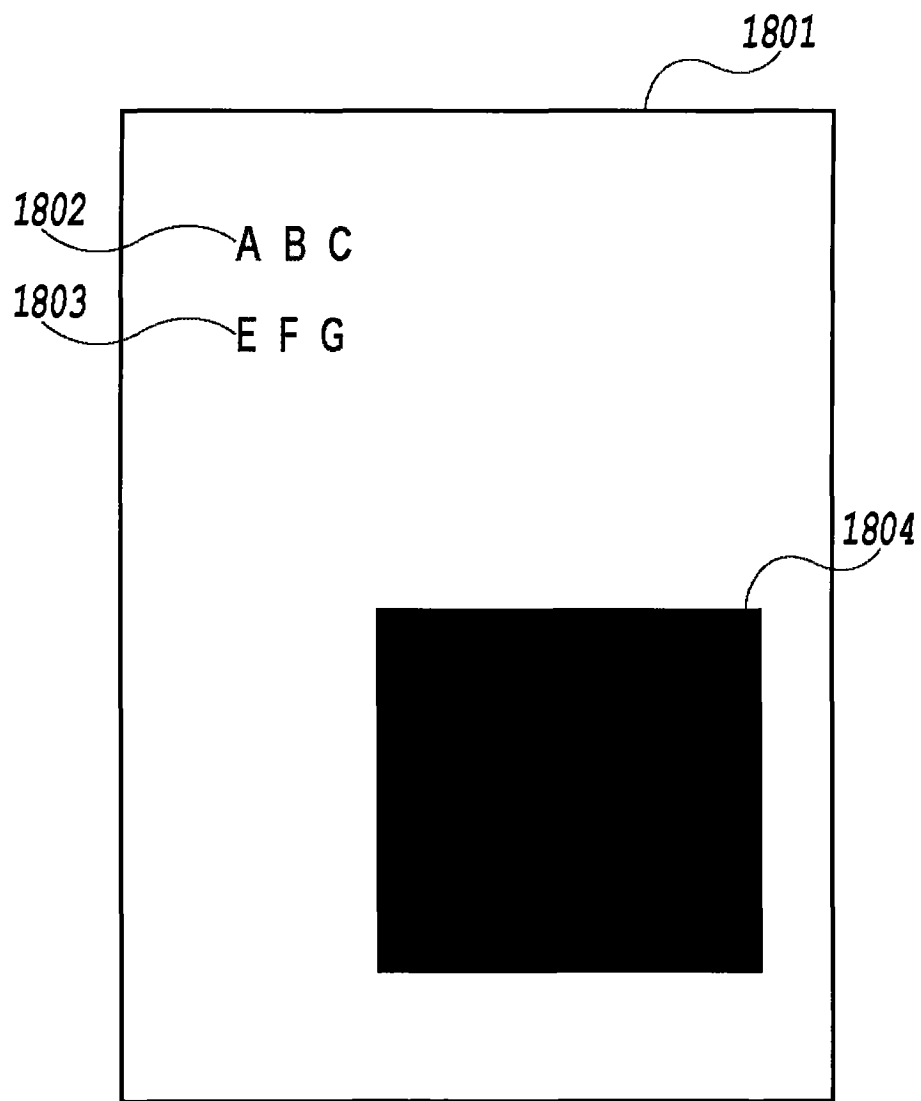


图 18

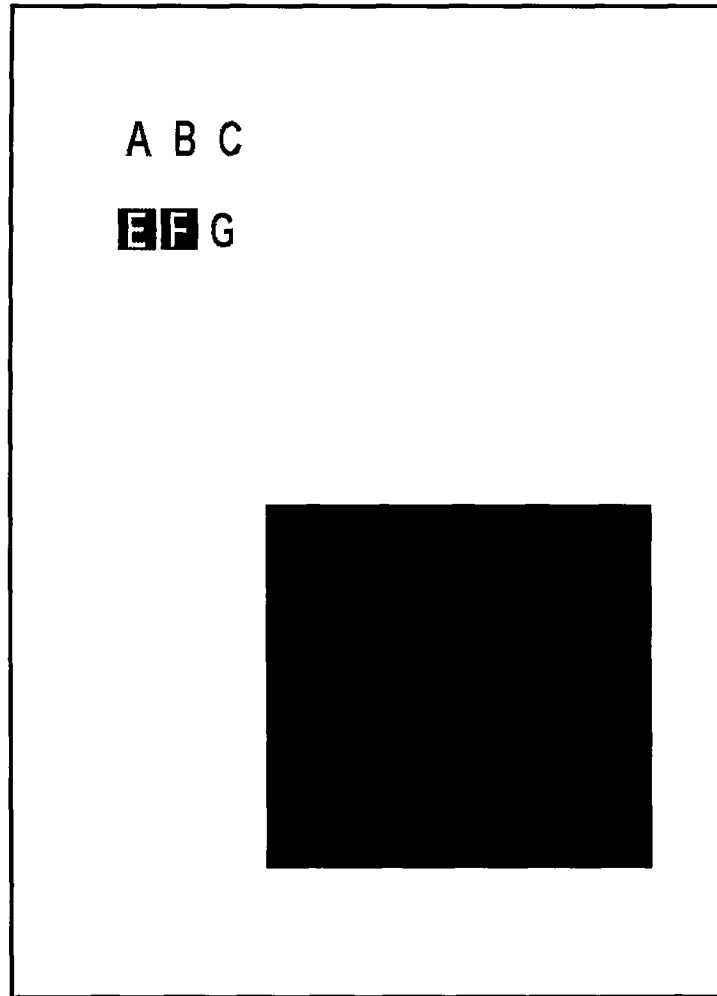


图 19

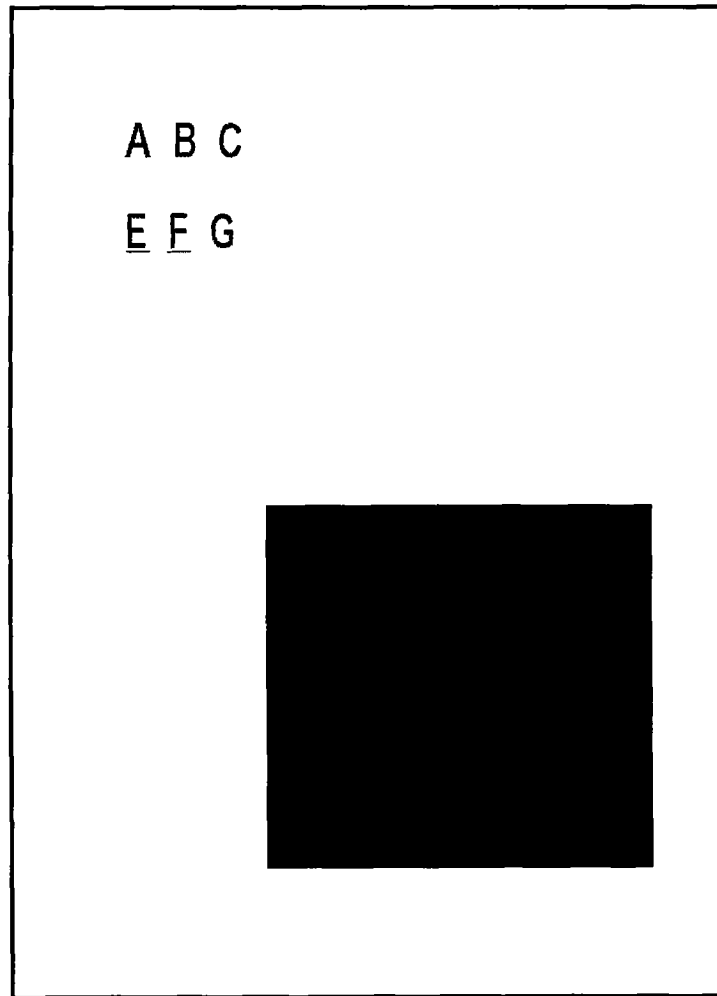


图 20