



[12] 发明专利说明书

专利号 ZL 03108345.5

[45] 授权公告日 2009 年 1 月 14 日

[11] 授权公告号 CN 100452094C

[22] 申请日 2003.3.25 [21] 申请号 03108345.5

[30] 优先权

[32] 2002.4.25 [33] US [31] 10/133,558

[73] 专利权人 微软公司

地址 美国华盛顿州

[72] 发明人 P·Y·希玛德 H·S·玛尔瓦

E·L·伦肖

[56] 参考文献

US4922545A 1990.5.1

JP10-198792 A 1998.7.31

US4606069A 1986.8.12

US5077807A 1991.12.31

审查员 王艳妮

[74] 专利代理机构 上海专利商标事务所有限公司

代理人 张政权

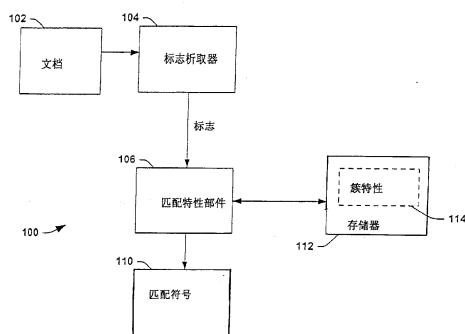
权利要求书 4 页 说明书 25 页 附图 12 页

[54] 发明名称

聚类

[57] 摘要

揭示了用来执行文档图像的聚类的系统和方法。把从文档中析取的标志的一个特性与现存簇的特性比较。如果标志的特性不能与现存簇的特性中的任一个特性相匹配，则把标志加到现存簇作为一个新的簇。一个能被使用的特性是 x 尺寸和 y 尺寸，它们分别是现存簇的宽度和高度。另一个可采用的特性是油墨尺寸，它表示簇中黑像素相对于所有像素的比例。又一个可使用的特性是简化的标志或图像，它是标志和/或簇的像素尺寸简化版本的位图。以上的特性能被使用来识别不匹配，以及减少执行的逐位比较的数目。



1. 一种聚类系统，其特征在于包括：

标志析取器，从文档中析出标志；

匹配部件，计算所述标志的至少一个特性的可接受值的阈值，并根据计算出的可接受值的阈值，将所述标志的所述至少一个特性与标志的现存簇的匹配特性相比较，以识别匹配的现存簇；

二维表，根据盒尺寸来存储现存簇；以及

匹配符号部件，操作上将标志与匹配的现存簇比较，以及识别匹配的簇。

2. 如权利要求 1 所述的聚类系统，其特征在于还包括库，用来存储现存簇。

3. 如权利要求 1 所述的聚类系统，其特征在于还包括现存簇的匹配特性表。

4. 如权利要求 2 所述的聚类系统，其特征在于库包括局部库和全局库，局部库包括从文档的当前页添加的现存簇，全局库包括从文档的先前页添加的现存簇。

5. 如权利要求 1 所述的聚类系统，其特征在于标志析取器操作上定位标志并从文档中析出标志。

6. 如权利要求 1 所述的聚类系统，其特征在于匹配特性包括现存簇的 x 尺寸和 y 尺寸。

7. 如权利要求 1 所述的聚类系统，其特征在于匹配特性包括现存簇的油墨尺寸。

8. 如权利要求 1 所述的聚类系统，其特征在于匹配特性包括现存簇的重新调整大小的图像。

9. 如权利要求 6 所述的聚类系统，其特征在于标志的至少一个特性是 x 尺寸和 y 尺寸。

10. 如权利要求 7 所述的聚类系统，其特征在于标志的至少一个特性是油墨尺寸。

11. 如权利要求 8 所述的聚类系统，其特征在于标志的至少一个特性是重

新调整大小的标志图像。

12. 如权利要求 1 所述的聚类系统，其特征在于匹配部件操作上计算特性的可接受范围。

13. 如权利要求 1 所述的聚类系统，其特征在于匹配部件还操作以添加标志作为不匹配现存簇的标志上的新的簇。

14. 如权利要求 1 所述的聚类系统，其特征在于匹配符号部件还操作以添加标志到最匹配的簇的一组匹配标志中。

15. 如权利要求 13 所述的聚类系统，其特征在于还包括库，库操作上用来存储现存簇和根据现存簇的出现频率来分类现存簇。

16. 一种影印机，其特征在于采用权利要求 1 所述的聚类系统。

17. 一种传真机，其特征在于采用权利要求 1 所述的聚类系统。

18. 一种数字相机，其特征在于采用权利要求 1 所述的聚类系统。

19. 一种图像编码系统，其特征在于采用权利要求 1 所述的聚类系统。

20. 一种聚类方法，其特征在于包括：

在文档中定位标志；

计算用于特性比较的阈值；

根据计算出的阈值，将该标志的第一特性与现存簇的第一特性比较，来识别匹配的和不匹配的簇；

当第一特性匹配时，将该标志的位图与匹配簇的位图相比较，来发现匹配簇的一个匹配的簇；以及

当第一特性不匹配以及位图不匹配时，添加该标志作为新的簇到现存簇。

21. 如权利要求 20 所述的方法，其特征在于正被比较的第一特性包括含有 x 尺寸、y 尺寸、油墨尺寸和重新调整大小的标志图像的组中的至少一个。

22. 如权利要求 20 所述的方法，其特征在于还包括从标志的位图中产生重新调整大小的标志图像。

23. 如权利要求 22 所述的方法，其特征在于通过将标志的位图划分成 9 个区域以及将所述 9 个区域的每一个中的大多数的像素用于一 9 像素的，即 3 乘 3 的，重新调整大小的标志图像的对应像素，，而产生重新调整大小的标志图像。

24. 如权利要求 20 所述的方法，其特征在于还包括计算标志的热点。

25. 如权利要求 20 所述的方法，其特征在于还包括将标志的第二特性和现存簇的第二特性相比较来识别匹配的和不匹配的簇。

26. 如权利要求 25 所述的方法，其特征在于还包括将标志的第三特性和现存簇的第三特性相比较来识别匹配的和不匹配的簇。

27. 如权利要求 20 所述的方法，其特征在于将标志的位图与匹配的簇的位图相比较的方法包括计算标志的重心和簇的重心，计算在标志和匹配簇之间的 xor 距离，以及选择具有与该标志最小 xor 距离的最匹配的簇，该最小 xor 距离处于可接受的范围内。

28. 如权利要求 20 所述的方法，其特征在于经常出现的现存簇被保存在全局库中，而较少出现的簇被保存在局部库中。

29. 一种文档编码系统，其特征在于包括：

表征码离析器，操作以从文档图像中产生二元表征码，该二元表征码包括文本信息；

后台前台分隔器，操作以根据所述二元表征码从文档图像分隔前台图像和后台图像；以及

聚类系统，操作用来根据将表征码的特性与簇的特性相比较的可接受阈值的范围，以计算有效的方法来识别表征码中的簇。

30. 一种聚类系统，其特征在于包括：

用来定位标志的部件；

用来计算可接受的匹配百分比的阈值范围的部件；

使用所述可接受的匹配百分比的阈值范围，将标志的第一特性和现存簇的第一特性相比较来识别匹配的和不匹配的簇的部件；

当第一特性匹配，将标志的位图与匹配簇的位图相比较来发现匹配簇的最匹配的簇的部件；以及

当第一特性不匹配以及位图不匹配，添加标志作为一个新的簇到现存簇的部件。

31. 一种聚类系统，其特征在于包括：

对于一文本的至少一页的每一页：

用来发现至少一个标志的部件；
用于得出可接受的匹配范围的阈值的部件；
根据所述可接受的匹配范围的阈值，将该至少一个标志的第一特性与现存簇的第一特性相比较来识别匹配的和不匹配的簇的部件；
当第一特性匹配，将该至少一个标志的位图与匹配簇的位图相比较来发现匹配簇的最匹配的簇的部件；以及
当第一特性不匹配以及位图不匹配，添加该至少一个标志作为一个新的簇到现存簇上的部件；以及用来更新一全局库的部件。

聚类

技术领域

本发明一般涉及文件图像处理，尤其涉及执行聚类的系统和方法。

背景技术

通过计算机的可利用信息的量随着计算机网络，因特网和数字存储方法的广泛传播而引人注目地增长。随着信息量如此增长，就需要快速传递信息和有效存储信息。数据压缩是一种有效促进信息的传输和存储的技术。

数据压缩减少了描述信息所需的空间量，并且可用于多种信息类型。对包括图像，文字，音频和视频的数字信息的压缩需求不断地增长。典型地，数据压缩用于标准计算机系统；然而，其它技术也可利用数据压缩，例如但不局限于数字和卫星电视，以及蜂窝/数字电话。

随着对操作、传输和处理大量信息的需求的增长，对数据压缩的需求也增长。尽管存储设备容量已显著增长，但是对信息的需求已经超过了容量增长的速度。例如，一个未压缩的图像可能需要 5 兆字节空间，而相同的图像可被压缩，并只需要 2.5 兆字节空间。因此，数据压缩促进了更大量的信息的传递。甚至伴随传输速率的增长，例如宽带，DSL，电缆调制解调器(cable modem) 因特网等等，用未压缩信息会很容易达到传输限制。例如，在 DSL 线上传输一个未压缩的图像可能需要十分钟。然而，当同一图像得到压缩，从而提供十倍的数据吞吐量增益时，就可能在大约一分钟内传输。

通常的，有两种形式的压缩，无损和有损。无损压缩允许在压缩后恢复精确的原始数据，而对于有损压缩，压缩后恢复的数据与原始数据不同。在两种压缩模式中存在折衷，相比无损压缩，有损压缩提供更好的压缩率，因为一定程度的数据完整性的折衷是可以容忍的。例如，当压缩重要文本时，可使用无损压缩，因为不能正确地重建数据可能会显著地影响文本质量和可读性。有损压缩可能用于图像或不重要文本，其中对人类感官来说，一定数量的失真或干扰可以被接受或不能感知。数据压缩尤其可用于文档的数字表示（数字文档）。

代表性地，数字文档包括文本，图像，和/或文本和图像。对当前数字数据除使用较少存储空间之外，没有显著质量退化的紧凑存储器将促使文档的当前硬拷贝的数字化，以便无纸化办公更为可行。对许多行业来说，向无纸化办公奋斗是一个目标，因为无纸化办公提供诸如允许信息方便存取、减少环境成本、减少存储成本等等之类的好处。此外，通过压缩减少数字文档的文件尺寸允许更有效地使用因特网的带宽，从而允许更多信息的更快速传输，以及减少网络阻塞。减少信息所需的存储空间，向有效的无纸化办公推动，以及增加因特网带宽效率仅是与压缩技术相关的许多重要利益中的一些。

数据文档压缩将满足一定目标，以便使使用数字文档更吸引人。首先，压缩应能够以少量时间压缩和解压缩大量信息。其次，压缩应提供数字文档的精确再现。另外，数字文档的数据压缩应该利用文档的预期目标或最终用途。一些数字文档使用于文档归档或提供硬拷贝。其它文档可被修订和/或编辑。许多传统数据压缩方法不能处理在观看时文本和/或图像的回流，并且不能提供有效方法来使压缩技术能够识别字符并将它们回流到文字处理器、个人数字助理（PDA）、蜂窝电话等。因此，如果办公文档硬拷贝被扫描为数字格式，当前压缩技术可能使更新、修正或一般改变数字文档（如果可能的话）变得困难。

数字文档通常包括大量的文本信息。没有任何压缩情况下，每英寸 200 点 (dpi) 的 8.5 乘 11 英寸的一个单文档占据大约 2MB 的存储空间。不过，文本信息具有可承受压缩的特性。压缩文本信息的一个途径是在文本上执行光符识别 (OCR)，并通过使用诸如 ASCII 码的标准字符表中的字符代码的序列来表示文档。然而，使用 OCR 有一些缺陷，在于 OCR 不是完全可靠的，尤其是对于低质量文档。在文档中的干扰、变化的打印字体和不寻常的字符都可能导致 OCR 错误。另外，特殊字体，外国语言和数学公式会引起特殊问题。

压缩数字文档的另外一个途径是使用聚类。聚类包括寻找文档的连通成分（一个连通成分是连接成一组的一给定色彩的像素），并且连通成分被搜索和被分析来定位被称作为簇的类似的连通成分。通常簇能大大增加压缩并且能避免 OCR 的一些可靠性问题。例如，200dpi 的单页面的 8.5 乘 11 英寸文档占据大约 2MB 未压缩的存储空间，但是使用聚类它只需要大约 200k。文件尺寸锐减的原因在于每个连通成分可由一位置以及一个指向属于形状字典的一个形

状的指针来概括。算法的聚类部分决定哪个形状应属于此字典，和哪个形状最接近于每个连通成分。典型地，形状字典是原始文档图形尺寸的一部分，并且甚至能够在页面间被共享。指向形状的指针可能通过页面上一个位置（X 和 Y）和一个形状号码来表征。X 和 Y 位置可能通过使用前一位置来压缩，而使用上下文或语言模型来压缩图像索引。因此，聚类能够大大增加压缩；然而分析连通成分来发现相似连通成分（簇）通常是一个计算密集的处理。一个单页面或多页面文档能简单地拥有上千个连通成分或更多，它们被比较以便能找到相似的连通成分。

发明内容

以下是本发明的概述，以便提供对本发明一些方面的基本理解。此概述不打算标识本发明的关键/重要元素，或描绘本发明的范围。它的唯一目的是以一个简化形式来表现本发明的一些概念，以作为随后给出的更详细描述的前序。

本发明一般涉及文档图像编码解码系统和方法，尤其涉及聚类。聚类包括将相似标志聚合成簇以增加压缩，相似标志主要是连通像素。聚类能够消耗大量存储器和处理器资源，因为每个标志必须和其它标志逐位比较，以便识别簇。本发明减少识别不匹配簇所占用的存储器和处理器资源，并且因此避免对这种不匹配簇的逐位比较。识别不匹配的过程可被称为筛选。虽然一些传统聚类方法利用筛选，但是本发明的筛选测试或特性的选择使它尤其有效。本发明利用现存簇特性，并且把从一个文档中析取的标志的特性与簇特性相比较。此比较不需要费时的和花费大的逐位比较来识别不匹配。如果此标志特性不能和任何簇特性相匹配（如果此标志被认为是远离现存簇），则此标志被添加为一个新簇，而可避免逐位比较。

本发明提供一种计算有效的筛选，来检测不匹配，以及一种计算有效的算法来检测聚类的正确匹配。

根据本发明的一个方面，簇由插入于簇中的第一个元素或标志进行表征。第一个优点是增加一个新标志到簇不需要重新计算簇的特性。此方法的另外一个优点是它避免了“簇漂移”，当一个簇中心移动或伴随着每次增加一个新标志到簇中而更改时，簇漂移发生。避免簇漂移，使在同一个簇的两个元素之间

的最大距离得到保证。此保证允许更多有效的和主动的筛选。此方法的主要缺点在于在感觉上此聚类不是最佳的，对于相同的平均的簇到标志的距离，它会比其他的聚类算法产生更多的簇。主要优点是大大增加速度。在文本压缩情况下，在簇的数量上有 10% 的增长是不必要关心的，因为传递每个簇的信息的字典大小一般是压缩文档尺寸的一部分。

依照本发明的一个方面，例如，簇特性能被存储于一张表中。一个能被使用的簇的某一特定特性是 x 尺寸和 y 尺寸。x 尺寸和 y 尺寸提供现存簇的尺寸信息。可把一个标志的 x 尺寸和 y 尺寸与现存簇的 x 尺寸和 y 尺寸比较，来识别不匹配。另外，本发明的另外一个方面是，将簇组织成 2D 表存储段，通过 x 尺寸和 y 尺寸来索引。当一个新标志被发现时，相同 x 尺寸和 y 尺寸的存储段从 2D 表存储段中被析取，并且将标志与存储段中的簇相比较。也可搜索邻近存储段，进行更精确的匹配（所有其它存储段将被忽略或筛选）。在簇中心上假设“无漂移”在这里是有用的，因为它保证簇保持在它们的存储段中。

簇的另外一个能被利用的特性是油墨尺寸。通常油墨尺寸指在一个标志或簇中黑像素对所有像素的比率。类似的，一个标志的油墨尺寸能够与现存簇的油墨尺寸比较，以识别不匹配。

标志的另外一个特性是它们的“热点”。热点是在标志上的一个位置，它可能是重心，或通过其它方法计算（例如，周围字符的线性方程）。在比较过程中，标志和簇的热点在比较之前被对齐，经常导致转化。另外一个簇特性是一个简化的标志或图像，它是像素尺寸简化版本的标志和/或簇的一个位图。本发明的一个方面中，简化的标志以热点为中心，并且这个标志被重定比例，使这个简化的标志具有一个固定尺寸。同样，对该簇的不漂移假设在担保一个簇的简化版是在簇中的所有标志的一个良好表示中是有用的。简化的标志可与现存簇的简化标志或简化图像比较，来识别不匹配。

依照本发明的另外一个方面，当一个标志成功通过所有的特征测试时，随后它以更直接的方式与簇比较。计算标志和簇之间的一个距离，并且与第一阈值比较。这个距离可以是一个 XOR 距离，一个更复杂的距离，或诸距离的组合，各距离被使用为对更复杂和有意义的距离的一个筛选。如果距离低于它们所对应的阈值（可实验地确定该阈值），则这个标志被加到当前簇中。

在本发明的一个方面中，簇以大小顺序保存，以便首先将最可能匹配的簇与标志相比较，从而当一次匹配被较早发现时减少比较次数。

本发明的另外一个方面是一种聚类系统。此聚类系统包括一标志析取器、库、表、匹配部件和匹配符号部件。标志析取器从一个文档中析取一个标志。库可被运行用来存储现存簇和关于那些现存簇的信息。表存储现存簇的匹配特性。匹配部件被运行用来把至少一个标志特性与匹配特性表中的匹配特性比较，来识别匹配的现存簇。匹配符号部件被运行用来把标志与匹配的现存簇比较，以及识别最匹配的簇，或当速度是必需的时候，用来识别第一簇，使得标志和该簇之间的距离小于第一阈值。

本发明的另外一个方面涉及如何计算每个簇的最终位图。一旦聚类完成，通过根据热点对齐簇中的所有标志，和对每个位置平均簇中所有标志的像素来估计每个簇的位图。对每个位置，如果平均值大于 0.5，则在那个位置上的簇的位图像素被设置为 1，而如果平均值小于 0.5，在那个位置上的簇的位图像素被设置为 0。如果平均值正好是 0.5，8 个邻近平均数也被平均，并且如果新的平均值大于 0.5，则像素值被设置为 1；反之为 0。结果产生的位图输入形状字典，并且当图像待被编码时，在那个簇中的所有标志都指向它。

本发明的另外一个方面提供一种聚类方法。提供一个文档。这个文档是二进制的并且典型的包括文本信息。在文档中发现一标志。该标志的第一特性和现存簇的第一特性相比较，以识别匹配和不匹配的簇。标志的位图和匹配簇的位图相比较，以发现匹配簇中最匹配的一个簇，或者提高聚类速度，来发现第一簇，使得标志和该簇之间的距离小于第一阈值。如果没有匹配簇，则这个标志作为一个新簇被添加。

为了前述的和有关的结尾的完成，本发明的某些说明性方面连同下列描述和附图在这里被描述。这些方面指示了各种各样的可能实践本发明的方法，它们都由本发明所覆盖。当连同附图一起考虑时，本发明的其它优点和新颖特征可从本发明的下面的详细描述中变得显而易见。

附图说明

图 1 是根据本发明的一个方面的一个聚类系统的框图。

图 2 说明根据本发明的一个方面的一个示例性标志。

图 3 说明 4 连通特性。

图 4 说明 8 连通特性。

图 5 说明根据本发明的一个方面的示例性组合的标志。

图 6 是根据本发明的一个方面的聚类系统的框图。

图 7 是根据本发明的一个方面的二维表的框图。

图 8 是根据本发明的一个方面的二维表的条目的方框图。

图 9 说明根据本发明的一个方面的示例性簇。

图 10 说明根据本发明的一个方面的示例性错误映射。

图 11 是根据本发明的一个方面的执行聚类的一个方法的流程图。

图 12 是根据本发明的一个方面的利用聚类的一种图像编码系统的框图。

图 13 说明可在其中运行本发明的一个示例性操作环境。

图 14 是根据本发明的示例性通信环境的示意框图。

具体实施方式

现在参照附图描述本发明，其中，相同的参考标号被用来全程地指示相似元素。在下列表述中，为了解释目的，阐明很多具体细节，以便提供对本发明的彻底理解。然而，本发明可能在没有这些具体细节的情况下被实行，这是是明显的。在其它的一些实例中，知名的结构和设备以框图形式被显示以便描述本发明。

如同在本申请中所用的，术语“部件”意图指一个计算机相关的实体，或者是硬件、硬件和软件的组合、或执行中的软件。例如，一个部件可以是但不局限于运行在处理器上的一个过程、处理器、一个对象、一个可执行的线程、程序以及计算机。通过举例说明的方法，运行在服务器上的一个应用程序和这个服务器都能够是一个部件。一个或多个部件能够驻留于一个过程和/或执行线程中，并且一个部件可能定位于一个计算机上和/或分布在两个或多个计算机之间。

此外，“文档图像”意图指包含一个或多个颜色的文档（例如，二元的（黑/白），灰度级和/或彩色文档）的数字表示。另外，一个文档图像能具有图像，

文本，和/或有图像的，有文本和图像可能叠印的文本。文档图像能包含二元，RGB，YUV 和/或其它文档的表示。一个 RGB 文档图像被表现为红色，绿色，和蓝色成分。一个 YUV 文档图像使用用 Y 表示的亮度成分和用 U 和 V 表示的色度成分来表示。该 YUV 表示通常更适合于压缩，因为人类的眼睛对 U 和 V 失真是较不敏感的，并且因此 U 和 V 能够通过因子 2 被二次采样，并且因为 Y 在 R，G 和 B 之间捕捉相关性。为了文本聚类的目的，Y 表示特别引起注意，因为当亮度改变时，文本通常比较容易去读取。从色度的一个改变中得到的文本，例如在一个给定亮度上从红色变到绿色，一般较难读取。一个彩色文档，因此能被转换为一个 YUV 文档，它能随后被压缩或二进制化，而充分保持文本信息。一个文档图像包含一般被称为“像素”的图片元素。一个文档图像可基于任何形状或尺寸的单个或多个页面的文档。

图 1 说明了根据本发明的一个方面的一个聚类系统 100。这个聚类系统 100 从一个文档中析取标志，通过避免已发现的标志对现存簇的大量逐位比较，以计算有效的方式，从标志中创建簇。聚类系统 100 执行许多比较来识别标志和现存簇中的匹配和不匹配，从而避免许多计算开销大的逐位比较。

逐位标志析取器 104 从文档 102 中析取一个标志（例如，它典型地包括文本信息）。通常地，文档 102 是一个二元的文档图像，然而可以理解的是本发明的替代方面能包括具有任意适合数量的等级或色彩的文档。文档 102 有一页或更多页，并且可以是从另一个文档图像中析取的层或表征码。标志通常是一个给定颜色的多个连通像素，并且也能够被称为连通成分。常常，一个标志是一个例如“e”的字母数字字符。然而，不同于 OCR，一个标志不被识别为一个具体字符。

在文档中的文本的普通表示表现为白底黑字。为了简单起见，系统 100 通过文档 102 的文本表示为黑色来描述。然而，可以理解的是文本可以表示为任何适当颜色或底纹或任何其它颜色或底纹。

标志析取器 104 扫描文档 102（例如，从左到右，从上到下）来识别或定位一个标志的“种子”。种子是标志析取器 104 识别/定位的第一个非白色像素。标志析取器 104 从一个开始位置或像素开始扫描。该开始位置一般在文档的左上角，并且在每个标志被析出后被更新。因此，例如，最先定位最高最左边标

志。一旦种子被找到，应用算法去定位连通像素。标志析取器 104 随后析出标志，包括连通像素，并且从文档 102 中移出它。

匹配特性部件 106 接收来自标志析取器的标志。为了识别不匹配，匹配特性部件 106 把一个或多个标志特性与相对应的现存簇 114 的一个或多个特性相比较。这个比较也可称为是筛选。一个或多个特性可包括，例如，标志的位图图像，标志的重心，x 尺寸，y 尺寸，油墨尺寸，位置信息和重新调整大小的标志图像。存储部件 112 存储现存簇 114 的一个或多个特性，并提供诸特性到匹配特性部件 106。存储部件 112 可以是存储设备，例如，存储器，硬盘驱动器，闪存存储器等等。如果需要，可由匹配特性部件 106 计算标志的一个或多个特性。类似地，如果需要的话，现存簇 114 的一个或多个特性也能由匹配特性部件 106 计算。

匹配特性部件 106 使用一个阈值为一个或多个特性计算可接受的值的范围。这个阈值可以是，例如，80%，意思是现存簇的特性能够落在标志的对应特性的 80% 内。可调整阈值来提高系统的性能。然而，阈值越大，标志将被添加到一个现存簇中的可能性越小，并且将获得越少的压缩。可对诸特性中的每一个使用多个阈值。

一个或多个特性能被顺序地或并行地比较。例如，为了顺序比较，标志的 x 尺寸特性可与现存簇的 x 特性相比较，随后，标志的油墨尺寸可与现存簇的油墨尺寸相比较。顺序比较受益于这样一个事实，即由于从许多簇中寻找一个适当的簇而造成预期大多数比较都会失败。因此最先进行计算最低廉的特征比较，一旦一次比较失败，对这个簇的剩余比较就不需要被执行了。该簇就是一个不匹配簇，考虑下一个簇。如另外一个例子，为了并行比较，标志的 x 尺寸和油墨尺寸可与现存簇的 x 尺寸和油墨尺寸相比较。并行比较从同时执行多个比较中受益。并行比较能利用较大量的比较，但能通过执行并行计算来提高速度。

如果一个不匹配被识别出，意为现存簇的一个或多个特性中的一个不能落在可接受的范围内，或者标志和簇之间的距离也未能落在可接受的范围内，则标志作为一个新的簇被添加到现存簇中。此外，系统进行的标志处理完成。文档的另一个标志可由系统 100 来处理。

如果识别一个匹配，意为至少有一个匹配的现存簇，匹配符号部件 110 把标志的位图与至少一个匹配簇的位图比较。位图是逐位比较，来识别具有相似位图的至少一个匹配的现存簇的簇。这个比较也称为匹配。可计算不止一个距离的比较。例如，首先计算 XOR 距离，意味着对每个不匹配位，把 1 添加到该距离，如果第一距离落在一定范围内，则继之以计算更复杂和计算开销更大的距离。根据本发明可使用其它适合的匹配过程，如加权异或（WXOR），它根据被设置的许多邻近像素来加权像素；加权的与非(WAN)，类似于 WXOR，当计算加权时区分对待从白到黑的错误和从黑到白的错误；式样匹配和替换（PMS），如果发现在错误映射中的任何位置具有 4 个或更多的被设置的邻居，则拒绝匹配；组合的尺寸无关策略（CSIS），它通过利用试探方法来增大 PMS 过程来探测稀少的笔划或缝隙；和基于压缩的模板匹配（CTM）。匹配过程或比较能在第一个可接受的匹配的现存簇作为匹配簇时被终止。选择一个阈值，它允许在标志和至少一个匹配的现存簇的位图上有较小偏差。关于图 6，比较位图的方法在下面被进一步详细地解释。作为替代，匹配过程能够贯穿所有的至少一个匹配的现存簇来继续，以识别最佳匹配或最匹配的簇为匹配簇。在发现或识别的匹配簇上的标志，被添加到匹配簇的类似标志组中。能通过平均利用标志的位图来更新匹配簇的位图。

如果识别关于位图的一个不匹配，意为至少一个匹配簇无位图可被接受，则标志被添加到现存簇中作为一个新的簇。另外，系统 100 对标志的处理完成。系统 100 能随着文档中的所有剩余标志而继续。

图 2 说明根据发明的一个方面的示例性标志 201。利用连通特性来发现连通像素。这个示例性标志 201 代表一个文本字符（例如，字符“z”）。典型地，使用 8 连通特性来定位连通像素。然而，也可使用 4 连通特性。4 连通特性仅识别在四个主要包围方向中的连通像素。图 3 通过示出在四个主要包围方向 310 上由 4 个像素包围的一个像素，来说明 4 连通特性。8 连通特性在每个可搜索的邻居方向上识别连通像素，即八个直接的邻居，分别在 8 个主要方向上的每一个方向上。图 4 通过显示在八个主要方向 411 上由 8 个像素包围的一个像素，来说明 8 连通特性。例如，图 2 中的像素 203 根据 4 连通特性和 8 连通特性被识别。然而，4 连通特性将不能把像素 204 识别为连通像素。

再次涉及图 1，标志析取器 104 能通过使用下列的算法来找到标志。首先，把二元图像转换成游程长度编码（RLE）表示。假设这里有 n 个游程，创建 n 个组，在每个组中放置一个游程。随后对每行以及每个游程，使用 4 连通或 8 连通，在前一行中找到与当前游程接触的所有游程。执行包含接触的游程的所有这些组的一个联合。在页的末端，剩余的组就是标志。这个算法可被称为“联合-发现”算法。复杂性为 $O(n \log^*(n))$ ，n 为游程个数， \log^* 是 Ackerman 函数的反运算。因此，复杂性近似为 $O(n)$ ，在这里 n 是游程个数。另外，可对图像和它的底片都计算标志，以便使反白显示文本形成簇。

如上陈述，标志通常由连通像素构成。然而，标志析取器 104 能被适应于识别不严格地为连通像素的标志。例如，图 5 连同上部点 502 一起显示字母“i”的下部 501。下部 501 和点 502 是明显的连通像素。然而，标志析取器 104 能通过从原始标志中扫描附加区域来将下部 501 和点 502 识别为一单个的或组合的标志。

为了结合连通成分或标志，可执行行分析。例如连通成分可基于它们在页中的位置而被组合成单词，单词同样基于它们在页中位置和它们的角而被组合成行。一旦识别了一行，相对于行的方程，就可能合并垂直位于彼此顶部上的所有标志，使得重音和点标志附加到它们所属的单词的标志上。结果产生的标志理所当然是不连通的，但能如同以前一样在聚类算法中被使用。

在标志已被识别之后，标志析取器 104 能在标志上执行附加的计算。标志析取器 104 能计算标志的特性，例如，标志的重心，x 尺寸，y 尺寸，油墨尺寸，区域，位置信息，热点，以及重新调整大小的标志图像。x 尺寸按标志的最大宽度计算。y 尺寸按标志的最大高度计算。如果标志是两个标志的组合（例如，“i”和它的点），则 y 尺寸是所产生的组合的高度。油墨尺寸是在标志中的黑像素数，而区域是标志的边界框中的像素总数（x 尺寸乘以 y 尺寸）。另外一个有时有用的用来探测不匹配的特征是油墨百分比。下面的方程式能用来计算标志或簇的油墨百分比：

$$\text{油墨百分比} = \frac{\text{油墨尺寸}}{\text{区域}} \quad \text{方程 1}$$

热点是在标志上的一个位置，它可以是重心，或通过其它手段计算。作为一个例子，对于文本，就可能进行行分析和找到标志所位于的行的方程式（这

是用于寻找行来连接例如“i”上的一点之类的标志的相同算法）。标志的热点具有对于x坐标水平重心，以及对于y坐标，用行方程式，穿过相同的x坐标的垂直线的交叉点。在比较过程中，标志和簇的热点在比较开始之前对齐，常导致在比较之前的标志的转化。

位置信息可包括诸如标志的坐标，偏移量，基准点和/或空间信息之类的信息。

重新调整大小的标志图像是标志经重新调整大小的版本。这个重新调整大小的标志图像被计算为一个较小的版本，例如标志的一个3乘3像素图像。为了帮助符号比较，计算重新调整大小的标志图像。重新调整大小的符号一般在像素尺寸上比标志小得多。许多算法都可用于将标志缩小成重新调整大小的标志图像。其中一个算法是将标志划分为9个区域（3乘3），以热点为中心，并且按比例缩放以覆盖标志的区域，并且当它们和这些区域中的每块区域相交时，计算标志的平均值。

另外一个算法是从位图中移去或删除若干像素。然而，可使用其它算法来生成重新调整大小的图像，但仍依照本发明。可以理解的是上面的计算能被延迟，直到或如果系统需要为止，以便增加计算效率。

另外，根据本发明，可以理解的是现存簇和它们的特性在一局部库和一全局库中保持。可以使用局部库为当前页存储现存簇，以及使用全局库为文档102全局地存储现存簇。在另外一个方面中，频繁出现的现存簇（例如，具有相对较大量相关联的标志的簇）保持在全局库中，而很少出现的现存簇保持在局部库中。

图6是根据发明的一个方面的聚类系统600的框图。聚类系统600从一个文档中析取标志，并且通过避免对现存簇的大量逐位比较已发现的标志，以计算有效的方式从标志中创建簇。聚类系统600执行若干识别不匹配的比较或筛选操作，从而避免许多计算开销大的逐位比较。

通常，通过插入到簇中的第一个元素或标志来表征簇。它的一个优点是增加一个新的标志到簇不需要重新计算簇的特性。这个方法的另外一个优点是它避免了“簇漂移”，当簇中心移动或随着每次添加新标志到簇而被更新时发生簇漂移。避免簇漂移保证了在相同簇的两个元素之间的最大距离。这个保证允

许更有效和更主动的筛选。此方法的一个缺点在于在感觉上该聚类不是最佳的，比其它聚类算法，对于相同的平均簇到标志距离，它会产生更多的簇。然而，处理速度上的进步一般比上述缺点重要。作为一个例子，在文本压缩的情况下，在簇的数量上有 10% 的增长是不必要关心的，因为传递每个簇的信息的字典大小一般是压缩文档尺寸的一部分。

二维表 612 存储和保持现存簇。现存簇是先前发现的或识别的簇。如上所述，一个簇是至少一个相似标志的组。二维表 612 根据现存簇的盒尺寸或边界框来容纳实体。表 612 的各个实体，称为存储段，包含对应于某一特定盒尺寸的簇的列表。该盒尺寸称为标志或簇的 x 尺寸和 y 尺寸。可以理解的是实体或存储段可能是空的。另外，可对表建立最大盒尺寸，因为簇一般落在一定的盒尺寸以下。在最大盒尺寸之上的簇可存储在一个单独的表中（例如，公司的标识或印章）。典型的，具有相对较大盒尺寸的标志不太可能类似于其它簇或标志。

图 7 是说明可用于图 6 的系统 600 中的二维表的框图。如所描述的那样，实体或存储段按 x 尺寸和 y 尺寸排列。一个存储段 710 具有的 x 尺寸为 1，y 尺寸为 N。在存储段 710 中的簇说明于图 8 中。

标志析取器 604 运行用来从文档 602 中获得标志。文档 602 是一个数字文档，一般包括文字信息。标志析取器 604 一个一个的移除和处理标志。标志析取器 604 采用一合适的算法从文档 602 中析取和移除标志。对于各个标志，标志析取器 604 计算和/或获得标志的特征或特性，包括但不限于标志的位图图像，标志的重心，x 尺寸，y 尺寸，油墨尺寸，位置信息和重新调整大小的标志图像。

匹配尺寸部件 606 接收来自标志析取器的标志。该标志包括标志的位图图像，标志的重心，x 尺寸，y 尺寸，油墨尺寸，位置信息和重新调整大小的标志图像。该匹配尺寸部件 606 利用一个阈值来计算可接受的 x 尺寸和 y 尺寸值的范围。该阈值可以是，例如 10%，意味着可接受的 x 尺寸和 y 尺寸的值应当在标志的 x 尺寸和 y 尺寸的 10% 内。可以调整阈值来提高系统 600 的性能。然而，阈值越大，现存簇将被发现匹配的可能性越小，并且一般获得的压缩越少。

匹配尺寸部件 606 参照尺寸表 612 获得一个或多个在上述的可接收范围内

的存储段。该一个或多个存储段包含现存簇的一部分，可称为尺寸匹配的簇。如果没有尺寸匹配的簇，则标志作为一个新的簇被添加到二维表 612 中。另外，匹配尺寸部件 606 可扩大尺寸值的范围或包括邻近的存储段，以便获得更多尺寸匹配的簇，尤其是如果最初没有识别出任何尺寸匹配的簇时。上述的在簇中心的“无漂移”假设是有用的，因为它允许现存簇保留在它们的存储段中。

如果标志作为一个新的簇被添加，标志的信息，包括但不限于，唯一的标识符，标志的位图图像，标志的重心，x 尺寸，y 尺寸，油墨尺寸，位置信息和重新调整大小的标志图像，被加到在二维表 612 中的一个存储段中。标志的图像（现在是一新簇的图像）能单独地存储在形状字典中，该字典中库条目仅容纳指向形状字典或形状号的指针。形状字典可以是局部的和/或全局的。图 9 说明对于一个单独的簇的一个示例性的条目 901。可以理解的是条目 901 并不全面，额外的信息可包括在现存簇的一个实体中。

再次谈及图 6，如果有至少一个尺寸匹配的簇，对标志的处理则由匹配油墨部件 608 继续。标志的油墨尺寸分别和尺寸匹配的簇的下一个簇的油墨尺寸相比较。如果尺寸匹配的簇中没有下一个簇，则如上描述的那样把标志作为一个新的簇添加到现存簇中。油墨尺寸信息是标志中黑像素数对全部像素数的百分比或比值。如果标志的油墨尺寸匹配下一个簇的油墨尺寸，则对标志的处理由匹配符号部件继续。如果标志的油墨尺寸不匹配该下一个簇的油墨尺寸，则匹配油墨部件用尺寸匹配簇的接着的下一个簇继续对标志的处理。

匹配符号部件 610 继续对标志的处理。首先，匹配符号部件 610 将标志的重新调整大小的图像和下一个簇的重新调整大小的图像相比较。重新调整大小的图像是标志或簇的位图的一个缩小尺寸的版本（例如，3 乘 3 的像素表示）。缩小尺寸的版本能够用灰度等级像素表示，这比二元好。这个比较能相对较快地执行，因为仅有 9 个像素被比较。可建立一阈值来允许缩小尺寸的表示的有限变化。如果重新调整大小的图像不匹配（例如，不匹配），标志的处理返回到匹配油墨部件 608，在那里处理接着的下一个簇。

在缩小尺寸之前，使用“基准点”对齐图像是有利的。参考的簇和标志的基准点在计算向下采样的图像之前被计算。基准点，也被称为热点，能计算为重心或通过其它方法计算（例如，周围的标志的行方程式）。按每个位图上的

黑像素的平均位置计算标志或簇的重心。其它的基准点和多重基准点能用来对齐图像。在图像被向下采样之前，图像得到对齐，使得热点对应于向下采样的图像的中心。

如果重新调整大小的图像确实匹配，则执行对标志的位图和下一个簇的位图的逐位比较。为了比较标志和参考的簇，匹配符号部件 610 计算标志和下一个簇之间的距离。如果该距离在一阈值之内或小于一阈值，那么这个参考的簇是一个匹配簇。否则，对标志的处理返回到匹配油墨部件 608，并且用接着的下一个簇继续。

能由匹配符号部件利用的一个合适的距离是“xor”距离。通过产生一个错误映射来计算“xor”距离，它是标志和参考的簇之间的按位异或。“xor”距离是错误映射中的像素数量。在计算 xor 距离之前，图像通过叠加它们各自的热点来得到对齐。

图 10 说明一示例性的错误映射。标志的位图示于 1001。现存簇的位图示于 1002。标志 1001 和现存簇 1002 上的按位异或操作示于错误映射 1003。

根据本发明，可使用其它距离计算，如加权异或（WXOR），它根据设置的若干邻近像素来加权像素；加权与非（WAN），它类似于 WXOR，当计算加权时区分对待从白到黑的错误和从黑到白的错误；式样匹配和替换（PMS），如果在错误映射中发现任何位置具有 4 个或更多的被设置的邻居时，则拒绝一个匹配；组合的尺寸无关策略（CSIS），它通过利用探索方法增大 PMS 过程来探测稀少的笔划或缝隙；以及基于压缩的模板匹配（CTM）。

另外，对于对下一个簇的各个标志比较，可计算和采用多重距离。例如，可首先计算上述的 XOR 距离，意为对于每个不匹配位把 1 添加到该距离，如果第一距离落在一定范围内，则继之以一更复杂和计算开销更大的距离。

一旦发现一个匹配，标志由匹配符号部件 610 添加到下一个簇上。该标志被分配一个唯一的标志号，它连同位置信息一起存储在匹配簇的标志组中。该位置信息包括但不限于，x 和 y 位置坐标，与前一坐标的距离，偏移量信息等。现存簇可包括标志计数，它指示与簇相关联的标志的数量。如果匹配簇位于局部库 616 中，则在局部库 616 中的簇可由属于每个簇的标志数，即标志计数再分类。

匹配尺寸部件 606 和匹配油墨部件 608 一起称为匹配特性部件 605。匹配特性部件 605 筛选现存簇，以识别不匹配。在本发明的其它方面中，匹配特性部件 605 可包括任何合适的标志和现存簇的特性或特征，提供用于有用的比较，以识别不匹配。此外，替换方面中，匹配特性部件 605 可包括尺寸和油墨尺寸特性中的一个或两者，或两者都不包括。

在文档中的所有或充分的标志被析取之后，可删除二维表，并且现存簇可被转换到一个全局库 618。另外，二维表 612 中的簇在文档 602 的各个页面的处理之后能被转化为全局库 618。全局库 618 包括来自文档的现存簇。先前与现存簇一起存储的一些信息可被删除，例如油墨尺寸，因为对于编码和解码而言并不需要。通常，簇的位置信息和位图图像使用适合该信息类型的压缩方案来编码。一旦经过编码，位置信息和位图图像就组合成一个文件或位流。另外，为了适当解码这个文件，库中的簇的数量应当包括在该文件或位流中。从而，该位流应当最终包括库、标志顺序以及位置信息或偏移量。可以理解的是库也可称为字典，簇可称为符号，以及标志顺序可称为符号顺序。

同样，在所有的标志已被移去之后，文档可被称为残余图像。对于有损压缩，该残余图像可被删除。然而，对于无损压缩，以及甚至某些有损实现，残余图像被压缩并以位流传送。残余图像可能看起来是简单的噪音干扰，并且因此对于绝大多数压缩方案来说压缩质量可能很差。然而，利用现存簇的压缩方案极大地提高了残余图像的压缩。可进行文档布局的附加分析，以进一步增加文档的压缩。例如，可用水平线识别簇和标志，来减少所需的位置信息的量。另外，可把共同出现的簇序列结合为一组合的簇或单词（例如，“the”）。

根据发明的一个替代方面，OCR 部件接收库、标志顺序和位置信息，并且从簇中获得和识别字符。OCR 部件可把一个或多个簇组合成一个字符。这是因为若干簇能表示一个字符，例如“e”，并且在不同的簇上也是。另外，OCR 部件能利用已接收的信息从已接收到的信息中产生单词、句子和段落。字符和/或单词随后由类似于公共 ASCII 库的一个库表示，进一步提高压缩。

在解码结束时，位流被解码来重建一重建的文档。如果压缩是无损的，则重建文档和文档完全相同。即使压缩是有损的，重建的文档一般和文档基本相似。可以理解的是重建的文档可以是一表征码或层，它与至少一个其它图像或

层结合，来形成文档图像。

已相对于具有白色背景上的黑色文本的文档而一般地描述了聚类系统 600。然而，可以理解的是聚类系统 600 能被用于文档图像，以识别任何颜色的簇，包括但不限于黑色。

鉴于上述示出的示例性系统，可根据本发明实现的方法将参照图 11 的流程图来更好地理解。虽然，出于简单解释的目的，按一系列块来显示和描述方法，但是要理解是本发明不限于块的次序，根据本发明，一些块可与这里所示和所述的其它块以不同次序和/或同时出现。而且，根据本发明，不是所有说明的块都被要求来执行方法。

本发明可在计算机可执行指令的一般上下文环境中描述，例如通过一个或多个部件执行的程序模块。通常地，程序模块包括例行程序、程序、对象、数据结构等，它们执行特定任务或实现特定的抽象数据类型。程序模块的功能性一般可以组合或分布，这按不同的实施例的需要。

图 11 是根据本方明的一个方面的聚类方法的流程图。在 1102 上提供了文档。该文档是二元文本图像并包括一定量的文本信息。文档可能是一任何尺寸的单页或多页文档。另外，文档可能另一个文档图像的的层或组成成分。文档中的文本信息一般是白色背景上的黑色，然而，文本信息也可以包括黑色背景上的白色(如反白显示)。

在 1104 上从文档中获得下一个的标志或标志。可以采用一个适合的方法或算法来获得标志，如关于图 1 和图 6 所描述的方法。另外，可对标志计算盒尺寸、油墨尺寸、重新调整大小的图像和重心。也可对标志计算其它的特性或特征。盒尺寸包括标志的 x 尺寸和 y 尺寸。另外，设置一指向簇表 1112 的起始的指针，以便于处理。如果在 1106 上不能获得标志，指示出文本中没有剩余的标志待被处理，则方法在 1108 结束或退出。从而，在 1108，文本中的标志已经被适当地处理和聚类。如果能获得标志，则在 1110 上检索现存簇中的下一个簇。该下一个一般从簇表 1112 中获得。簇表 1112 保存着现存簇。如果在 1114 没有现存簇的簇，则在 1116 建立一个新的簇。

新簇的建立指示出标志不能正确地匹配现存簇，或该标志是从文档中析取的第一标志。以上计算的标志的特性或特征部分地包括新的簇。新的簇通常被

加到簇表 1112 的底部，因为新的簇（作为新的）通常地不像其它现存簇那样可能被碰到。

否则，在 1118 执行盒尺寸匹配或比较。盒尺寸匹配将标志的 x 尺寸和 y 尺寸与下一个标志的 x 尺寸和 y 尺寸比较。一个二维表可被用来保存现存簇的盒尺寸信息，如关于图 7 所描述和描写的那样。典型地，对比较设置一阈值，如百分之十，或一固定尺寸的变化(如 1 或 0)，以允许标志和下一个簇的尺寸中有微小偏差。另外，可调节该阈值，来增加该方法的速度和/或减少发现的簇数目，从而增加压缩。如果标志的盒尺寸不能匹配下一个簇的盒尺寸（如，不在百分之十内），则该方法在 1110 继续，获得新的下一个簇。

如果在下一个簇的盒尺寸和标志的盒尺寸之间存在匹配，则在 1120 执行油墨尺寸匹配。油墨尺寸匹配把标志的油墨尺寸与下一个簇的油墨尺寸比较，如果它们是相同的和/或基本相似，则认为是匹配。油墨尺寸是在标识中的黑像素的数目对所有像素数目的百分比。从而，可出现标志的油墨尺寸和下一个簇的油墨尺寸之间的偏差，并且只要标志的油墨尺寸和下一个簇的油墨尺寸都在一个可接受的阈值内（即，基本相似），则被认为是一个匹配。阈值确定可接受的值的范围，并能够被调节来允许较大的或较小的油墨尺寸变化。另外，阈值能够被调节来增加方法的速度和/或减少发现的簇的数目，从而增加压缩。如果标志的油墨尺寸不能匹配下一个簇的油墨尺寸（如，不在百分之十内），也被称为是不匹配，则方法在 1110 继续，获得新的下一个簇。

如果在下一个簇的油墨尺寸和标志的油墨尺寸之间存在匹配，则在 1122 执行特征映射匹配。特征映射匹配把标志的重新调整大小的图像与下一个簇的重新调整大小的图像比较，如果它们是相同的和/或基本相似，则认为是匹配。重新调整大小的图像是标志或簇的 3 乘 3 的像素表示，也可称为重新调整大小的标志图像。把重新调整大小的图像计算为一个较小的版本，例如，标志或簇的 3 乘 3 的像素图像。重新调整大小的图像一般在像素尺寸上远远小于标志的图像。多种算法可以被用来将标志图像缩小成重新调整大小的标志图像。其中一个算法是将标志分成 9 个区域（3 乘 3），以热点为中心，并按比例缩放以覆盖标志的范围，并计算标志的平均数，当它们与这些区域的每一个相交时。另一个算法是从位图中移除或删除若干像素。然而，其它的算法可以被用来生

成重新调整大小的图像，而仍然按照本发明。

从而，执行在标志的重新调整大小的图像的像素和下一个簇的重新调整大小的图像的像素之间的比较。如果所有 9 个像素或在其变化（阈值）匹配（如 7 或更多的匹配），标志的重新调整大小的图像和下一个簇的重新调整大小的图像就被认为是匹配。如用盒尺寸匹配和油墨尺寸匹配一样，可调整特征映射匹配的阈值，来增加方法的速度和/或减少发现的簇数目，从而增加压缩。如果标志的重新调整大小的图像不能匹配下一个簇的重新调整大小的图像，也被称为不匹配，则方法在 1110 继续，获得新的下一个簇。

如果下一个簇的重新调整大小的图像和标志的重新调整大小的图像之间存在匹配，则在 1124 执行 xor 油墨匹配。xor 油墨匹配计算标志的位图和下一个簇的位图之间的距离，也称为 xor 距离。对标志和下一个簇生成错误映射。错误映射是标志的位图图像和下一个簇的位图图像之间的按位异或，并以重心为中心。接着，通过对每个不匹配位添加 1 来计算距离。如果距离小于一阈值，那么就存在匹配。如果距离大于一阈值，则是不匹配，该方法就在 1110 处继续，获得新的下一个簇。可按需设置和/或调节阈值。

如果对于 xor 油墨匹配存在匹配，则在 1126 执行 “wan” 匹配。该 wan 匹配，称为加权与非(WAN)，类似于 1124 上的 xor 油墨匹配而被执行。该 wan 匹配也产生错误映射，但当计算加权时把从黑到白的错误与从白到黑的错误区分对待。如果距离大于一阈值，就是不匹配，该方法就在 1110 处继续，获得新的下一个簇。如果距离小于一阈值，那么就存在匹配，该方法就继续执行到 1128，在那里标志被添加到下一个簇。

该标志被添加到下一个簇，在其中于该簇相关联的标志数加 1。此外，位置信息和一唯一的标识符可以与下一个簇一起被存储。簇表 1112 可被再分类，使得由该方法首先采用更多地公共的现存簇。在标志在 1128 处被加到下一个簇后，该方法在 1104 继续，获得另一个标志。

可以理解的是在本发明的一个替换方面中，可把现存簇被存储在局部和/或全局库中。该方法能在局部库中存储的现存簇上执行，直到处理了当前页为止。接着，在通过该方法继续处理下一个页之前，可把局部库中的簇与全局库中的簇相合并。在全局和局部库中的基本相似的簇可以被合并成一个簇，而

不相似的簇则被加到全局库。

图 12 是根据本发明的一个方面的使用计算有效的聚类的一个图像编码系统的框图。该系统包括表征码离析器 1202，前台后台分隔器 1204，聚类部件 1205，表征码编码器 1206，前台编码器 1208，后台编码器 1210 和组合部件 1212。

表征码离析器 1202 接收文档图像，并产生一表征码。文档图像是文档的数字化表示。文档图像可具有一页或多页。文档图像一般从一文档中被扫描。文档图像能够有任何的分辨率，一般被表示为“点每英寸”(dpi)。例如，传真文档通常用大约 100-200dpi 的分辨率。此外，文档图像能够有任何的像素尺寸或尺寸。例如，文档图像可以是 640 像素乘 480 像素和/或 A4 尺寸。

由表征码离析器 1202 所生成的表征码随后被用来将文档图像分割或划分成前台和后台图像。表征码，也称为表征码图像，是二元图像，其中每个像素的值确定该像素是属于前台图像还是属于后台图像。表征码离析器 1202 生成减少表征码、前台图像和后台图像的组合尺寸的表征码。

许多方法可以被用来生成减少表征码、前台图像和后台图像的组合尺寸的表征码。相似区域或具有小的或减少数目的变化的区域趋向于比具有较大变化的区域能更好地压缩。例如，单色图像的压缩将好于在强度和颜色中变化大的图像的压缩。对于一个具有 N 像素的文档图像，有 2^N 个可能的表征码。从而，有可能仔细检查每个可能的表征码，并确定哪一个产生最小的整体的组合图像。然而，仔细检查每个可能的表征码在计算上开销很大。小的子区域，例如 2 乘 2 个像素或 4 乘 4 个像素，可分析变化。子区域的像素能够被分成前台或后台，来减少在前台和后台中的变化。然后可把子区域合并到一起来生成减少表征码、前台图像和后台图像的组合尺寸的表征码。可使用能减少表征码、前台图像和后台图像的整体尺寸的其它方法。

前台后台分隔器 1204 接受来自表征码离析器 1202 的表征码和文档图像。前台后台分隔器 1204 用表征码来从文档图像中产生前台图像和后台图像。对于文档图像的每一个像素，参考对应的表征码的像素。根据该表征码的对应的像素，把该像素分配给前台图像或后台图像。例如，如果表征码的对应的像素是“1”，则把像素分配给前台图像。相反地，如果表征码的对应的像素是“0”，则把像素分配给后台图像。可以理解的是不管是“0”还是“1”指示的前台或后台

是可以变化的。

后台图像和前台图像现在有空洞或空区域，在那里像素针对着其它图像。能够以任何方法处理空区域，来减少前台和后台图像的整体压缩尺寸。一个方法是将这些空的区域填满无关紧要的像素。无关紧要的像素被选出来，以便增加压缩和减少图像的尺寸。其它的方法可以被使用，并仍然按照本发明。

另外，前台图像和后台图像相互脱节。然而，这可能导致可视的边缘以一个最终的重新结合的文档图像而再现。很多方法可以用来减少这些可视边缘。前台和后台图像可被伸展一定数量的像素到彼此之中，使它们不再脱节。从而，可减少重新结合的文本图像中的可视边缘。

聚类部件 1205 接收来自表征码离析器 1202 的表征码，并产生簇的库、表征码顺序和位置信息。聚类部件 1205 从表征码中一次一个地发现和移除标志。聚类部件 1205 使用至少一个存储现存簇的特性的表。从一文档中析出的标志的特性迅速与至少一个表中的现存簇的特性相比较。此比较不需要标志位图对现存簇位图的费时的和开销大的逐位比较，来识别不匹配。如果此标志的特性未能和表中的任何特性相匹配，则此标志被添加为一个新簇，避免了逐位比较。

可能在所述至少一个表中的一个表中的簇的一个特性是 x 尺寸和 y 尺寸。x 尺寸和尺寸提供了现存簇的尺寸信息。标志的 x 尺寸和 y 尺寸能迅速地与现存簇的 x 尺寸和 y 尺寸比较，以识别不匹配。在所述至少一个表中的一个表中的簇的另一个特性是油墨尺寸。通常油墨尺寸指在标志或簇中相对于所有像素的黑像素的比率。类似地，标志的油墨尺寸能够迅速地与现存簇的油墨尺寸相比较，以识别不匹配。在所述至少一个表中的一个表中的簇的最后一个特性是一个简化的标志或图像。简化的标志或图像是标志和/或簇的像素尺寸简化版本位图。同样地，简化的标志能迅速与现存簇的简化标志或简化图像比较，以识别不匹配。这些特性关于图 1 和图 6 而被更进一步详细地讨论。

如果任何现存簇保持没有由使用所述至少一个表的比较不匹配，则标志的位图与剩余的现存簇的位图相比较。该位图比较是逐位比较。通过逐位比较来计算一个距离，如“xor”距离。距离越短，匹配越接近。其它适合的距离计算也能够被使用，并仍然按照本发明。如果识别了最匹配的现存簇，则标志被加到属于该匹配的现存簇的一组标志。

表征码编码器 1206 接收来自聚类部件 1205 的表征码、簇的库、标志顺序，以及位置信息，并编码表征码，以从表征码中产生压缩位或压缩的表征码位流。任何编码技术可用于表征码编码器 1206。然而，可以理解的是表征码是二元的，并且因而应选择利用表征码的二元特征的压缩方案。典型地，使用二值无损压缩方案来编码表征码。然而，可使用其它压缩方法或方案。

能被用来编码此表征码的一个典型的压缩方案是 CCITT（国际电话和电报咨询委员会）。CCITT 当前被认为是 ITU-T 国际电信联盟-电信部门（在 10994 中改名），它是一个标准组以及对于传真/调制解调器通信的一个无损压缩技术的名字。这个类型的压缩更适合工作于黑色和白色文本和图像，二值图像。对于老版本的 V.42bis 的典型压缩率是 4:1，而对于基于 Lempel-Ziv-Jeff-Heath(LZJH)压缩算法的新版本的 V.44 2000，压缩率是 6:1。

前台编码器 1008 接收前台图像，并编码前台图像，以从前台图像产生压缩位或压缩的前台图像位流。任何编码技术能用于前台编码器 1208。例如，可使用渐进波长编码或渐进变换编码来编码前台图像。

后台编码器 1210 接收后台图像，并编码此后台图像，以从后台图像中产生压缩位或压缩的后台图像位流。例如，可使用渐进波长编码或渐进变换编码来编码后台图像。

如上所述，前台图像和后台图像有无关紧要的区域。无关紧要区域能被许多方式所处理。一个方法是用数据填充此无关紧要区域并随后使用一个规范的压缩技术。最简单的方法将用对于该图像的平均像素值来填充图像的此无关紧要区域。然而，此方法在表征码边界造成尖锐的间断。此方法还增加了给定的峰值信噪比（PSNR）所需要的比特率，并在接近表征码或无关紧要区域的边界处产生值得注意的振荡。另外一个方法是使用最接近的非表征码（或非无关紧要区域）像素的颜色来给每个像素涂颜色。一个标准的形态算法允许通过仅仅通向表征码下的 Voronoi-填充的无关紧要区域的所有像素上的两条通路来执行处理。随后，当前台或后台图像被重建时，被重建图像是经低通的，并且已知的像素随后被恢复到它们的正确值。如果低通滤波器的截断频率很低，会出现尖锐边缘，导致所需的比特率的增加和在接近边界处的值得注意的振荡。

组合部件 1212 接收来自表征码编码器 1206、前台编码器 1208 和后台编码

器 1210 的压缩位，并将位组合成一输出流或者输出文件。组合部件 1212 可在输出文件中包括头部信息，标识或提供诸如编码类型、字典、库等之类的信息，这些信息能被解码器用来重建文档图像。

为了对本发明的各种方面提供额外的上下文环境，图 13 和随后的讨论意在提供一个可能合适的计算环境 1310 的一个简洁的，通用的描述，在其中本发明的各种方面可能被实现。可以理解的是计算环境 1310 只是一个可能的计算环境，并且不意图用可采用本发明的环境来限制计算环境。虽然已在上面可运行于一个或多个计算机上的计算机可执行指令的一般上下文环境中描述了本发明，但是可认识到也可能以与其它程序模块的组合来实现本发明，和/或实现成硬件和软件组合。通常地，程序模块包括例行程序、程序、部件、数据结构等，它执行特定任务或实现特定的抽象数据类型。而且，可以理解的是发明方法可实践于其它计算系统配置，包括单处理器或多处理器计算机系统、小型机、大型机、以及个人计算机、手持计算设备、基于微处理或可编程的用户电子设备等，它们的每一个可操作上与一个或多个的相关设备相连接。本发明的说明的各方面也可实践于分布式计算环境中，在那里由通过一通信网络连接的远程处理设备执行特定任务。在一个分布式计算环境中，程序模块可位于本地和远程存储器存储设备中。

图 13 说明可能的硬件配置来支持在此描述的系统和方法。可以理解的是尽管一个独立的体系结构被举例说明，但是任何合适的计算环境根据本发明能被使用。例如，根据本发明可采用的计算体系结构包括但不限于独立的、多处理器、分布式、客户/服务器、小型机、大型机、巨型机、数字和模拟结构。

参考图 13，用于实现本发明的各种方面的一个示例性环境 1310 包括计算机 1312，它包括处理单元 1314，系统存储器 1316，以及系统总线 1318，系统总线将包括系统存储器的各种系统部件耦合到处理单元 1314。处理单元 1314 可以是各种商业上可获得的处理器中的任何一种。双微处理器和其它多处理器体系结构也能作为处理单元 1314 被使用。

系统总线 1318 可以是若干类型总线结构中的任何一种，包括存储器总线或存储控制器，外围总线，和使用各种商业上可获得的总线体系结构中的任何一种的局部总线。计算机存储器 1316 包括只读存储器（ROM）1320 和随机存

取存储器 (RAM) 1322。基本输入/输出系统 (BIOS)，包含存储于 ROM 1320 中的帮助在计算机 1312 内的元素之间转移信息的基本例行程序，如在启动时。

计算机 1312 还可包括硬盘驱动器 1324，磁盘驱动器 1326，例如，从一个可拆卸磁盘 1328 上读取或写入，以及光盘驱动器 1330，例如用于读取 CD-ROM 盘片 1332，或读取或写入其它光学媒体。硬盘驱动器 1324、磁盘驱动器 1326，和光盘驱动器 1330 通过硬盘驱动器接口 1334，磁盘驱动器接口 1336，和光盘驱动器接口 1338 分别连接到系统总线 1318。计算机 1312 一般包括至少一些形式的计算机可读媒体。计算机可读媒体可以是任何可用媒体，它能由计算机 1312 存取访问。作为例子，但不是限制，计算机可读媒体可包括计算机存储媒体和通信媒体。计算机存储媒体包括以任何方法或技术实现的用于诸如计算机可读指令、数据结构、程序模块或其它数据之类的信息的存储的易失性和非易失性、可拆卸和不可拆卸的媒体。计算机存储媒体包括但不限于，RAM，ROM，EEPROM，闪存或其它存储技术，CD-ROM，数字通用磁盘 (DVD) 或其它磁性存储设备，或任何其它能被用来存储期望信息和能被计算机 1312 存取访问的媒体。通信媒体一般具体化为计算机的可读指令，数据结构、程序模块或在已调数据信号中的其它数据，例如一个载波或其它传送机械装置和包括任何信息传输媒体。术语“已调数据信号”意为以编码信号中的信息的方式使其一个或多个特性得到设置或改变的信号。作为例子，但不是限制，通信媒体包括有线媒体，例如有线网络或直接电缆连接，以及无线媒体，例如声音，RF，红外线和其它无线媒体。以上任何的结合也应包括在计算机可读媒体之内。

许多程序模块可存储在驱动器和 RAM 1322 中，包括操作系统 1340，一个或多个应用程序 1342，其它程序模块 1344，和程序不可中断数据 1346。在计算机 1312 中的操作系统 1340 可以是许多商业上可获得的操作系统中的任何一个。

用户可通过键盘 1348 和例如一个鼠标 1350 的指示设备输入命令和信息到计算机 1312 中。其它输入设备 (未显示) 可能包括麦克风，红外线远程控制，操纵杆，游戏手柄，圆盘式卫星电视天线，扫描仪等。这些和其它的输入设备常常通过串行端口接口 1352 连接到处理单元 1314，串行端口接口 1352 连接到系统总线 1314，但也可通过其它接口连接，例如并行端口，游戏端口，通用串

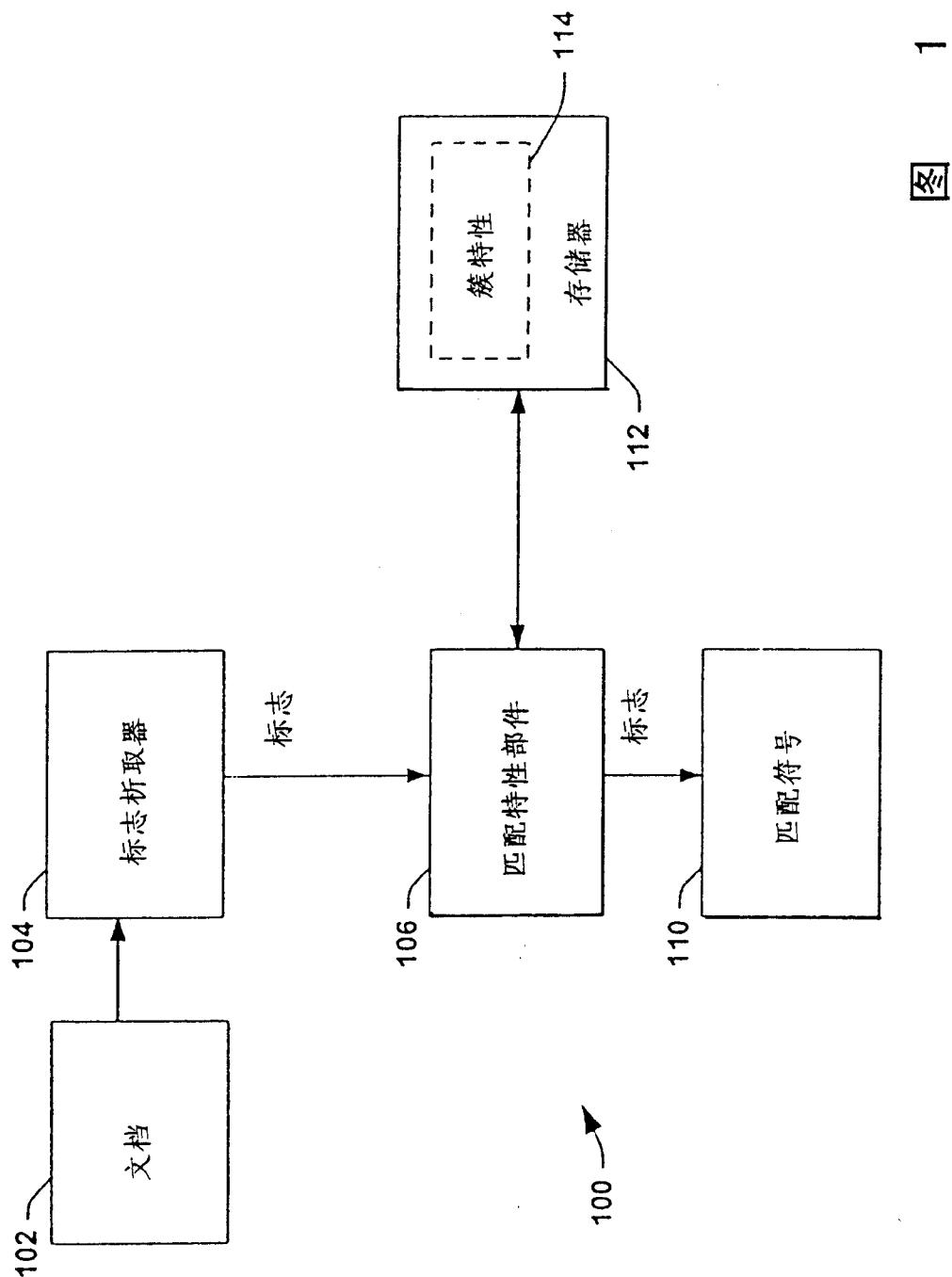
行总线（“USB”），红外线接口等。监视器 1354，或其它类型的输出设备通过一接口也连接到系统总线 1318 上，例如通过视频适配器 1356。除监视器之外，计算机一般还包括其它外围输出设备（未显示），例如扬声器，打印机等。

计算机 1312 可操作于使用逻辑和/或物理连接到一个或多个远程计算机的网络环境中，远程计算机例如远程计算机 1358。远程计算机 1358 可以是工作站，服务器计算机，路由器，个人计算机，基于娱乐应用的微处理器，对等设备或其它普通网络节点，一般包括关于计算机 1312 所描述的许多或所有元件，尽管为了简洁的目的，仅有一个存储器存储设备 1360 被举例说明。描述的逻辑连接包括局域网（LAN）1362 和广域网（WAN）1364。这种网络环境在办公室、企业范围的计算机网络、企业内部互联网、因特网中是很普通的。

当在 LAN 网络环境中使用时，计算机 1312 通过网络接口或适配器 1366 连接到局域网 1362。当在 WAN 网络环境中使用时，计算机 1312 一般包括调制解调器 1368，或连接到 LAN 上的通信服务器，或有其它手段在 WAN 1364 上建立通信，WAN 例如因特网。调制解调器 1368，可以是内置或外置的，通过串行端口接口 1352 连接到系统总线 1318。在连网环境中，相对于计算机 1312 而描述的程序模块或其一部分，可存储在远程存储器存储设备 1360 中。可以理解的是所示的网络连接是示例性的，可使用在计算机之间建立通信连接的其它方法。

图 14 是样本计算环境 1400 的示意性框图，本发明可与之相互作用。系统 1400 包括一个或多个客户端 1410。客户端 1410 可以是硬件和/或软件（例如，线程，过程，计算设备）。系统 1400 还包括一个或多个服务器 1430。服务器 1430 也可以是硬件和/或软件（例如，线程，过程，计算设备）。服务器 1430 能提供线程执行环境，通过使用本发明去执行转换。在一个客户端 1410 和一个服务器 1430 之间的一个可能的通信可以适应于在两个或多个计算机过程之间传送的形式。系统 1400 包括通信架构 1450，它被使用来促进客户端 1410 和服务器 1430 之间的通信。客户端 1410 操作上连接到一个或多个客户数据存储器 1460，它能被使用来存储客户 1410 本地信息。类似的，服务器 1430 可操作上连接到一个或多个服务器数据存储器 1440，它能被使用来存储服务器 1430 的本地信息。

上述已经被描述的包括本发明的例子。当然，为了描述本发明的目的描述每个可能的部件或方法论的联合是不可能的，不过本领域的普通技术人员可认识到本发明的许多更进一步的组合和改变是可能的。从而，本发明意在包括所有类似的变更，修正和变化，并都出于权利要求书的精神和范围内。此外，就术语“包括”的范围来说，或被使用在详细描述中，或被使用在权利要求书中，此术语类似于术语“包含”，用作为权利要求书中的过渡单词。



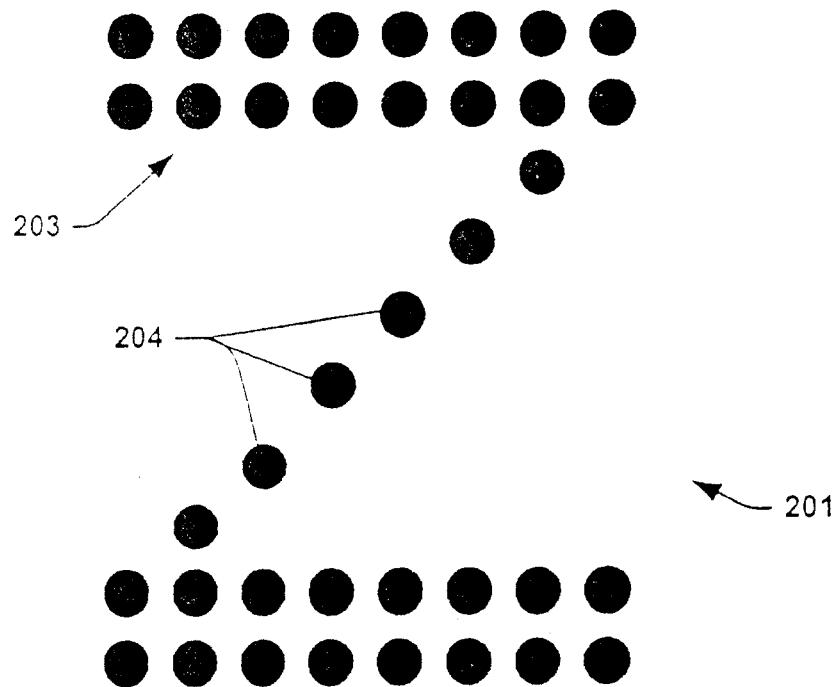
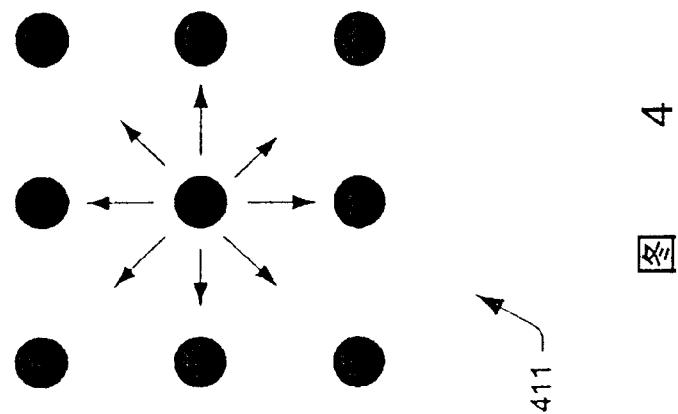


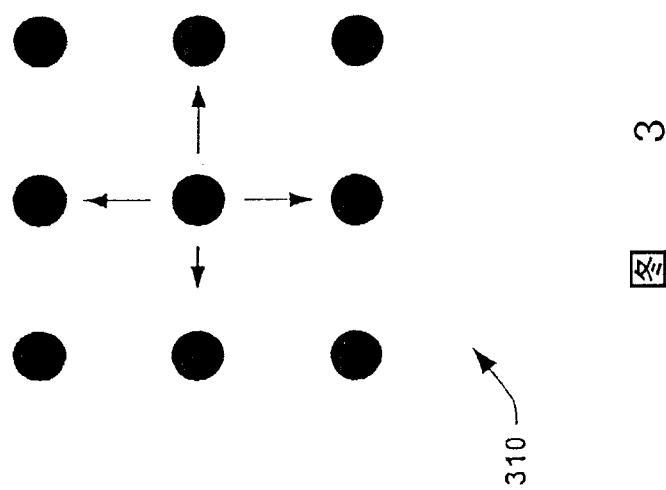
图 2



图

4

411



图

3

310

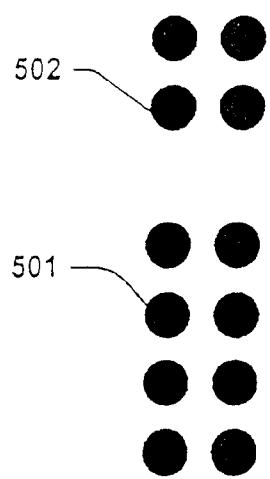


图 5

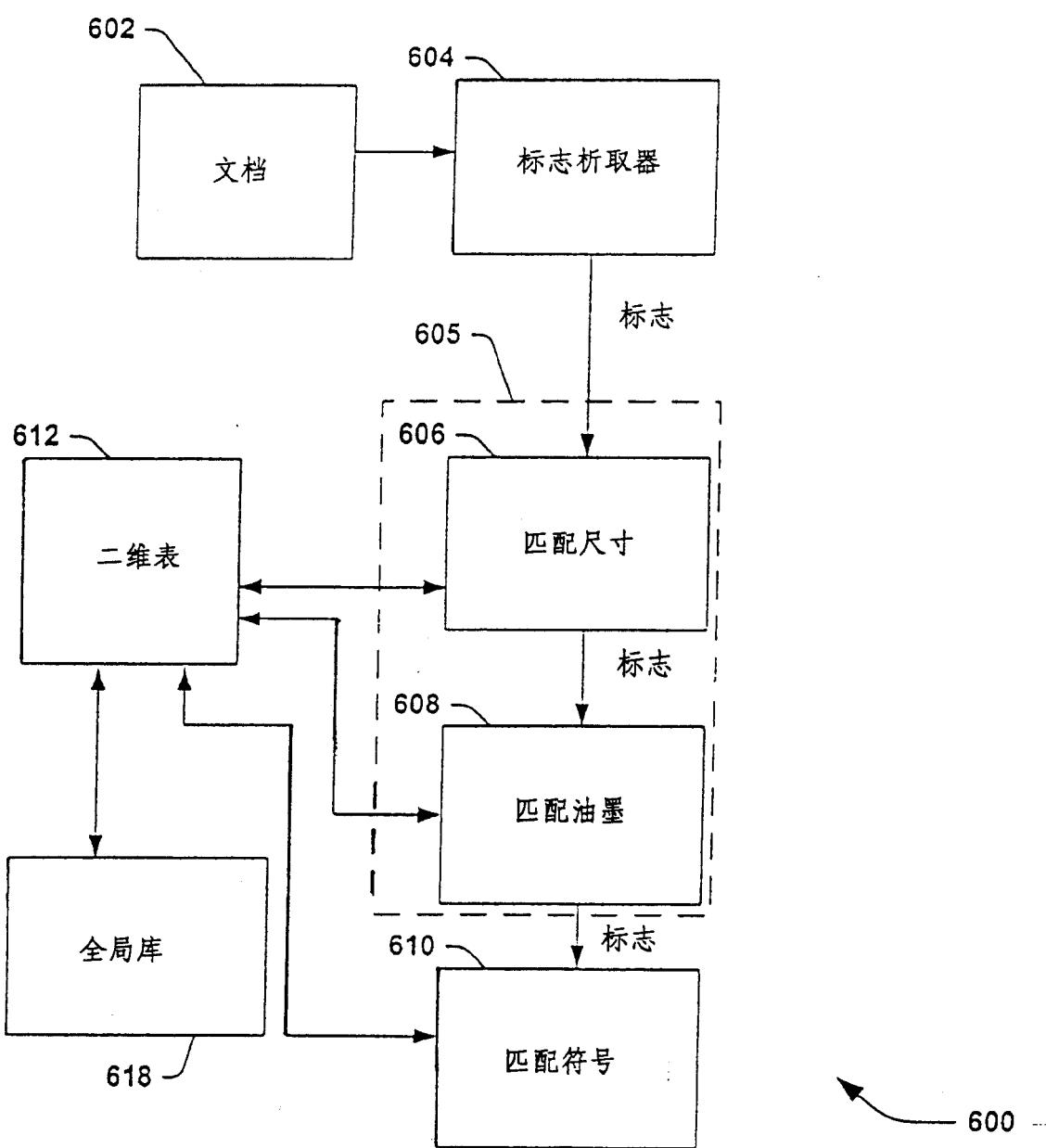
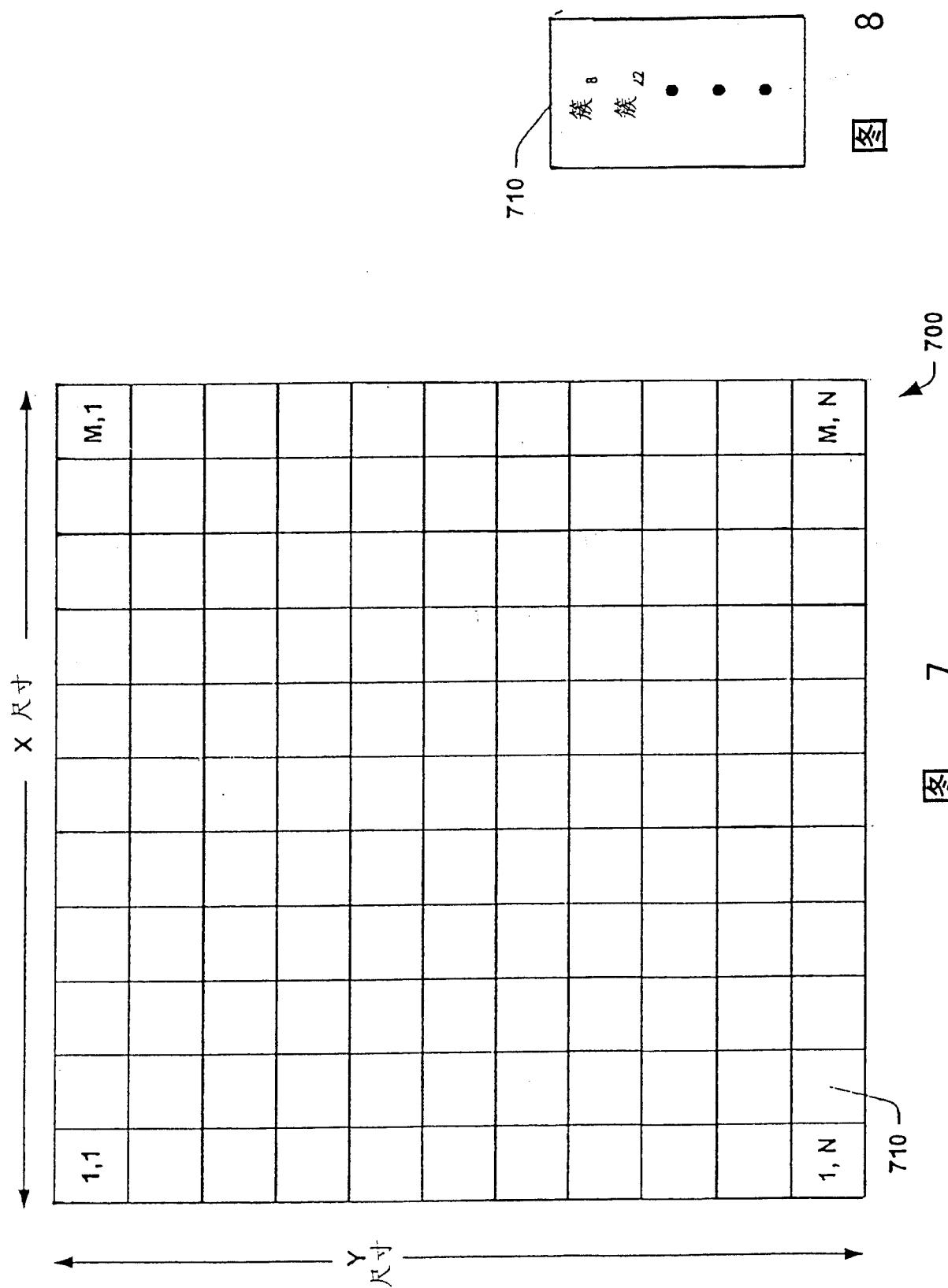


图 6

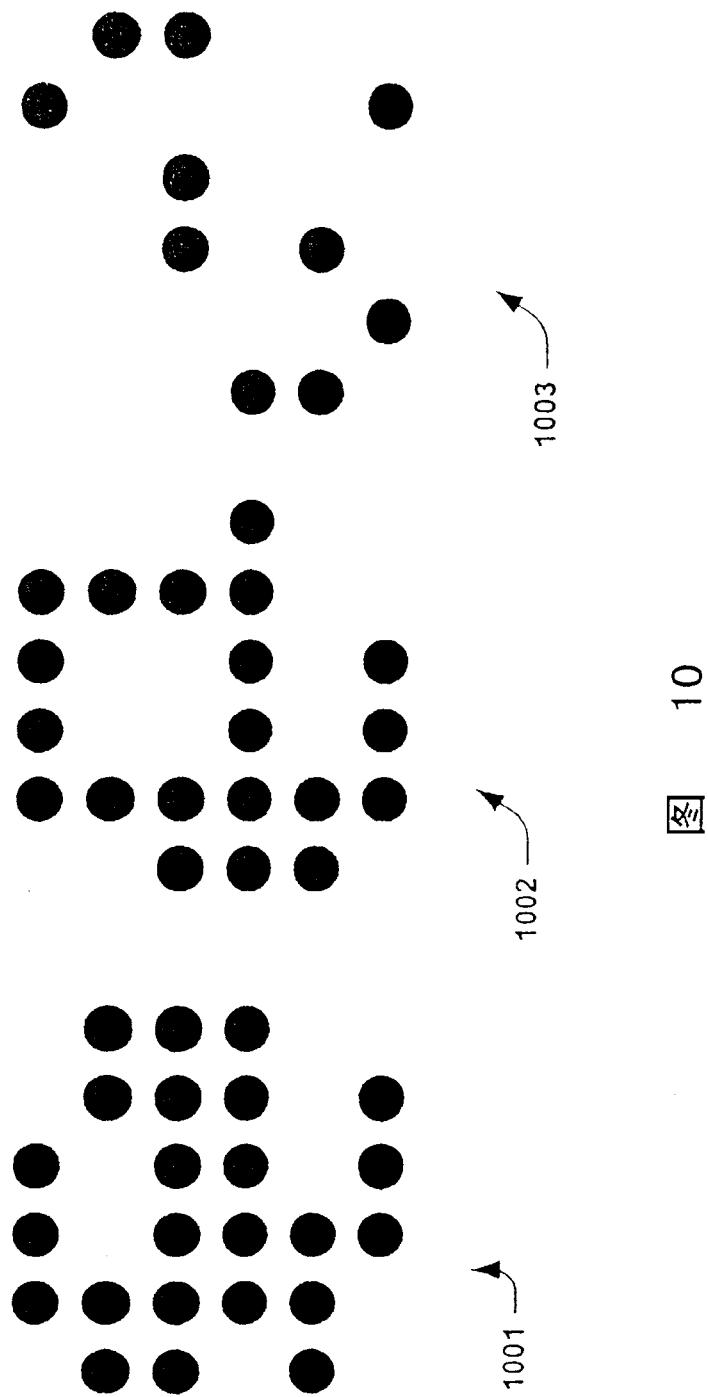


唯一标识符	X尺寸	Y尺寸	位图	油墨尺寸	重新调整大小的 图像
0	10	20		65	
4	12	15		65	
3	8	15		80	
2	11	20		75	
25	4	11		55	

901

9

冬



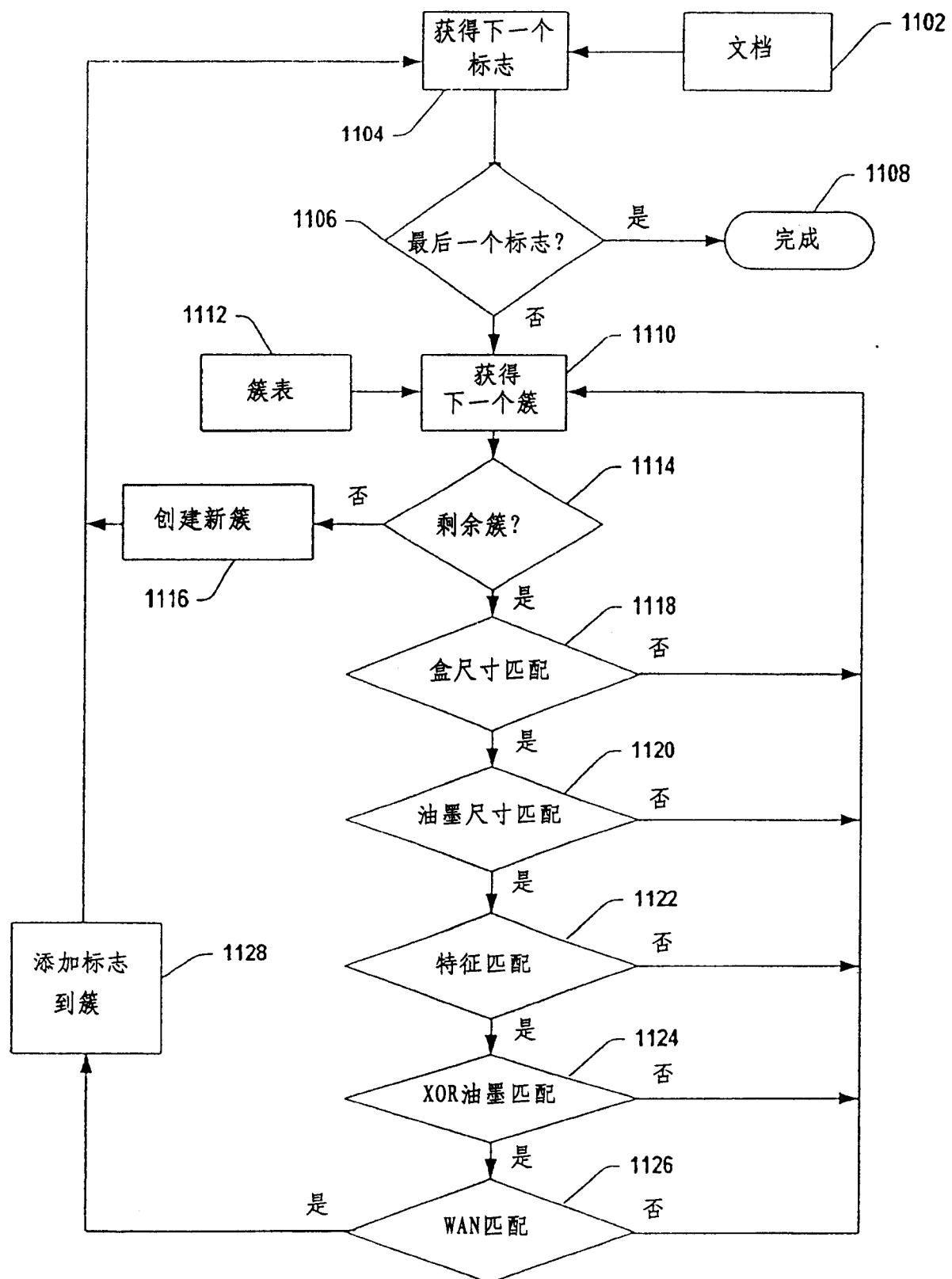


图 11

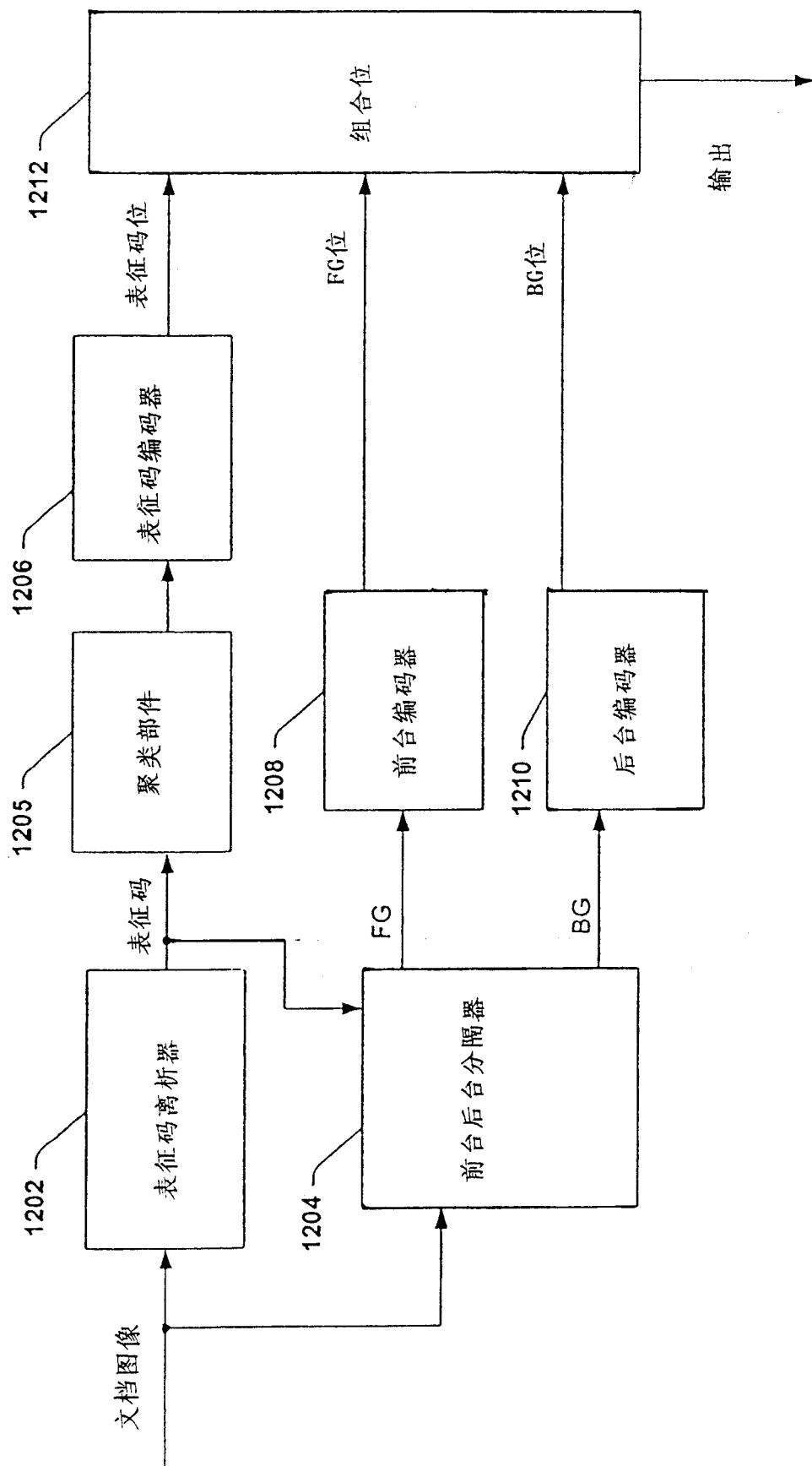


图 12

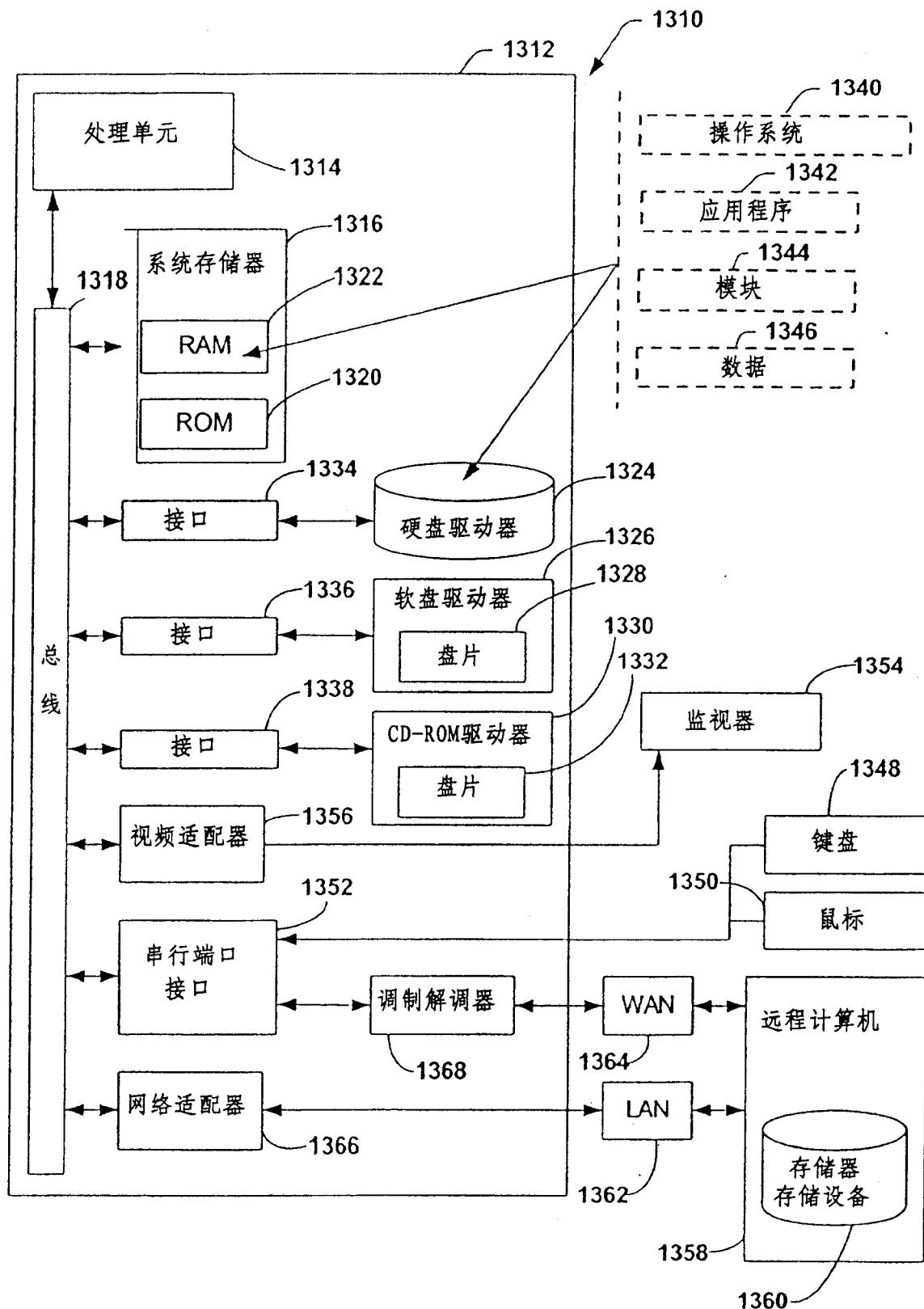


图 13

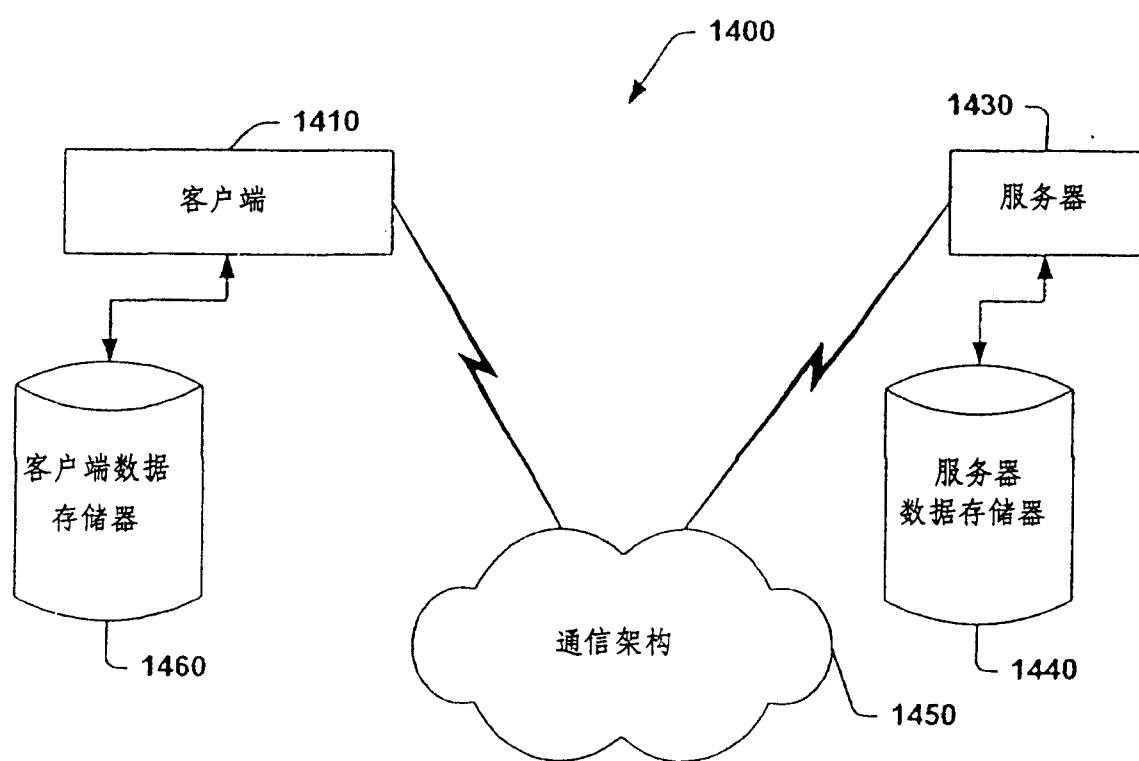


图 14