(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101576932 B
(45)授权公告日 2012.07.04

(21)申请号 200910146726.5

(22)申请日 2009.06.16

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛

(72)发明人 贾梦雷

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘

(51)Int.Cl.
G06F 17/30(2006.01)
G06K 9/68(2006.01)

(56)对比文件
CN 101359366 A,2009.02.04,
CN 101290634 A,2008.10.22,
US 2007127813 A1,2007.06.07,
刘金松 等．基于网页上下文分析的图片检索．《语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集》．2003,507-512.

审查员 刘琳

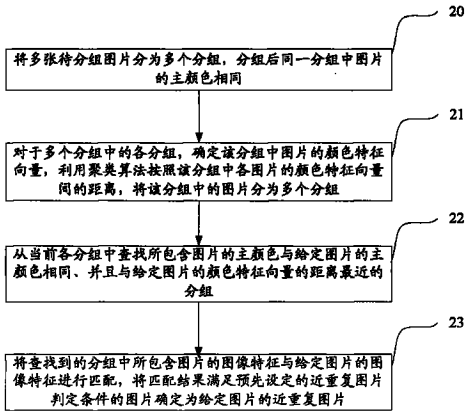
权利要求书 8 页 说明书 14 页 附图 5 页

(54)发明名称

近重复图片的计算机查找方法和装置

(57)摘要

本申请实施例公开了一种近重复图片查找方法，该方法为：将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多组；从多组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。本申请实施例还公开了一种近重复图片查找装置。本申请实施例还公开了近重复图片的计算机查找方法和装置。采用本申请，能够有效提高查找给定图片的近重复图片的效率。



CN 101576932 B

1. 一种近重复图片的计算机查找方法，其特征在于，该方法包括：

读取数据库中存储的多张待分组图片；

将读取的多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组，并将该多个分组进行储存；

读取给定图片；查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组，并读取该分组中的图片；

将所述读取的图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

2. 如权利要求 1 所述的方法，其特征在于，所述利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组并将该多个分组进行储存包括：

A1、将该分组作为当前图片分组及第一分组，将当前图片分组中图片的主颜色设置为图片签名树的子树的根节点，将该根节点作为当前父节点；

B1、利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

C1、对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 D1，否则，到步骤 E1；

D1、将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 B1；

E1、将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照第一分组中各 图片的颜色特征向量间的距离，将第一分组中的图片划分为多个分组后的一个分组；

F1、将建立的图片签名树进行储存。

3. 如权利要求 2 所述的方法，其特征在于，所述查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组包括：

A2、在所述存储的图片签名树中查找根节点为给定图片的主颜色的子树，将所述子树的根节点作为当前父节点；

B2、在所述子树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则到步骤 C2；若该子节点为叶子节点，则到步骤 D2；

C2、判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 B2，否则，停止查找操作；

D2、将该叶子节点所包含各图片构成的分组，确定为所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组。

4. 如权利要求 1 所述的方法，其特征在于，所述给定图片为用户输入的图片或用户通过互联网搜索得到的图片。

5. 如权利要求 1 所述的方法，其特征在于，在将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片之后，该方法进一步包括：

将所述近重复图片展现给用户。

6. 一种近重复图片的计算机查找方法，其特征在于，该方法包括：

读取多张待分组图片；

确定读取的各待分组图片的颜色特征向量；利用聚类算法按照读入内存的多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个 分组，并将该多个分组进行储存；

读取给定图片；查找与所述给定图片的颜色特征向量的距离最近的分组，并读取所述查找到的分组中的图片；

将所述读取的图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

7. 如权利要求 6 所述的方法，其特征在于，所述利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片分为多个分组并将该多个分组进行储存包括：

A1、设置图片签名树的根节点，并将该根节点作为当前父节点；将所述多张待分组图片构成的分组作为当前图片分组；

B1、利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

C1、对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 D1，否则，到步骤 E1；

D1、将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 B1；

E1、将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组后的一个分组；

F1、将建立的图片签名树进行存储。

8. 如权利要求 7 所述的方法，其特征在于，所述查找与给定图片的颜色特征向量的距离最近的分组包括：

A2、将所述存储的图片签名树的根节点作为当前父节点；

B2、在所述图片签名树中查找当前父节点的子节点，对于查找到的各子节 点，若该子节点为中间节点，则到步骤 C2；若该子节点为叶子节点，则到步骤 D2；

C2、判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 B2，否则，停止查找操作；

D2、将该叶子节点所包含各图片构成的分组，确定为与给定图片的颜色特征向量的距离最近的分组。

9. 如权利要求 6 所述的方法，其特征在于，所述给定图片为用户输入的图片或用户通过互联网搜索得到的图片。

10. 如权利要求 6 所述的方法，其特征在于，在将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片之后，该方法进一步包括：

将所述近重复图片展现给用户。

11. 一种近重复图片查找方法，其特征在于，该方法包括：

将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

从所述多个分组中查找所包含图片的主颜色与给定图片的主颜色相同的分组；

将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片；

所述近重复图片判定条件包括：

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的 距离小于设定的距离门限值、以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

12. 一种近重复图片查找方法，其特征在于，该方法包括：

将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组；

从当前各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；

将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

13. 如权利要求 12 所述的方法，其特征在于，所述确定该分组中图片的颜色特征向量包括：

将所述图片划分为 N 块，N 为大于 1 的整数；

对于所述 N 块中的每一块，统计该块上设定种颜色对应的像素点个数；

将统计得到的各像素点个数构成的向量确定为所述图片的颜色特征向量。

14. 如权利要求 12 所述的方法，其特征在于，所述利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组包括：

A1、将该分组作为当前图片分组及第一分组，将当前图片分组中图片的主颜色设置为图片签名树的子树的根节点，将该根节点作为当前父节点；

B1、利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

C1、对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 D1，否则，到步骤 E1；

D1、将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 B1；

E1、将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点

所包含各图片构成的分组，确定为利用聚类算法按照第一分组中各图片的颜色特征向量间的距离，将第一分组中的图片划分为多个分组后的一个分组。

15. 如权利要求 14 所述的方法，其特征在于，所述步骤 C1 中分组停止条件包括以下三种中的一种或任意组合：

分组中包含的颜色特征向量的个数小于设定的向量数门限值；

分组中各颜色特征向量到该分组中各颜色特征向量的聚类中心的距离均小于设定的距离门限值；

分组的分裂次数超过设定的分裂数门限值，所述分组的分裂次数是从所述待分组图片到得到该分组的时间段内执行分组操作的次数。

16. 如权利要求 14 所述的方法，其特征在于，所述从当前各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组包括：

A2、在所述图片签名树中查找根节点为给定图片的主颜色的子树，将所述子树的根节点作为当前父节点；

B2、在所述子树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则到步骤 C2；若该子节点为叶子节点，则到步骤 D2；

C2、判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 B2，否则，停止查找操作；

D2、将该叶子节点所包含各图片构成的分组，确定为所包含图片的主颜色 与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组。

17. 如权利要求 16 所述的方法，其特征在于，所述步骤 C2 中设定条件为：

所述距离小于预先设定的距离阈值；或者，

所述距离为查找到的各中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离中的最小值。

18. 如权利要求 12 所述的方法，其特征在于，在将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配之前，该方法进一步包括：

将选定的颜色空间量化到 M 种颜色，M 为大于 1 的整数；

对于查找到的分组中所包含图片和所述给定图片中的各图片，统计该图片上 M 种颜色中各颜色对应的像素点个数，计算统计得到的最大像素点个数占该图片上像素点个数总和的比例，将计算结果作为该图片的主颜色率；将该主颜色率和 / 或该图片的颜色特征向量确定为该图片的图像特征。

19. 如权利要求 18 所述的方法，其特征在于，所述近重复图片判定条件包括：

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，以及所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值，以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

20. 一种近重复图片查找方法，其特征在于，该方法包括：

确定多张待分组图片中各待分组图片的颜色特征向量；利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为 多个分组；

从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

21. 如权利要求 20 所述的方法，其特征在于，所述确定待分组图片的颜色特征向量包括：

将所述待分组图片划分为 N 块，N 为大于 1 的整数；

对于所述 N 块中的每一块，统计该块上设定种颜色对应的像素点个数；

将统计得到的各像素点个数构成的向量确定为所述待分组图片的颜色特征向量。

22. 如权利要求 20 所述的方法，其特征在于，所述利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片分为多个分组包括：

A1、设置图片签名树的根节点，并将该根节点作为当前父节点；将所述多张待分组图片构成的分组作为当前图片分组；

B1、利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

C1、对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 D1，否则，到步骤 E1；

D1、将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 B1；

E1、将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组后的一个分组。

23. 如权利要求 22 所述的方法，其特征在于，所述步骤 C1 中分组停止条件包括以下三种中的一种或任意组合：

分组中包含的颜色特征向量的个数小于设定的向量数门限值；

分组中各颜色特征向量到该分组中各颜色特征向量的聚类中心的距离均小于设定的距离门限值；

分组的分裂次数超过设定的分裂数门限值，所述分组的分裂次数是从所述待分组图片到得到该分组的时间段内执行分组操作的次数。

24. 如权利要求 22 所述的方法，其特征在于，所述从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组包括：

A2、将所述图片签名树的根节点作为当前父节点；

B2、在所述图片签名树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则到步骤 C2；若该子节点为叶子节点，则到步骤 D2；

C2、判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 B2，否则，停止查找操作；

D2、将该叶子节点所包含各图片构成的分组，确定为与给定图片的颜色特征向量的距

离最近的分组。

25. 如权利要求 24 所述的方法，其特征在于，所述步骤 C2 中设定条件为：

所述距离小于预先设定的距离阈值；或者，

所述距离为查找到的各中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离中的最小值。

26. 如权利要求 20 所述的方法，其特征在于，在将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配之前，该方法进一步包括：

将选定的颜色空间量化到 M 种颜色，M 为大于 1 的整数；

对于查找到的分组中所包含图片和所述给定图片中的各图片，统计该图片上 M 种颜色中各颜色对应的像素点个数，计算统计得到的最大像素点个数占该图片上像素点个数总和的比例，将计算结果作为该图片的主颜色率；将该主颜色率和 / 或该图片的颜色特征向量确定为该图片的图像特征。

27. 如权利要求 26 所述的方法，其特征在于，所述近重复图片判定条件包括：

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，以及所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值，以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

28. 一种近重复图片查找装置，其特征在于，该装置包括：

第一分组单元，用于将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

向量确定单元，用于对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量；

第二分组单元，用于对于所述多个分组中的各分组，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组；

查找单元，用于从所述第二分组单元分组后的各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；

匹配单元，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

29. 如权利要求 28 所述的装置，其特征在于，所述第二分组单元包括：

子树建立单元，用于对于所述多个分组中的各分组，将该分组作为当前图片分组及第一分组，将当前图片分组中图片的主颜色设置为图片签名树的子树的根节点，将该根节点作为当前父节点，触发聚类分组单元；

聚类分组单元，用于利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数，触发递归建立单元；

递归建立单元，用于对于 K 组中的每一组，判断该组是否满足设定的分组停止条件，若是，则触发叶子节点建立单元；否则，触发中间节点建立单元；

中间节点建立单元，用于将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，触发聚类分组单元；

叶子节点建立单元，用于将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照第一分组中各图片的颜色特征向量间的距离，将第一分组中的图片划分为多个分组后的一个分组。

30. 如权利要求 29 所述的装置，其特征在于，所述查找单元包括：

第一查找单元，用于在所述图片签名树中查找根节点为给定图片的主颜色的子树，将所述子树的根节点作为当前父节点；

第二查找单元，用于在所述子树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则触发中间节点处理单元；若该子节点为叶子节点，则触发叶子节点处理单元；

中间节点处理单元，用于判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，触发第二查找单元，否则，停止查找操作；

叶子节点处理单元，用于将该叶子节点所包含各图片构成的分组，确定为 所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组。

31. 一种近重复图片查找装置，其特征在于，该装置包括：

向量确定单元，用于确定多张待分组图片中各待分组图片的颜色特征向量；

分组单元，用于利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组；

查找单元，用于从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

匹配单元，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

32. 如权利要求 31 所述的装置，其特征在于，所述分组单元包括：

初始化单元，用于设置图片签名树的根节点，并将该根节点作为当前父节点；将所述多张待分组图片构成的分组作为当前图片分组；

聚类分组单元，用于利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

递归建立单元，用于对于 K 组中的每一组，判断该组是否满足设定的分组停止条件，若是，则触发叶子节点建立单元；否则，触发中间节点建立单元；

中间节点建立单元，用于将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，触发聚类分组单元；

叶子节点建立单元，用于将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片 划分为多个分组后的一个分组。

33. 如权利要求 31 所述的装置，其特征在于，所述查找单元包括：

第一查找单元，用于将所述图片签名树的根节点作为当前父节点，在所述图片签名树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则触发中间节点处理单元；若该子节点为叶子节点，则触发叶子节点处理单元；

中间节点处理单元，用于判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，触发第一查找单元，否则，停止查找操作；

叶子节点处理单元，用于将该叶子节点所包含各图片构成的分组，确定为与给定图片的颜色特征向量的距离最近的分组。

近重复图片的计算机查找方法和装置

技术领域

[0001] 本申请涉及数字图像处理领域，尤其涉及一种近重复图片的计算机查找方法和装置。

背景技术

[0002] 目前，对于给定的两张图片，判断这两张图片是否相同具体采用如下图像特征提取法：

[0003] 首先，提取两张图片的图像特征，图像特征可以视为图片的签名；

[0004] 然后，比较两张图片的签名是否完全匹配，若是，则判断两张图片相同，否则，判断两张图片不相同。

[0005] 上述方法中，提取的图片的图像特征为该图片的颜色直方图向量。颜色直方图向量的具体提取方法如下：

[0006] 首先，选择一种颜色空间，如 RGB 空间，并将颜色空间进行量化，量化后的结果是若干种颜色；

[0007] 然后，统计图片的全部区域或部分区域中每一种颜色对应的像素个数，形成颜色直方图；

[0008] 最后，将形成的所有颜色直方图拼成一个向量，作为图片的签名。

[0009] 在需要从多张图片中查找与给定图片相同的图片时，具体做法是，按照上述图像特征提取法判断给定图片与多张图片中的各张图片是否相同，并将判断相同的图片作为查找结果。

[0010] 在实现本申请的过程中，发明人发现现有技术中存在如下技术问题：

[0011] 其一，在从多张图片中查找与给定图片相同的图片时，需要将给定图片与多张图片中的每一张图片进行比较，比较过程涉及图像特征提取等复杂过程，实现效率较低。

[0012] 其二，利用上述图像特征提取法并不能判断两张图片是否为近重复图片，因为在图片的局部颜色发生不大的变化时，例如在图片中嵌入了水印，图片的图像特征也会发生变化。因此也就无法从多张图片中查找与给定图片为近重复图片的图片。近重复图片是指，两张图片的主体内容基本相同，只是由于人工加入小面积的标志或水印，或是由于图片缩放等原因而引起了少量差异，将这两张图片称为近重复图片。

[0013] 发明内容

[0014] 本申请实施例提供一种近重复图片查找方法和装置，以及一种近重复图片的计算机查找方法和装置，用于提高从多张图片中查找给定图片的近重复图片的效率。

[0015] 本申请实施例提供一种近重复图片的计算机查找方法，该方法包括：

[0016] 将数据库中存储的多张待分组图片读入内存；

[0017] 将读入内存的多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分

组，并将该多个分组保存在硬盘上；

[0018] 将给定图片读入内存；从硬盘上查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组，并将该分组中的图片读入内存；

[0019] 将从硬盘上读入内存的图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0020] 本申请实施例提供一种近重复图片的计算机查找方法，该方法包括：

[0021] 将多张待分组图片读入内存；

[0022] 确定读入内存的各待分组图片的颜色特征向量；利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组，并将该多个分组保存在硬盘上；

[0023] 将给定图片读入内存；从硬盘上查找与所述给定图片的颜色特征向量的距离最近的分组，并将查找到的分组中的图片读入内存；

[0024] 将从硬盘读入内存的图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0025] 本申请实施例提供一种近重复图片查找方法，该方法包括：

[0026] 将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

[0027] 从所述多个分组中查找所包含图片的主颜色与给定图片的主颜色相同的分组；

[0028] 将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片；

[0029] 所述近重复图片判定条件包括：

[0030] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为0，并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

[0031] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值、以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

[0032] 本申请实施例提供一种近重复图片查找方法，该方法包括：

[0033] 将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

[0034] 对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组；

[0035] 从当前各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；

[0036] 将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0037] 本申请实施例提供一种近重复图片查找方法，该方法包括：

[0038] 确定多张待分组图片中各待分组图片的颜色特征向量；利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组；

[0039] 从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

[0040] 将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0041] 本申请实施例提供一种近重复图片查找装置，该装置包括：

[0042] 第一分组单元，用于将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

[0043] 向量确定单元，用于对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量；

[0044] 第二分组单元，用于对于所述多个分组中的各分组，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组；

[0045] 查找单元，用于从所述第二分组单元分组后的各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；

[0046] 匹配单元，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0047] 本申请实施例提供一种近重复图片查找装置，该装置包括：

[0048] 向量确定单元，用于确定多张待分组图片中各待分组图片的颜色特征向量；

[0049] 分组单元，用于利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组；

[0050] 查找单元，用于从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

[0051] 匹配单元，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0052] 本申请中，根据图片的主颜色和／或颜色特征向量将多张待分组图片进行分组，在各分组中查找所包含图片的主颜色与给定图片的主颜色相同和／或与给定图片的颜色特征向量的距离最近的分组，将查找到的分组中所包含的各图片与给定图片进行图像特征的匹配，将匹配结果满足设定的近重复图片判定条件的图片确定为给定图片的近重复图片。由于首先将多张待分组图片分组，在各分组中查找到满足一定条件的分组后，只将满足条件的分组中的图片与给定图片进行匹配，而不是将所有待分组图片均与给定图片进行匹配来确定给定图片的近重复图片，能够有效地提高查找给定图片的近重复图片的效率。

[0053] 附图说明

[0054] 图 1 为本申请实施例一的方法流程示意图；

[0055] 图 2 为本申请实施例二的方法流程示意图；

[0056] 图 3 为本申请实施例二中建立的图片签名树的结构示例图；

[0057] 图 4 为本申请实施例三的方法流程示意图；

[0058] 图 5 为本申请实施例提供的一种装置的结构示意图；

[0059] 图 6 为本申请实施例提供的另一种装置的结构示意图；

[0060] 图 7 为本申请实施例提供的又一种装置的结构示意图。

[0061] 具体实施方式

[0062] 为了以较高的效率实现从多张图片中查找到与给定图片为近重复图片的图片，本申请实施例提供图片查找方法，该方法中，根据图片的主颜色和 / 或颜色特征向量将多张待分组图片进行分组，然后查找所包含图片的主颜色与给定图片的主颜色相同和 / 或与给定图片的颜色特征向量的距离最近的分组，最后将该分组中所包含的各图片与给定图片进行图像特征的匹配，将匹配成功的图片确定为给定图片的近重复图片。

[0063] 图片的主颜色，是指该图片上对应像素点个数最多的颜色，具体确定方法可以为：首先，选择一种 RGB 空间作为颜色空间，将该颜色空间量化到 M 种颜色；然后，统计量化后的每种颜色在图片上对应的像素点个数；最后，选择像素点个数最多的颜色作为该图片的主颜色。这里，M 的取值为大于 1 的整数， 例如 512、256、1024 等。颜色空间是指，为了使各种颜色能按照一定的排列次序并容纳在一个空间内，将三维坐标轴与颜色的三个独立参数对应起来，使每一个颜色都有一个对应的空间位置，反过来，在空间中的任何一点都代表一个特定的颜色，将这个空间称为颜色空间。

[0064] 图片的颜色特征向量，是指一种或多种颜色在该图片上对应像素点个数构成的向量。图片的颜色特征向量的确定方法可以为：

[0065] 首先，将图片划分为 N 块；然后，对于划分后 N 块中的每一块，统计该块上设定种颜色对应的像素点个数；最后，将统计得到的各像素点个数构成的向量确定为该图片的颜色特征向量。这里，N 的取值为大于 1 的整数，例如 9、4、16 等。设定种颜色可以是一种或多种颜色，例如红、黄、蓝三基色，也可以是量化后的 M 种颜色中的一种或多种颜色。

[0066] 图片的颜色特征向量的确定方法还可以为：直接统计图片上设定种颜色对应的像素点个数，将统计得到的各像素点个数构成的向量确定为该图片的颜色特征向量。

[0067] 图片的图像特征，是指对该图片内容的描述信息。图片的图像特征可以有多种，例如，图片的颜色特征向量和 / 或图片的主颜色率等。图片的主颜色率的确定方法如下：首先，将选定的颜色空间量化到 M 种颜色，M 为大于 1 的整数；然后，统计该图片上 M 种颜色中各颜色对应的像素点个数，计算统计得到的最大像素点个数占该图片上像素点个数总和的比例，将计算结果作为该图片的主颜色率。

[0068] 本申请实施例提供的图片查找方法具体包括以下三种实施例：

[0069] 实施例一：

[0070] 本实施例中，根据图片的主颜色将多张待分组图片进行分组，参见图 1，具体包括以下步骤：

[0071] 步骤 10：将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

[0072] 步骤 11：从多个分组中查找所包含图片的主颜色与给定图片的主颜色相同的分组；

[0073] 步骤 12：将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行

匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为给定图片的近重复图片。

[0074] 步骤 12 中，在将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配之前，需要提取查找到的分组中所包含图片和给定图片中各图片的主颜色率和 / 或颜色特征向量等作为该图片的图像特征。

[0075] 近重复图片判定条件可以有以下两种 ：

[0076] 第一种，给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，并且给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值 ；主颜色率门限值在 0 到 1 之间取值。

[0077] 第二种，给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值、以及给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。距离门限值的取值为大于 0 的自然数，第一主颜色率门限值和第二主颜色率门限值在 0 到 1 之间取值，并且第一主颜色率门限值大于第二主颜色率门限值。

[0078] 下面结合具体计算机应用对本实施例进行说明 ：

[0079] 步骤 a1 ：将数据库中存储的多张待分组图片读取到内存中 ；

[0080] 步骤 a2 ：对读取到内存中的每一张待分组图片，确定该图片的主颜色 ；

[0081] 根据图片的主颜色将将读入内存的多张待分组图片划分为多个分组，同一分组中图片的主颜色相同，将每一个分组存储在硬盘上不同的分组数据库中，并建立主颜色标识与分组数据库地址的对应关系表 A ；

[0082] 计算每一张待分组图片的图像特征，建立图片标识与图像特征的对应关系表 B，将对应关系表 A 和 B 存放在硬盘上 ；

[0083] 步骤 a3 ：将给定图片和对应关系表 A、B 读入内存 ；

[0084] 确定给定图片的主颜色，从对应关系表 A 中查找给定图片的主颜色的标识对应的分组数据库地址，从硬盘上读取该分组数据库地址对应的分组数据库中保存的所有图片到内存 ；

[0085] 步骤 a4 ：从对应关系表 B 中查找从分组数据库中读取到的各个图片的标识对应的图像特征，将给定图片的图像特征分别与查找到各个图像特征分别进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图像特征对应的图片，确定为给定图片的近重复图片。

[0086] 实施例二 ：

[0087] 本实施例中，根据图片的主颜色和颜色特征向量将多张待分组图片进行分组，参见图 2，具体包括以下步骤 ：

[0088] 步骤 20 ：将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同 ；

[0089] 步骤 21 ：对于多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片分为多组 ；

[0090] 步骤 22 ：从多组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给

定图片的颜色特征向量的距离最近的分组；

[0091] 步骤 23：将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为给定图片的近重复图片。

[0092] 步骤 21 中，利用聚类算法按照一分组中各图片的颜色特征向量间的距离，将该分组中的图片分为多组，假设该分组为分组 A，其具体实现方法如下：

[0093] 步骤 S01：将分组 A 作为当前图片分组，将当前图片分组中图片的主颜色设置为图片签名树的子树的根节点，将该根节点作为当前父节点；

[0094] 步骤 S02：利用聚类算法按照当前图片分组中各图片的颜色特征向量间的 距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

[0095] 步骤 S03：对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 S04，否则，到步骤 S05；

[0096] 步骤 S04：将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 S02；

[0097] 步骤 S05：将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照分组 A 中各图片的颜色特征向量间的距离，将分组 A 中的图片分为多组后的一个分组。

[0098] 步骤 S02 中，聚类算法是一种将多个同类元素进行分组的算法，具体的，将给定的一个有 N 个元素的数据集分成 K 组，分组后每一个分组至少包含一个元素，且每一个元素属于且仅属于一个分组。对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法将上一次的分组继续进行分组，使得本次的分组结果较之前一次的分组结果好，好的标准是：同一分组中元素的距离越来越近，而不同分组中元素的距离越来越远。聚类算法有 K-MEANS 算法、GCS 算法等。本申请中是利用聚类算法将图片的颜色特征向量进行分组，两个颜色特征向量的距离是指该两个颜色特征向量的分量差的平方和，例如，向量 A 为 (a1，b1，c1)，向量 B 为 (a2，b2，c2)，则向量 A 和向量 B 的距离为 $(a1-a2)^2+(b1-b2)^2+(c1-c2)^2$。

[0099] 步骤 S03 中，分组停止条件可以有多种，举例说明，可以包括以下三种中的一种或任意组合：

[0100] 第一种，分组中包含的颜色特征向量的个数小于设定的向量数门限值，该向量数门限值为大于 1 的整数；

[0101] 第二种，分组中各颜色特征向量到该分组中各颜色特征向量的聚类中心的距离均小于设定的距离门限值，该距离门限值为不小于 0 的自然数；

[0102] 第三种，分组的分裂次数超过设定的分裂数门限值，分组的分裂次数是指从待分组图片到得到该分组的时间段内执行分组操作的次数。分裂数门限值的取值为不小于 1 的整数。

[0103] 步骤 S04 中，分组的聚类中心是根据该分组中所包含的所有颜色特征向量确定的另一颜色特征向量，该颜色特征向量到该分组中所包含的所有颜色特征向量的距离小于其它分组中任意一个颜色特征向量到该分组中所包含的所有颜色特征向量的距离。

[0104] 步骤 22 中，从当前各分组中查找所包含图片的主颜色与给定图片的主颜色相同、

并且与给定图片的颜色特征向量的距离最近的分组，其实现方法如下：

[0105] 步骤 S11：在上述建立的图片签名树中查找根节点为给定图片的主颜色的子树，将该子树的根节点作为当前父节点；

[0106] 步骤 S12：在所述子树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则到步骤 S13；若该子节点为叶子节点，则到步骤 S14；

[0107] 步骤 S13：判断该中间节点的颜色特征向量与给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 S12，否则，停止查找操作；

[0108] 步骤 S14：将该叶子节点所包含各图片构成的分组，确定为所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组。

[0109] 步骤 S13 中，设定条件可以为以下两种：

[0110] 第一种，该中间节点的颜色特征向量与给定图片的颜色特征向量的距离小于预先设定的距离阈值，该距离阈值的取值为大于 0 的自然数；

[0111] 第二种，该中间节点的颜色特征向量与给定图片的颜色特征向量的距离为查找到的各中间节点的颜色特征向量与给定图片的颜色特征向量的距离中的最小值。例如查找到当前父节点的 4 个中间节点，该 4 个中间节点的颜色特征 向量与给定图片的颜色特征向量的距离分别为 1、2、3 和 4，则确定距离为 1 的中间节点为符合设定条件的节点。

[0112] 步骤 23 中，在将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配之前，需要提取查找到的分组中所包含图片和给定图片中各图片的主颜色率和 / 或颜色特征向量等作为该图片的图像特征。近重复图片判定条件与步骤 12 中的近重复图片判定条件相同，这里不再赘述。

[0113] 下面以具体实例对上述方法进行说明：

[0114] 假设待分组图片包含 10 张图片，建立这 10 张图片的图片签名树即将这 10 张图片进行分组的流程如下：

[0115] 步骤 S21：将待分组图片分为 2 组，每组中包含 5 张图片，第 1 组所包含图片的主颜色均为红色，第 2 组所包含图片的主颜色均为蓝色；确定每组中各图片的颜色特征向量；

[0116] 步骤 S22：将步骤 S21 中分组后的第 1 组中图片的主颜色设置为图片签名树的一个子树的根节点，将该根节点作为当前父节点；利用聚类算法按照第 1 组中各图片的颜色特征向量间的距离，将第 1 组中各图片的颜色特征向量分为 2 组，分组后第 1 组中包含 2 个图片的颜色特征向量，第 2 组中包含 3 个图片的颜色特征向量；

[0117] 步骤 S23：对于步骤 S22 中分组后的第 1 组，由于该组满足设置的分组停止条件：向量的数目小于 3，将第 1 组包含的 2 个图片的主颜色率和颜色特征向量设置为当前父节点的子节点；对于步骤 S22 中分组后的第 2 组，该组不满足设置的分组停止条件，将第 2 组中 3 个颜色特征向量的聚类中心设置为当前父节点的子节点，将该子节点作为当前父节点；

[0118] 步骤 S24：对于步骤 S22 中分组后的第 2 组，利用聚类算法按照该组中各图片的颜色特征向量间的距离，将该组中各图片的颜色特征向量分为 2 组，分组后第 1 组中包含 1 个图片的颜色特征向量，第 2 组中包含 2 个图片的颜色特征向量；

[0119] 步骤 S25：对于步骤 S24 中分组后的 2 组，由于这 2 组均满足设置的分组停止条件：向量的数目小于 3，将第 1 组包含的 1 个图片的主颜色率和颜色特征向量设置为当前父节点

的一个子节点，将第 2 组包含的 2 个图片的主颜色率和颜色特征向量设置为当前父节点的另一个子节点；

[0120] 步骤 S26：将步骤 S21 中分组后的第 2 组中图片的主颜色设置为图片签名树的一个子树的根节点，将该根节点作为当前父节点；利用聚类算法按照第 2 组中各图片的颜色特征向量间的距离，将第 2 组中各图片的颜色特征向量分为 2 组，分组后第 1 组中包含 1 个图片的颜色特征向量，第 2 组中包含 4 个图片的颜色特征向量；

[0121] 步骤 S27：对于步骤 S26 中分组后的第 1 组，由于该组满足设置的分组停止条件：向量的数目小于 3，将第 1 组包含的 1 个图片的主颜色率和颜色特征向量设置为当前父节点的子节点；对于步骤 S26 中分组后的第 2 组，该组不满足设置的分组停止条件，将第 2 组中 4 个颜色特征向量的聚类中心设置为当前父节点的子节点，将该子节点作为当前父节点；

[0122] 步骤 S28：对于步骤 S26 中分组后的第 2 组，利用聚类算法按照该组中各图片的颜色特征向量间的距离，将该组中 4 个图片的颜色特征向量分为 2 组，分组后第 1 组中包含 2 个图片的颜色特征向量，第 2 组中也包含 2 个图片的颜色特征向量；

[0123] 步骤 S29：对于步骤 S28 中分组后的 2 组，由于这 2 组均满足设置的分组停止条件：向量的数目小于 3，将第 1 组包含的 2 个图片的主颜色率和颜色特征向量设置为当前父节点的一个子节点，将第 2 组包含的 2 个图片的主颜色率和颜色特征向量设置为当前父节点的另一个子节点。

[0124] 至此，10 个待分组图片的图片签名树建立完毕，如图 3 所示，由于该图片签名树具有 6 个叶子节点，因此，10 个待分组图片被分为 6 组。

[0125] 现需要从如图 3 所示的图片签名树中查找到一主颜色为红色的给定图片的近重复图片，具体实现流程如下：

[0126] 步骤 S31：在图片签名树中查找根节点为红色的子树，将查找到的子树的根节点作为当前父节点；

[0127] 步骤 S32：查找当前父节点的子节点，查找到 2 个子节点，其中一个字节点为中间节点，另一个子节点为叶子节点；

[0128] 步骤 S33：对于步骤 S32 中查找到的叶子节点，将该叶子节点中各图片的图像特征与给定图片的图像特征进行匹配，匹配发现给定图片的颜色特征向量与该叶子节点中一个图片的颜色特征向量的距离为 0，并且给定图片的主颜色率与叶子节点中该图片的主颜色率之差小于设定的主颜色率门限值，将叶子节点中的该图片确定为查找到给定图片的近重复图片；

[0129] 步骤 S34：对于步骤 S32 中查找到的中间节点，判断该中间节点的颜色特征向量与给定图片的颜色特征向量的距离不满足设定条件：该距离小于设定的距离门限值，不在继续查找。

[0130] 下面结合具体计算机应用对本实施例进行说明：

[0131] 步骤 b1：将数据库中存储的多张待分组图片读取到内存中；

[0132] 步骤 b2：对读取到内存中的每一张待分组图片，计算每一张待分组图片的图像特征，建立图片标识与图像特征的对应关系表 A，将对应关系表 A 存放在硬盘上；

[0133] 步骤 b3：根据图片的主颜色将读入内存的多张待分组图片划分为多个分组，同一分组中图片的主颜色相同；

[0134] 对于各个分组，根据步骤 S01 ～步骤 S05 的算法，建立图片签名树；将建立的图片签名树保存在硬盘上；

[0135] 步骤 b4：将给定图片、对应关系表 A 和图片签名树读入内存；

[0136] 从对应关系表 A 中查找给定图片的标识对应的图像特征；根据步骤 S11 ～步骤 S14 的算法，在硬盘上保存的图片签名树中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的叶子节点，将该叶子节点中的图片读入内存；

[0137] 步骤 b5：从对应关系表 A 中查找上一步骤从叶子节点读入内存的各个图片的标识对应的图像特征，将给定图片的图像特征分别与查找到的各个图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图像特征对应的图片，确定为给定图片的近重复图片。

[0138] 实施例三：

[0139] 本实施例中，根据图片的颜色特征向量将多张待分组图片进行分组，参见图 4，具体包括以下步骤：

[0140] 步骤 40：确定多张待分组图片中各待分组图片的颜色特征向量；

[0141] 步骤 41：利用聚类算法按照多张待分组图片的颜色特征向量间的距离，将多张待分组图片划分为多个分组；

[0142] 步骤 42：从多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

[0143] 步骤 43：将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0144] 步骤 41 中，利用聚类算法按照多张待分组图片的颜色特征向量间的距离，将多张待分组图片划分为多个分组，其实现方法可以如下：

[0145] 步骤 S41：设置图片签名树的根节点，并将该根节点作为当前父节点；将多张待分组图片构成的分组作为当前图片分组；

[0146] 步骤 S42：利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

[0147] 步骤 S43：对于 K 组中的每一组，若该组不满足设定的分组停止条件，则到步骤 S44，否则，到步骤 S45；

[0148] 步骤 S44：将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，返回步骤 S42；

[0149] 步骤 S45：将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照多张待分组图片的颜色特征向量间的距离，将多张待分组图片划分为多个分组后的一个分组。

[0150] 步骤 S43 中，分组停止条件与步骤 S03 中的分组停止条件相同，这里不再赘述。

[0151] 步骤 42 中，从多个分组中查找与给定图片的颜色特征向量的距离最近的分组，其实现方法如下：

[0152] 步骤 S51：将建立的图片签名树的根节点作为当前父节点；

[0153] 步骤 S52：在图片签名树中查找当前父节点的子节点，对于查找到的各子节点，若

该子节点为中间节点，则到步骤 S53 ；若该子节点为叶子节点，则到步骤 S54 ；

[0154] 步骤 S53 ：判断该中间节点的颜色特征向量与给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，返回步骤 S52，否则，停止查找操作 ；

[0155] 步骤 S54 ：将该叶子节点所包含各图片构成的分组，确定为与给定图片的颜色特征向量的距离最近的分组。

[0156] 步骤 S53 中，设定条件与步骤 S13 中的设定条件相同，这里不再赘述。

[0157] 步骤 43 中，在将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配之前，需要提取查找到的分组中所包含图片和给定图片中各图片的主颜色率和 / 或颜色特征向量等作为该图片的图像特征。近重复图片判定条件与步骤 12 中的近重复图片判定条件相同，这里不再赘述。

[0158] 下面结合具体计算机应用对本实施例进行说明 ：

[0159] 步骤 c1 ：将数据库中存储的多张待分组图片读取到内存中 ；

[0160] 步骤 b2 ：对读取到内存中的每一张待分组图片，计算每一张待分组图片的图像特征，建立图片标识与图像特征的对应关系表 A，将对应关系表 A 存放在 硬盘上 ；

[0161] 步骤 b3 ：根据步骤 S41 ～步骤 S45 的算法，建立图片签名树 ；将建立的图片签名树保存在硬盘上 ；

[0162] 步骤 b4 ：将给定图片、对应关系表 A 和图片签名树读入内存 ；

[0163] 从对应关系表 A 中查找给定图片的标识对应的图像特征 ；根据步骤 S51 ～步骤 S54 的算法，在硬盘上保存的图片签名树中查找与给定图片的颜色特征向量的距离最近的叶子节点，将该叶子节点中的图片读入内存 ；

[0164] 步骤 b5 ：从对应关系表 A 中查找上一步骤读入内存的各个图片的标识对应的图像特征，将给定图片的图像特征分别与查找到的各个图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图像特征对应的图片，确定为给定图片的近重复图片。

[0165] 下面举例说明本申请的具体应用场景 ：

[0166] 步骤 c1 ：用户将输入的给定图片或通过互联网搜索到的给定图片提交给客户端的近重复图片查找系统 ；

[0167] 步骤 c2 ：近重复图片查找系统按照实施例一～实施例三中的方法连入互联网的服务器上搜索给定图片的近重复图片，连入互联网的服务器中保存有按照实施例一～实施例三中的方法建立的图片签名树 ；

[0168] 步骤 c3 ：近重复图片查找系统将搜索到的近重复图片返回并展现在用户所在的客户端上。

[0169] 参见图 5，本申请实施例提供一种图片查找装置，该装置包括 ：

[0170] 分组单元 50，用于将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同 ；

[0171] 查找单元 51，用于从所述多个分组中查找所包含图片的主颜色与给定图片的主颜色相同的分组 ；

[0172] 匹配单元 52，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片 判定条件的图片

确定为所述给定图片的近重复图片。

[0173] 所述近重复图片判定条件包括：

[0174] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为0，并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

[0175] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值、以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

[0176] 参见图6，本申请实施例还提供一种图片查找装置，该装置包括：

[0177] 第一分组单元60，用于将多张待分组图片划分为多个分组，划分为多个分组后同一分组中图片的主颜色相同；

[0178] 向量确定单元61，用于对于所述多个分组中的各分组，确定该分组中图片的颜色特征向量；

[0179] 第二分组单元62，用于对于所述多个分组中的各分组，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片划分为多个分组；

[0180] 查找单元63，用于从所述第二分组单元分组后的各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组；

[0181] 匹配单元64，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0182] 所述第二分组单元62包括：

[0183] 子树建立单元，用于对于所述多个分组中的各分组，将该分组作为当前图片分组及第一分组，将当前图片分组中图片的主颜色设置为图片签名树的子树 的根节点，将该根节点作为当前父节点，触发聚类分组单元；

[0184] 聚类分组单元，用于利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为K组，K为大于1的整数，触发递归建立单元；

[0185] 递归建立单元，用于对于K组中的每一组，判断该组是否满足设定的分组停止条件，若是，则触发叶子节点建立单元；否则，触发中间节点建立单元；

[0186] 中间节点建立单元，用于将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，触发聚类分组单元；

[0187] 叶子节点建立单元，用于将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照第一分组中各图片的颜色特征向量间的距离，将第一分组中的图片划分为多个分组后的一个分组。

[0188] 所述分组停止条件包括以下三种中的一种或任意组合：

[0189] 分组中包含的颜色特征向量的个数小于设定的向量数门限值；

[0190] 分组中各颜色特征向量到该分组中各颜色特征向量的聚类中心的距离均小于设

定的距离门限值；

[0191] 分组的分裂次数超过设定的分裂数门限值，所述分组的分裂次数是从所述待分组图片到得到该分组的时间段内执行分组操作的次数。

[0192] 所述查找单元 63 包括：

[0193] 第一查找单元，用于在所述图片签名树中查找根节点为给定图片的主颜色的子树，将所述子树的根节点作为当前父节点；

[0194] 第二查找单元，用于在所述子树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则触发中间节点处理单元；若该子节点为叶子节点，则触发叶子节点处理单元；

[0195] 中间节点处理单元，用于判断该中间节点的颜色特征向量与所述给定图片 的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，触发第二查找单元，否则，停止查找操作；

[0196] 叶子节点处理单元，用于将该叶子节点所包含各图片构成的分组，确定为所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组。

[0197] 所述设定条件为：所述距离小于预先设定的距离阈值；或者，所述距离为查找到的各中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离中的最小值。

[0198] 所述近重复图片判定条件包括：

[0199] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，以及所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

[0200] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值，以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

[0201] 参见图 7，本申请实施例还提供一种图片查找装置，该装置包括：

[0202] 向量确定单元 70，用于确定多张待分组图片中各待分组图片的颜色特征向量；

[0203] 分组单元 71，用于利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组；

[0204] 查找单元 72，用于从所述多个分组中查找与给定图片的颜色特征向量的距离最近的分组；

[0205] 匹配单元 73，用于将所述查找单元查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片。

[0206] 所述分组单元 71 包括：

[0207] 初始化单元，用于设置图片签名树的根节点，并将该根节点作为当前父节点；将所述多张待分组图片构成的分组作为当前图片分组；

[0208] 聚类分组单元，用于利用聚类算法按照当前图片分组中各图片的颜色特征向量间的距离，将当前图片分组中各图片的颜色特征向量分为 K 组，K 为大于 1 的整数；

[0209] 递归建立单元，用于对于 K 组中的每一组，判断该组是否满足设定的分组停止条件，若是，则触发叶子节点建立单元；否则，触发中间节点建立单元；

[0210] 中间节点建立单元，用于将该组中各颜色特征向量的聚类中心设置为当前父节点的子节点，并将该分组作为当前图片分组，将所述子节点作为当前父节点，触发聚类分组单元；

[0211] 叶子节点建立单元，用于将该组中各颜色特征向量对应的图片设置为当前父节点的子节点，并将该子节点所包含各图片构成的分组，确定为利用聚类算法按照所述多张待分组图片的颜色特征向量间的距离，将所述多张待分组图片划分为多个分组后的一个分组。

[0212] 分组停止条件包括以下三种中的一种或任意组合：

[0213] 分组中包含的颜色特征向量的个数小于设定的向量数门限值；

[0214] 分组中各颜色特征向量到该分组中各颜色特征向量的聚类中心的距离均小于设定的距离门限值；

[0215] 分组的分裂次数超过设定的分裂数门限值，所述分组的分裂次数是从所述待分组图片到得到该分组的时间段内执行分组操作的次数。

[0216] 所述查找单元 72 包括：

[0217] 第一查找单元，用于将所述图片签名树的根节点作为当前父节点，在所述图片签名树中查找当前父节点的子节点，对于查找到的各子节点，若该子节点为中间节点，则触发中间节点处理单元；若该子节点为叶子节点，则触发叶子节点处理单元；

[0218] 中间节点处理单元，用于判断该中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离是否满足设定条件，若是，则将该中间节点作为当前父节点，触发第一查找单元，否则，停止查找操作；

[0219] 叶子节点处理单元，用于将该叶子节点所包含各图片构成的分组，确定为与给定图片的颜色特征向量的距离最近的分组。

[0220] 所述设定条件为：所述距离小于预先设定的距离阈值；或者，所述距离为查找到的各中间节点的颜色特征向量与所述给定图片的颜色特征向量的距离中的最小值。

[0221] 所述近重复图片判定条件包括：

[0222] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离为 0，以及所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的主颜色率门限值；或者，

[0223] 所述给定图片的颜色特征向量与查找到的分组中图片的颜色特征向量的距离小于设定的距离门限值，以及所述给定图片和查找到的分组中图片的主颜色率均小于设定的第一主颜色率门限值、并且所述给定图片的主颜色率与查找到的分组中图片的主颜色率之差小于设定的第二主颜色率门限值。

[0224] 综上，本申请的有益效果包括：

[0225] 本申请实施例提供的方案中，首先根据图片的主颜色和 / 或颜色特征向量将多张待分组图片进行分组，然后，在各分组中查找所包含图片的主颜色与给定图片的主颜色相同和 / 或与给定图片的颜色特征向量的距离最近的分组，最后，只将查找到的分组中所包含的各图片与给定图片进行图像特征的匹配，将匹配结果满足设定的近重复图片判定条件

的图片确定为给定图片的近重复图片。由于首先将多张待分组图片分组，在各分组中查找到满足一定条件的分组后，只将满足条件的分组中的图片与给定图片进行匹配，从而确定给定图片的近重复图片，而不是将所有待分组图片均与给定图片进行匹配来确定给定图片的近重复图片，能够有效地提高查找给定图片的近重复图片的效率。

[0226] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

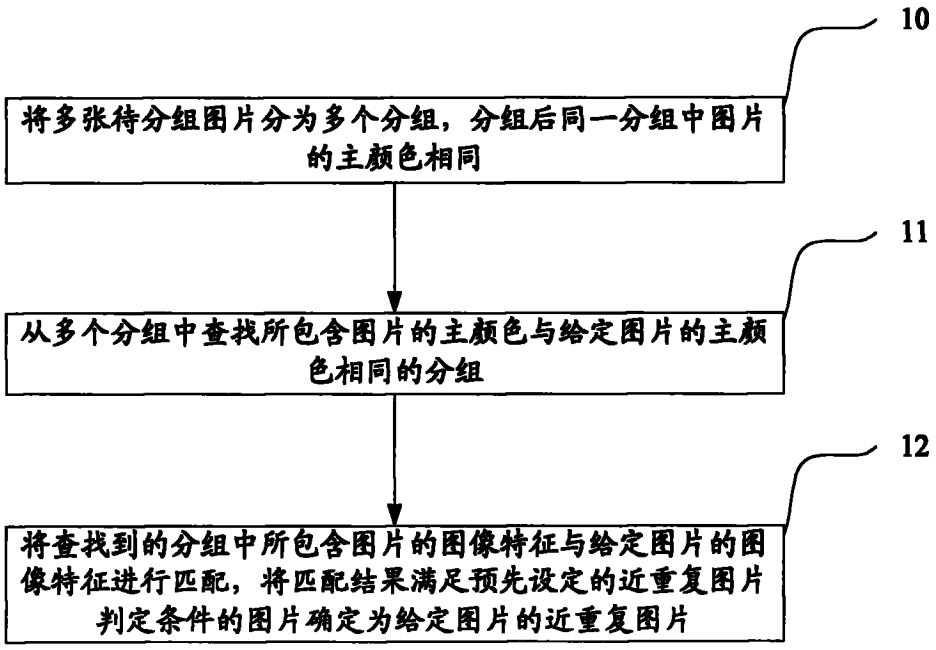

10
将多张待分组图片分为多个分组，分组后同一分组中图片的主颜色相同
11
从多个分组中查找所包含图片的主颜色与给定图片的主颜色相同的分组
12
将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为给定图片的近重复图片

图 1

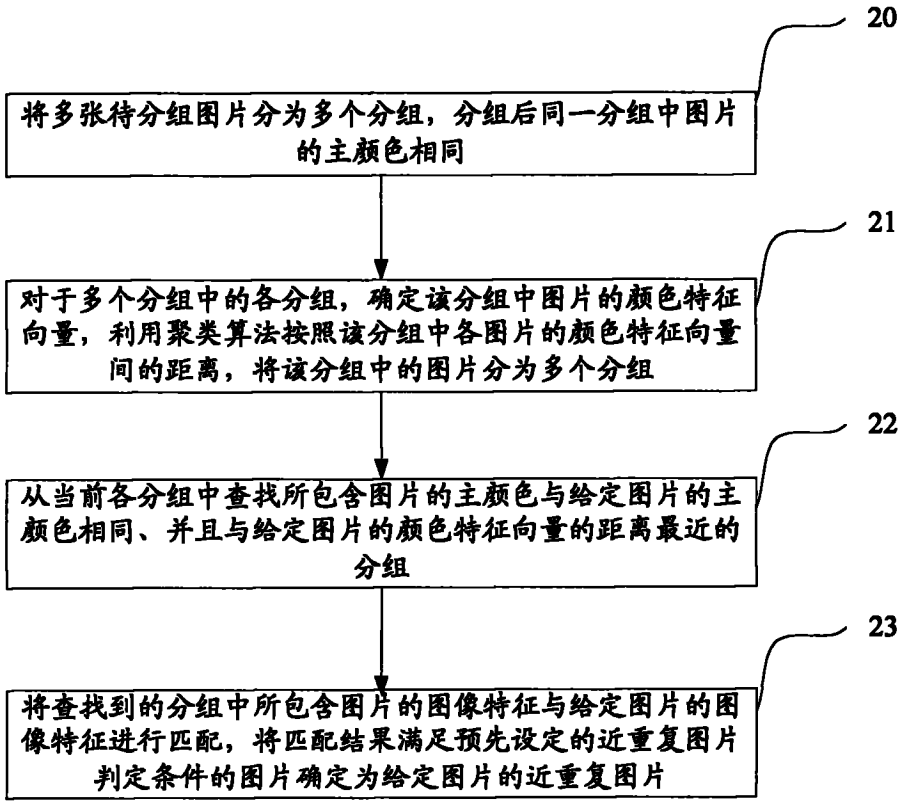

20
将多张待分组图片分为多个分组，分组后同一分组中图片的主颜色相同
21
对于多个分组中的各分组，确定该分组中图片的颜色特征向量，利用聚类算法按照该分组中各图片的颜色特征向量间的距离，将该分组中的图片分为多个分组
22
从当前各分组中查找所包含图片的主颜色与给定图片的主颜色相同、并且与给定图片的颜色特征向量的距离最近的分组
23
将查找到的分组中所包含图片的图像特征与给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为给定图片的近重复图片

图 2

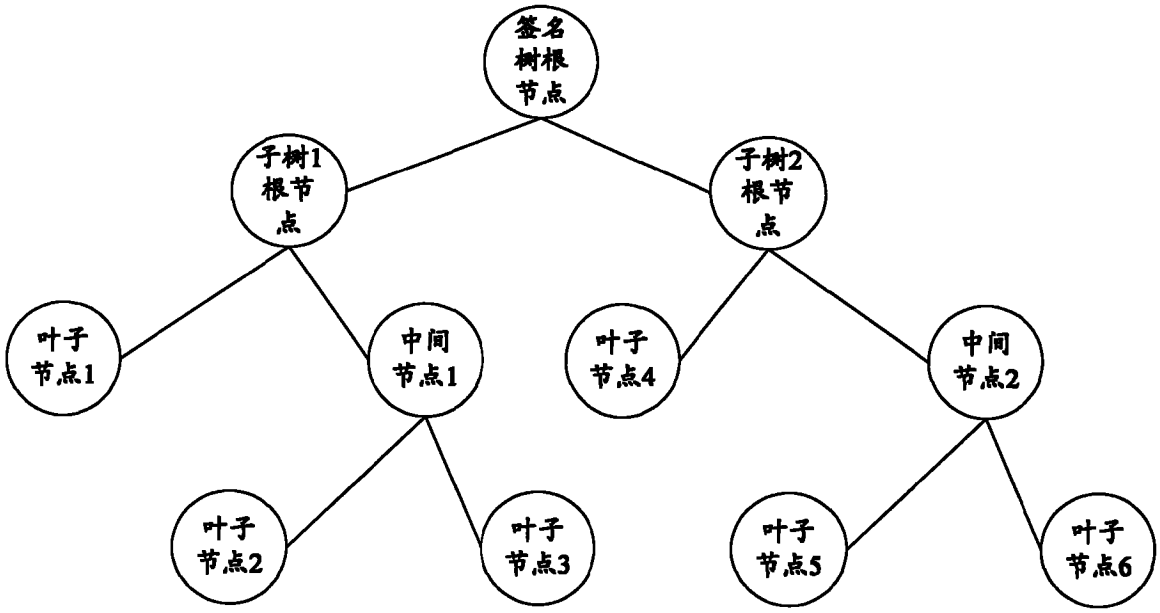

签名树根节点
子树1根节点
子树2根节点
叶子节点1
中间节点1
叶子节点4
中间节点2
叶子节点2
叶子节点3
叶子节点5
叶子节点6

图 3

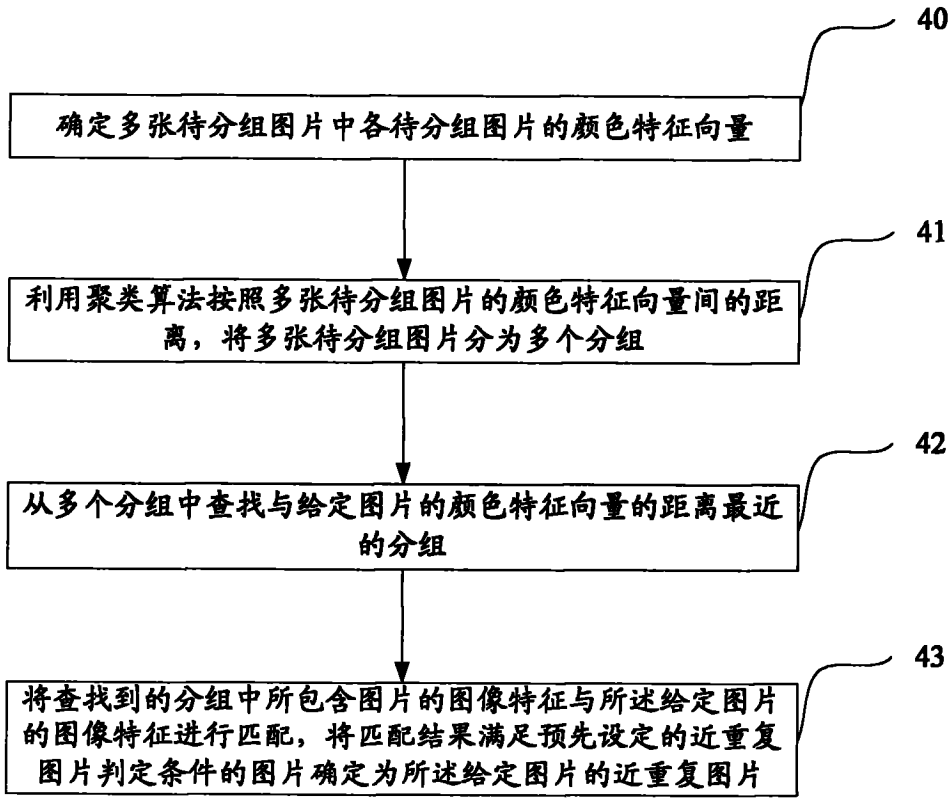

40
确定多张待分组图片中各待分组图片的颜色特征向量
41
利用聚类算法按照多张待分组图片的颜色特征向量间的距离，将多张待分组图片分为多个分组
42
从多个分组中查找与给定图片的颜色特征向量的距离最近的分组
43
将查找到的分组中所包含图片的图像特征与所述给定图片的图像特征进行匹配，将匹配结果满足预先设定的近重复图片判定条件的图片确定为所述给定图片的近重复图片

图 4

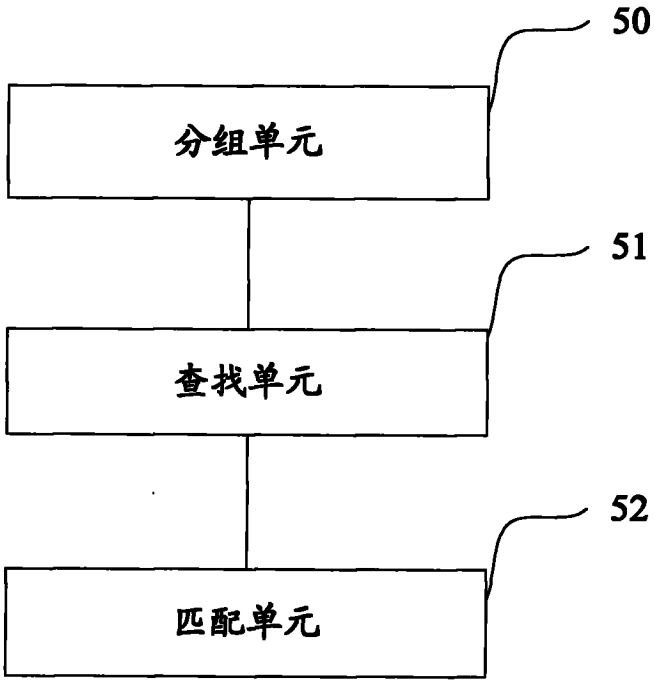


图 5

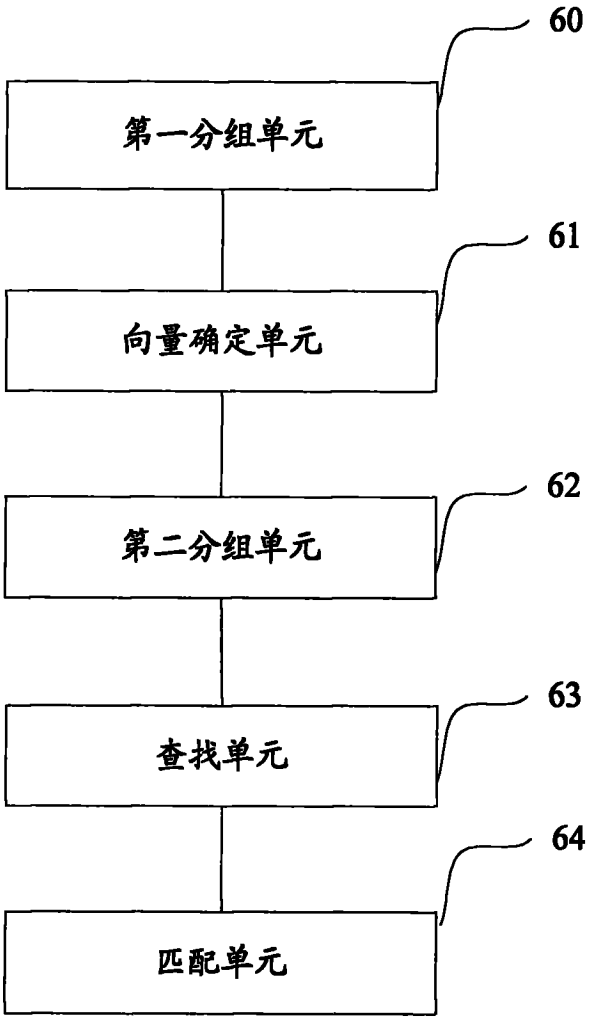


图 6

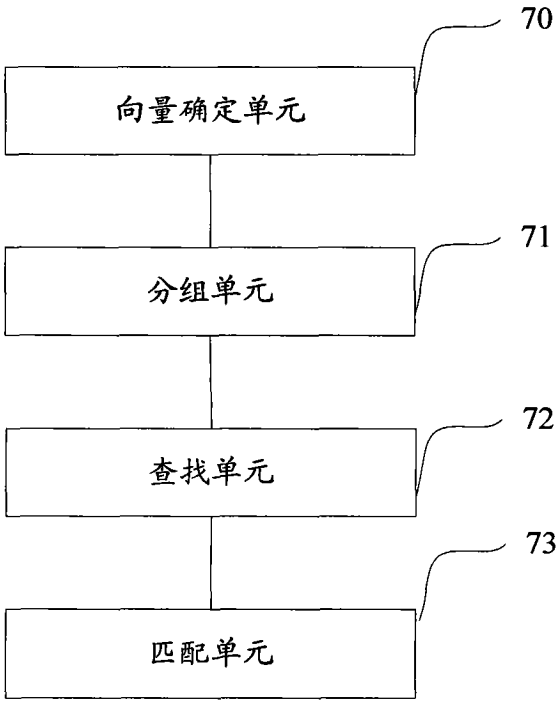


图 7