(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2018 年 6 月 14 日 (14.06.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/103179 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2017/070197

(22) 国际申请日: 2017 年 1 月 5 日 (05.01.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201611130891.8 2016年12月9日 (09.12.2016) CN

(71) 申请人: 西 北 大 学 (NORTHWESTERN UNIVERSITY) [CN/CN]; 中国陕西省西安市太白北路西北大学39号信箱, Shaanxi 710069 (CN)。

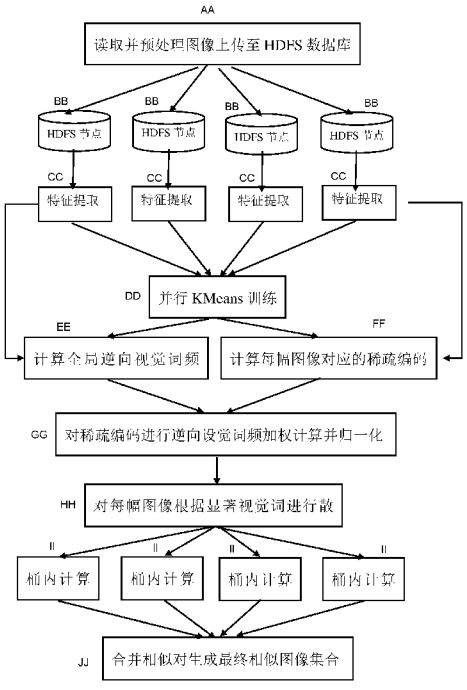
(72) 发明人: 赵 万 青 (ZHAO, Wanqing); 中 国 陕 西 省 西 安 市 太 白 北 路 西 北 大 学 39 号 信 箱, Shaanxi 710069 (CN)。 罗 远 哉 (LUO, Hangzai); 中 国 陕 西 省 西 安 市 太 白 北 路 西 北 大 学 39 号 信 箱, Shaanxi 710069 (CN)。 范 建 平 (FAN, Jianping); 中 国 陕 西 省 西 安 市 太 白 北 路 西 北 大 学 39 号 信 箱, Shaanxi 710069 (CN)。 彭 进 业 (PENG, Jinye); 中 国 陕 西 省 西 安 市 太 白 北 路 西 北 大 学39号信箱, Shaanxi 710069 (CN)。

(74) 代理人: 西安恒泰知识产权代理事务所等 (HENG TAI INTELLECTUAL PROPERTY AGENCY et al.); 中国陕西省西安市莲湖区唐延路北段22号金辉国际广场12-05室, Shaanxi 710068 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,

(54) **Title:** NEAR-DUPLICATE IMAGE DETECTION METHOD BASED ON SPARSE REPRESENTATION

(54) 发明名称: 一种基于稀疏表示的近似重复图像检测方法



AA Acquire and preprocess images, and upload the same to an HDFS database
BB HDFS node
CC Feature extraction
DD Parallel K-Means training
EE Compute a global inverse visual word frequency
FF Compute sparse encoding results of respective images
GG Perform a weighting operation on the sparse encoding results based on the inverse visual word frequency, and perform normalization
HH Perform hashing on the images according to salient visual words
II Within-bucket computation
JJ Combine near-duplicate pairs, and generate a final near-duplicate image set

图 2

(57) **Abstract:** A near-duplicate image detection method based on sparse representation. The method is proposed on the basis of a Hadoop distributed computing framework, and comprises the following steps: acquiring an image set *I*, wherein sparse encoding results of all images are *g'*; extracting non-zero elements in *g'*, and hashing the sparse encoding result *gi'* of the image Ii to groups corresponding to subscripts of the non-zero elements; computing, for each reduce function, a similarity level *Y* of the sparse encoding results for each pair of images <*Iw*, *Iz*>,

WO 2018/103179 A1

[见续页]

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

and if *Y* is greater than 0.7, then outputting said near-duplicate image pair <*Iw*, *Iz*>; and combining near-duplicate image pairs having the image *Iw*, and generating a near-duplicate image subset. The technical solution of the present invention employs parallel computing to greatly improve computational efficiency of a K-Means clustering algorithm for a large-scale data set, and introduces the sparse representation concept to increase the speed of the method and eliminate excessive computation for solution finding and optimization.

(57) 摘要：一种基于稀疏表示的近似重复图像检测方法，该方法基于hadoop分布式计算框架提出，该检测方法包括如下步骤，先获取图像集I，其中所有图像的稀疏编码为g'；提取g'中非零元素，将图像I_i的稀疏编码gi'散列到非零元素的下标对应的组中，计算每个Reduce函数中每对图像$<I_w, I_z>$稀疏编码的相似度Y，若Y大于0.7，则输出相似图像对$<I_w, I_z>$；将具有图像I_w的相似图像对合并，生成相似图像子集。上述技术方案通过并行化的计算方式大大提高了针对大规模数据集的KMeans聚类算法的计算效率，并引入稀疏表示理论，具有更快的实现方法，不需要过多的求解优化过程。

一种基于稀疏表示的近似重复图像检测方法

技术领域

本发明属于图像近似重复检测领域，涉及一种基于稀疏表示的并行化的图像近似重复检测方法，可以高效并准确的对海量图像集提取近似重复图像集合。

背景技术

随着移动互联网和数码相机的发展，人们越来越多的将拍摄的多媒体数据分享到互联网上，由于拍摄者的位置、拍摄的对象、角度的相同，从而导致了互联网上出现了大量的相似的图片。通过提取这些相似图像集不仅可以对图像检索结果进行去重过滤，同时在许多图像处理领域如图像聚类、图像识别、图像分类等也是重要一步。

通常近似重复图像是由某幅原图像通过某些近似重复图像变换得到的，一般可以产生近似重复图像的变换包括平移、缩放、选择、图像色调的变化、添加文字、格式变化、分辨率变化等等。而近重复图像检测是指给定查询图像，在数据集中找到与此图像的近重复图像，或是提取出数据集中所有近重复图像子集。目前，大多数的近重复图像检测是采用 Bag-of-words 和 LSH 方法构建系统。Bag-of-words 模型是将每幅图像的局部特征利用训练好的字典映射为一个视觉词频直方图向量。基于 Bag-of-words 的图像表示模型方法一般包括 3 部分：1）提取图像的局部特征；2）通过聚类图像集的局部特征，构建视觉字典；3）映射每幅图的局部向量为一个词频直方图。LSH（Locality-Sensitive Hashing）是一种对高维数据建立索引的随机方法，以一定的查找准确率为代价，在高维数据空间中进行近似线性的查找，返回查询数据的近似最近邻数据。它的基本

思想是通过一组哈希函数将输入数据点映射到各个桶中，并保证近邻的数据点以较大的概率映射到同一个桶中，相距较远的数据点以较小的概率映射到同一个桶中，这样一个查询数据点所在桶中的其它数据点就可以被看做是这个查询数据的近邻点。然而由于 BOW 模型对于局部特征过于严格的界定和 LSH 以准确度换效率的特性往往导致检测的结果无法让人满意。此外，已有的近重复图像检测系统一般是采用单节点进行运算，随着数据量爆炸式的增长，单节点系统已经远远不能满足目前的应用需要。因此，多节点的并行计算就成为了必然选择，在众多的分布式框架中，以 HADOOP 系统最为稳定和高效。

发明内容

针对上述现有技术中存在的问题，本发明的目的在于，提出一种针对大规模图像集的基于稀疏表示的分布式图像近似重复集提取系统及方法。该方法不仅能够提升处理大规模图像集的效率，并且与传统方法相比可以有效的提高检测结果的正确率。

为了实现上述任务，本发明采用以下技术方案：

一种基于稀疏表示的近似重复图像检测方法，该方法基于 hadoop 分布式计算框架提出，该检测方法包括如下步骤，获取图像集 I 中所有图像的 IDF 加权稀疏编码 g′，其中 $I=(I_1,I_2,...,I_i,...,I_w,...,I_z,...,I_R)$，$I_i$ 的 IDF 加权稀疏编码为 g_i'，$g_i' \in g'$, i 为大于等于 1 的自然数，w 为大于 i 的自然数, z 为大于 w 的自然数, R 为大于 z 的自然数, 其特征在于，方法还包括：

（1）提取图像 I_i 的 IDF 加权稀疏编码 g_i' 中的非零元素； $g_{ik}' \in g_i'$，k 为大于等于 1 的自然数，g_i' 内的非零元素为 $(g_{iu}', \ldots, g_{iv}')$，设非零元素为 m 个，m 为大于等于 1 的自然数，$m \leqslant k$, $g_{iu}' \neq 0$, $g_{iv}' \neq 0$, u 为大于等于 1 的自然数，v 大于

等于 1 的自然数，k>v>u；

（2）建立 k 个组，分别命名为：$\begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$；其中，$\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$为空矩阵；

（3）利用（式 1）的矩阵变换，将图像 I_i 的 IDF 加权稀疏编码 g_i' 分别散列到非零元素的下标(u, . . . , v)对应的 m 个组里；

$$g'=\begin{bmatrix} g'_{11} & g'_{12} & \cdots & g'_{1k}\\ g'_{21} & g'_{22} & \cdots & g'_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ g'_{i1} & g'_{i2} & \cdots & g'_{ik}\\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \Rightarrow \begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots\\ \ddots\\ \cdots\ g'_i\\ \ddots\\ \cdots\ g'_i\\ \ddots\\ \cdots\end{bmatrix} \qquad \text{（式 1）}$$

（4）利用 $Y=J(g_i',g_j')=\dfrac{|g_i'\cap g_j'|}{|g_i'\cup g_j'|}$ 计算步骤（3）所得 m 组中的每个组中每对图像<I_i,I_j>IDF 加权稀疏编码的相似度 Y，若 Y 大于 0.7，则图像<I_i,I_j>为相似图像对；

其中，j 为大于等于 1 的自然数，且 i≠j；g_i'和g_j'分别表示图像 I_i 和 I_j 的 IDF 加权稀疏编码；

（5）将步骤（4）所得结果中具有相同图像的相似图像对合并，生成相似图像子集。

进一步地，所述获取图像集 I 中所有图像的 IDF 加权稀疏编码 g′，包括以下步骤：

并行化提取每副图像的局部特征，得到图像集 I 中所有图像的局部特征 S；

提取图像聚类中心，得到特征字典 E；

计算 E 中每个聚类中心权重；

根据 E 中每个聚类中心权重，提取图像的 IDF 加权稀疏编码 g′。

进一步地，所述并行化提取图像局部特征为提取图像集 I 所有图像的 SIFT 特征。

进一步地，所述提取所有图像的 SIFT 特征，具体步骤为：

将图像集 I 中每副图像 I_i 的大小标准化并灰度处理，得到标准大小的灰度图像集；其中，$I=(I_1,I_2,...,I_i,...,I_w,\ldots,I_z,\ldots,I_R)$；

将标准大小的灰度图像集分割给各集群结点，并行化提取每个图像的 SIFT 特征，所有图像的 SIFT 特征表示为 S，其中 $S_a \in S$，S_a 代表向量,a 为大于等于 1 的自然数，第 I_i 个图像的 SIFT 特征为 F_i，其中 $F_{ib} \in F_i$，F_{ib} 代表向量，b 为大于等于 1 的自然数。

进一步地，所述提取图像聚类中心，具体步骤为：

(11)计算 S_a 到聚类中心 A^d 的欧式距离，将 S_a 作为 value，将到 S_a 欧氏距离最近的聚类中心作为 key 值；$A_{dk} \in A^d$，d 为大于等于 0 的非负整数，k 为大于等于 1 的自然数，A_{dk} 代表向量；从 S 中随机选取额的 k 个 SIFT 特征作为初始 k 个聚类中心,形成初始聚类中心集合 A^0，$A_{0k} \in A^0$；

(21)求取 key 值相同的 S_a 的平均值，d=d+1，将各平均值作为新的聚类中心 A^d；

(31)计算新的聚类中心 A^d 与上一次循环的聚类中心 A^{d-1} 的欧氏距离均值，若该欧氏距离均值＞0.05，则跳回至步骤(11)执行；若该欧氏距离均值＜0.05，则新的聚类中心 A^d 作为图像集 I 的特征字典 E 输出，其中 $E_k \in E$，k 为大于等于 1 的自然数，E_k 为向量，所述特征字典 E 即为图像聚类中心。

进一步地，所述计算 E 中每个聚类中心权重，具体步骤为：采用 $idf_k = \log \frac{D}{\sum_a S_a \in E_k}$ 计算 E 中每个聚类中心权重，其中：D 为 S 中所有 SIFT 特征的总数，D 为大于等于 1 的自然数，$\sum_a S_a \in E_k$ 表示归属于 E_k 中心的所有 SIFT 特征总数。

进一步地，所述提取 IDF 加权图像稀疏编码，具体步骤为：

分别计算图像 I_i 中每个特征向量 F_{ib} 与特征字典 E 的欧氏距离 h，其中 $E_k \in E$，$h_k \in h=（h_1,h_2,...,）$，k 为大于等于 1 的自然数，从 E 中选取 h_k 最小的 m 个 E_k，组成特征字典 E′，E′=（E_f，...，E_g），其中 E′ 中有 m 个向量，f 为大于等于 1 的自然数，g 为大于等于 1 的自然数，g＞f；

利用 $C_{ib} = (E' - 1F_{ib}^{\ T})(E' - 1F_{ib}^{\ T})^T$ 计算 F_{ib} 与特征字典 E′ 的平方差矩阵 C_{ib}；

计算 F_{ib} 在特征字典 E′ 中的稀疏编码 c_{ib}，$c_{ib} = \tilde{c}_{ib} / 1^T \tilde{c}_{ib}$，$\tilde{c}_{ib} = (C_{ib} + diag(h'))\backslash 1$，其中 $h_m' \in h' =（h_1',h_2',...,）$ 为特征向量 F_{ib} 分别到 E′ 中视觉词的欧氏距离，$\mathrm{diag}(h')$ 表示将向量 h′ 的元素作为矩阵的主对角线；

提取 c_{ib} 中的 k 个最大值，得到图像 I_i 的图像稀疏编码 g_i，其中 $g_{ik} \in g_i$；

根据 $g_i' = g_i \cdot idf / \sum_k g_{ik} \cdot idf_k$ 对图像稀疏编码 g_i 进行逆向视觉词频加权归一化，得到 IDF 加权稀疏编码 g′。

本发明的有益效果是：

（1）本发明通过 MapReduce 并行方式运行 KMeans 算法在线学习过完备字典特征，使其尽可能地提取全局特征结构，从而可以从过完备字典中找到具有最具表征能力的原子组合来表示图像特征。在线字典学习通过每次迭代 MapReduce Jop，并根据设定阈值判断终止，这种并行化的计算方式大大提高了针对大规模数据集 KMeans 聚类算法的计算效率；

（2）本发明引入稀疏表示理论，并通过查找字典集中与表征特征最近邻的 K 个码字进行重构，这种重构方式跟原始 BOW 的向量量化相比，用多个码字重构能够具有更小的重构误差，通过选取局部近邻进行重构更能达到局部平滑稀疏性同时这种方式与传统稀疏表示方法具有更快的实现方法不需要过多的求解优化过程；

（3）本发明通过统计全局逆向视觉词频率 IDF，对图像稀疏表示向量进行加权，使其表征性更强，同时降低稀疏编码中表征性较低的码字的权重，升高表征性较高的码字权重，使其稀疏性更强，从而使相似图像的高表征特征具有更高的概率相同或相似；

（4）本发明通过提取稀疏表示中高表征特征进行散列到候选相似集合，同时在每个散列桶中通过计算两两 Jaccard index 相似度进行匹配，该方法可以简单、高效的得到相似图像对，同时此方法可以充分利用稀疏表示的特点和 Mapreduce 模型进行并行计算，大幅度提升大规模图像数据集的匹配效率和准确度；

（5）本发明基于 HADOOP 的分布式存储系统 HDFS 和并行计算模型 MapReduce 的分布式图像近重复检测系统，已有的图像近重复检测系统一般是只支持单节点的检测框架，但随着移动互联网的发展，图像数据成指数级的增长，以往的

系统根本不能满足如此大的数据量上的存储和运算。通过本发明不但具有针对海量数据的高扩展性，同时可以大幅度提高近重复检测的计算效率。

附图说明

图 1 所示为本发明基于稀疏表示的近似重复图像检测方法结构示意图；

图 2 为本发明实现基于 MapReduce 的并行 KMeans 算法流程图；

图 3 为本发明实现基于局部优先搜索的图像稀疏编码算法流程图；

图 4 为本发明实现基于 MapReduce 和显著维度特征的并行相似集合检测算法流程图；

图 5 为实施例 2 中分别对 5 个关键字从不同方法得到的聚类结果；

图 6 为实施例 2 的聚类结果。

具体实施方式

以下给出本发明的具体实施例，需要说明的是本发明并不局限于以下具体实施例，凡在本申请技术方案基础上做的等同变换均落入本发明的保护范围。

实施例 1

本方法基于 hadoop 分布式计算框架提出，图 1 中，一种基于稀疏表示的近似重复图像检测方法，其包括特征提取模块、特征字典构造模块，图像稀疏编码表示模块，相似图匹配过滤模块、近重复图像对合并模块。特征提取模块用于并行化提取图像集合中的所有原始局部图像特征；特征字典构造模块用于将特征提取模块提取到的图像特征采用并行 K-Means 在线字典学习算法构建原始特征字典集。所诉图像稀疏编码表示模块用于将每幅图像的原始局部特征利用所构造的字典集和稀疏编码方法映射为一个稀疏向量来表示每幅图像；所诉相似图匹配过滤模块用于并行化计算过滤后的图像对的相似度并输出相似度大于

某一阈值的图像对；所诉近重复图像对合并模块用于将相似图匹配过滤模块所输出的所有相似图像对进行合并成近重复图像集合。

该检测方法包括如下步骤：

步骤 1，将图像集 I，其中 $I=(I_1,I_2,...,I_i,...,I_w,\ldots,I_z,\ldots,I_R)$中每副图像 I_i 的大小标准化，比如设置成标准化（128*128、256*256 或者 512*512）大小，本发明中选用的图像统一标准化为大小为 256*256，之后进行灰度处理，得到标准大小的灰度图像集，利用自定义 imageWritable 将图像进行序列化操作，输出图像的二进制数据流，并将所有图像以键值对形式<key:图像 ID,value:imageWritable>压缩存储于自定义 ImageBundle，最终上传至分布式存储框架 HDFS 上；

步骤 2，从 HDFS 上下载 ImageBundle 文件，hadoop 会自行根据集群内节点个数分割 ImageBundle 文件给不同节点的 Map 函数，Map 函数获取图像的键值对形式，基于公开的 SIFT 算法提取出图像集中每个图像的多个 SIFT 特征向量，组成图像的 SIFT 特征向量序列，所有图像的 SIFT 特征表示为 S，其中 $S_a \in S$，S_a 代表向量,a 为大于等于 1 的自然数，第 I_i 个图像的 SIFT 特征为 F_i，其中 $F_{ib} \in F_i$, F_{ib} 代表向量，b 为大于等于 1 的自然数，并将图像集中每个图像分别以<key:图像 ID，value：Fi> 键值对形式存储在 HDFS 上；

步骤 3，计算 S_a 到聚类中心 A^d 的欧式距离，将 S_a 作为 value，将到 S_a 欧氏距离最近的聚类中心 A_q 作为 key 值；$A_{dk} \in A^d$，d 为大于等于 0 的非负整数，k 为大于等于 1 的自然数，A_{dk} 代表向量；从 S 中随机选取额的 k 个 SIFT 特征作为初始 k 个聚类中心,形成初始聚类中心集合 A^0，$A_{0k} \in A^0$；以键值对<key:A_q,value:S_a>形式输出。

步骤 4，求取 key 值相同的 S_a 的平均值，d=d+1，将各平均值作为新的聚类

中心 A^d；

步骤 5，计算新的聚类中心 A^d 与上一次循环的聚类中心 A^{d-1} 的欧氏距离均值，若该欧氏距离均值＞0.05，则跳回至步骤 3 执行；若该欧氏距离均值＜0.05，则新的聚类中心 A^d 作为图像集 I 的特征字典 E 输出，其中 $E_k \in E$，k 为大于等于 1 的自然数，E_k 为向量，所述特征字典 E 即为图像聚类中心。

步骤 6，采用 $idf_k = \log \frac{D}{\sum_a S_a \in E_k}$ 计算 E 中每个聚类中心权重，其中：D 为 S 中所有 SIFT 特征的总数，D 为大于等于 1 的自然数，若 S_a 到 E_k 的欧式距离最小，即 S_a 归属于 E_k 中心，$\sum_a S_a \in E_k$ 表示归属于 E_k 中心的所有 SIFT 特征总数。

步骤 8，分别计算图像 I_i 中每个特征向量 F_{ib} 与特征字典 E 的欧氏距离 h，其中 $E_k \in E$，$h_k \in h=(h_1,h_2,...,)$，k 为大于等于 1 的自然数，从 E 中选取 h_k 最小的 m 个 E_k，组成特征字典 E′，E′=（E_f，...，E_g），其中 E′中有 m 个向量，f 为大于等于 1 的自然数，g 为大于等于 1 的自然数，g＞f；

利用 $C_{ib} = (E' - 1F_{ib}^{\ T})(E' - 1F_{ib}^{\ T})^T$ 计算 F_{ib} 与特征字典 E′的平方差矩阵 C_{ib}；

步骤 9，计算 F_{ib} 在特征字典 E′中的稀疏编码 c_{ib}，$c_{ib} = \tilde{c}_{ib} / 1^T \tilde{c}_{ib}$，$\tilde{c}_{ib} = (C_{ib} + diag(h'))\backslash 1$，其中 $h_m' \in h' =（h_1',h_2',...,）$ 为特征向量 F_{ib} 分别到 E′中视觉词的欧氏距离，$\mathrm{diag}(h')$ 表示将向量 h′的元素作为矩阵的主对角线；

提取 c_{ib} 中的 k 个最大值，得到图像 I_i 的图像稀疏编码 g_i，其中 $g_{ik} \in g_i$；

根据 $g_i' = g_i \cdot idf / \sum_k g_{ik} \cdot idf_k$ 对图像稀疏编码 g_i 进行逆向视觉词频加权归一化，得到 IDF 加权稀疏编码 g′，其中 $g_i' \in g'$。

步骤 10，提取图像 I_i 的 IDF 加权稀疏编码 g_i' 中的非零元素；$g_{ik}' \in g_i'$，k 为大于等于 1 的自然数，g_i' 内的非零元素为（g_{iu}'，...，g_{iv}'），设非零元素为 m

个，m 为大于等于 1 的自然数，m≤k, $g_{iu}'\neq 0$, $g_{iv}'\neq 0$,u 为大于等于 1 的自然数，v 大于等于 1 的自然数，k>v>u；

步骤 11，建立 k 个组，分别命名为：$\begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$；其中，$\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$为空矩阵；

步骤 12，利用（式 1）的矩阵变换，将图像 I_i 的 IDF 加权稀疏编码 g_i' 分别散列到非零元素的下标(u, . . . , v)对应的 m 个组里；

$$g'=\begin{bmatrix} g'_{11} & g'_{12} & \cdots & g'_{1k}\\ g'_{21} & g'_{22} & \cdots & g'_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ g'_{i1} & g'_{i2} & \cdots & g'_{ik}\\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \Rightarrow \begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots & \\ \ddots & \\ \cdots & g'_i\\ \ddots & \\ \cdots & g'_i\\ \ddots & \\ \cdots & \end{bmatrix} \qquad (式\ 1)$$

步骤 13，利用 $Y=J(g_i',g_j')=\frac{|g_i'\cap g_j'|}{|g_i'\cup g_j'|}$ 计算步骤（3）所得 m 组中的每个组中每对图像<I_i,I_j>IDF 加权稀疏编码的相似度 Y，若 Y 大于 0.7，则图像<I_i,I_j>为相似图像对；

其中，j 为大于等于 1 的自然数，且 i≠j；g_i'和g_j'分别表示图像 I_i 和 I_j 的 IDF 加权稀疏编码；

步骤 14，将步骤 13 所得结果中具有相同图像的相似图像对合并，生成相似

图像子集。

实施例 2

本实施例利用百度图片搜索引擎分别检索 30 个全球著名景点或建筑（例如：埃及埃菲尔铁塔、白宫、大雁塔等）的图片，并人工从对每个类型检索到的图像中挑选清晰、准确的 100 幅图像，组成 3000 幅近重复图像，同时从公开数据集 Flickr-100M（来源：http://webscope.sandbox.yahoo.com）中随机选择 17,000 副照片作为干扰项，连同近重复图像组成实验的图像数据集，共 20,000 张图像。

选择 Hadoop 2.4.0 版本作为实验平台，共 10 个节点计算机构成本实施例的 Hadoop 集群。由于 Hadoop 自身并不支持图像数据的读取和处理，所以我们基于 Hadoop 的 java 开源框架自定义了两个类型：ImageBundle 和 ImageWritable。ImageBundle 类似于 Hadoop 自带 SequenceFile，它可以将大量的图像格式文件合并为一个大的文件，并以固定键值对形式<key:imageID,value:imageWritable>序列化存储于 HDFS。自定义 ImageWritable 继承与 Hadoop 的 Writable，用于对 ImageBundle 的编解码。继承与 Writable 的两个关键函数 encoder()和 decoder()分别用于将图像的二进制文件解码为为键值对形式和将键值对编码为二进制格式。下面为实验步骤：

1、对 20,000 张图像进行标准化的缩放并进行灰度化处理，本实施例中图像标准化后尺寸为 256*256。

2、对 20,000 张图像利用 ImageBundle 将图像编码为键值对形式并统一存储于 ImageBundle 文件，上传至 HDFS。

3、基于 Mapreduce 对步骤 2 上传的 ImageBundle 文件进行并行化特征提取，并以键值对<key:图像 ID，value：图像特征>形式存储于 SequenceFile 文件并上

传至 HDFS。本实施例中提取的图像为 SIFT 特征，SIFT 特征为图像局部特征，每幅特征包含的特征数量不定，每个 SIFT 特征 128 维。

4、基于实施例 1 所述的 MapReduce-Kmeans 算法，对步骤 3 所生成的图像特征 SequenceFile 文件进行并行化 Kmeasn 聚类。本实施例中聚类中心 K=512，循环终止条件为 0.01，当两次循环生成的聚类中心的欧氏距离均值小于 0.01 时循环结束。本步骤将产生 512 个聚类中心作为视觉字典，每个聚类中心为 128 维向量作为视觉词。

5、基于实施例 1 所述步骤 7，对步骤 4 产生的视觉字典中每个视觉词进行 IDF 权重评估。

6、从 HDFS 下载步骤 3 所生成的图像特征 SequenceFile 文件进行 MapReduce 并行化稀疏编码和相似度计算，其中 Map 函数利用实施例 1 所述步骤 8,9 对图像集进行稀疏编码并按实施例 1 步骤 10 散列给 Reduce 函数，Reduce 函数接收从 Map 函数发送的稀疏编码后按照实施例 1 步骤 11 所述方法进行相似度计算并输出大于阈值的相似对。其中本实施例采用稀疏度 L=10，每个图像的稀疏编码 512 维，相似度阈值为 0.7。

7、对本实施例步骤 6 产生的相似图像对合并，最终输出近重复相似图像集。

通过对测试数据的实验结果表明，我们的算法可以在 recall 为 0.86 时 Precision 为 0.9，总耗时 3.24 kilosecond。其中实验结果图 5 为分别对 5 个关键字（Flower,Iphone,Colosseum,Elephant,Cosmopolitan）从不同方法得到的聚类结果中随机抽样提取的 9 幅图像，F 值为 F1-measure 指标。对比的方法分别为 Partition min-Hash 算法(PmH)、Geometric min-Hash 算法(GmH)、min-hash 方法(mH)、标准 LSH 算法(st.LSH)和基于 Bag-of-Visual-Words 的树查找算法

(baseline)。实验结果图 6 是本方法对 17,000 副 Flickr 照片聚类的结果，Cluster size 为相应的聚类集合的照片数量。

权 利 要 求 书

1、一种基于稀疏表示的近似重复图像检测方法，该方法基于 hadoop 分布式计算框架提出，该检测方法包括如下步骤，获取图像集 I 中所有图像的 IDF 加权稀疏编码 g′，其中 $I=(I_1,I_2,...,I_i,...,I_w,...,I_z,...,I_R)$，$I_i$ 的 IDF 加权稀疏编码为 g_i'，$g_i' \in g'$，i 为大于等于 1 的自然数，w 为大于 i 的自然数，z 为大于 w 的自然数，R 为大于 z 的自然数，其特征在于，方法还包括：

（1）提取图像 I_i 的 IDF 加权稀疏编码 g_i' 中的非零元素；$g_{ik}' \in g_i'$，k 为大于等于 1 的自然数，g_i' 内的非零元素为 $(g_{iu}',...,g_{iv}')$，设非零元素为 m 个，m 为大于等于 1 的自然数，$m \leqslant k$，$g_{iu}' \neq 0$，$g_{iv}' \neq 0$，u 为大于等于 1 的自然数，v 大于等于 1 的自然数，$k>v>u$；

（2）建立 k 个组，分别命名为：$\begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$；其中，$\begin{bmatrix}\cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\\ \ddots\\ \cdots\end{bmatrix}$ 为空矩阵；

（3）利用（式 1）的矩阵变换，将图像 I_i 的 IDF 加权稀疏编码 g_i' 分别散列到非零元素的下标 $(u,...,v)$ 对应的 m 个组里；

$$g'=\begin{bmatrix} g'_{11} & g'_{12} & \cdots & g'_{1k}\\ g'_{21} & g'_{22} & \cdots & g'_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ g'_{i1} & g'_{i2} & \cdots & g'_{ik}\\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \Rightarrow \begin{matrix}1\\ \vdots\\ u\\ \vdots\\ v\\ \vdots\\ k\end{matrix}\begin{bmatrix}\cdots & \\ \ddots & \\ \cdots & g'_i\\ \ddots & \\ \cdots & g'_i\\ \ddots & \\ \cdots & \end{bmatrix} \qquad \text{（式 1）}$$

（4）利用$Y = J(g_i^{'}, g_j^{'}) = \frac{\left|g_i^{'} \cap g_j^{'}\right|}{\left|g_i^{'} \cup g_j^{'}\right|}$计算步骤（3）所得 m 组中的每个组中每对图像<$I_i$,$I_j$>IDF 加权稀疏编码的相似度 Y，若 Y 大于 0.7，则图像<I_i,I_j>为相似图像对；

其中，j 为大于等于 1 的自然数，且 i≠j；$g_i^{'}$和$g_j^{'}$分别表示图像 I_i 和 I_j 的 IDF 加权稀疏编码；

（5）将步骤（4）所得结果中具有相同图像的相似图像对合并，生成相似图像子集。

2、如权利要求 1 所述基于稀疏表示的近似重复图像检测方法，其特征在于，

所述获取图像集 I 中所有图像的 IDF 加权稀疏编码 g′，包括以下步骤：

并行化提取每副图像的局部特征，得到图像集 I 中所有图像的局部特征 S；

提取图像聚类中心，得到特征字典 E；

计算 E 中每个聚类中心权重；

根据 E 中每个聚类中心权重，提取图像的 IDF 加权稀疏编码 g′。

3、如权利要求 2 所述基于稀疏表示的近似重复图像检测方法，其特征在于，所述并行化提取图像局部特征为提取图像集 I 所有图像的 SIFT 特征。

4、如权利要求 3 所述基于稀疏表示的近似重复图像检测方法，其特征在于，所述提取所有图像的 SIFT 特征，具体步骤为：

将图像集 I 中每副图像 I_i 的大小标准化并灰度处理，得到标准大小的灰度图像集；其中，I=(I_1,I_2,...,I_i,...,I_w,…,I_z,…,I_R)；

将标准大小的灰度图像集分割给各集群结点，并行化提取每个图像的 SIFT

特征，所有图像的 SIFT 特征表示为 S，其中 $S_a \in S$，S_a 代表向量,a 为大于等于 1 的自然数，第 I_i 个图像的 SIFT 特征为 F_i，其中 $F_{ib} \in F_i$，F_{ib} 代表向量，b 为大于等于 1 的自然数。

5、如权利要求 2 所述基于稀疏表示的近似重复图像检测方法，其特征在于，所述提取图像聚类中心，具体步骤为：

(11)计算 S_a 到聚类中心 A^d 的欧式距离，将 S_a 作为 value，将到 S_a 欧氏距离最近的聚类中心作为 key 值；$A_{dk} \in A^d$，d 为大于等于 0 的非负整数，k 为大于等于 1 的自然数，A_{dk} 代表向量；从 S 中随机选取额的 k 个 SIFT 特征作为初始 k 个聚类中心,形成初始聚类中心集合 A^0，$A_{0k} \in A^0$；

(21)求取 key 值相同的 S_a 的平均值，d=d+1，将各平均值作为新的聚类中心 A^d；

(31)计算新的聚类中心 A^d 与上一次循环的聚类中心 A^{d-1} 的欧氏距离均值，若该欧氏距离均值＞0.05，则跳回至步骤(11)执行；若该欧氏距离均值＜0.05，则新的聚类中心 A^d 作为图像集 I 的特征字典 E 输出，其中 $E_k \in E$，k 为大于等于 1 的自然数，E_k 为向量，所述特征字典 E 即为图像聚类中心。

6、如权利要求 2 所述基于稀疏表示的近似重复图像检测方法，其特征在于，所述计算 E 中每个聚类中心权重，具体步骤为：采用 $idf_k = \log \frac{D}{\sum_a S_a \in E_k}$ 计算 E 中每个聚类中心权重，其中：D 为 S 中所有 SIFT 特征的总数，D 为大于等于 1 的自然数，$\sum_a S_a \in E_k$ 表示归属于 E_k 中心的所有 SIFT 特征总数。

7、如权利要求 2 所述基于稀疏表示的近似重复图像检测方法，其特征在于，所述提取 IDF 加权图像稀疏编码，具体步骤为：

分别计算图像 I_i 中每个特征向量 F_{ib} 与特征字典 E 的欧氏距离 h，其中 $E_k \in E$，$h_k \in h=（h_1,h_2,...,）$，k 为大于等于 1 的自然数，从 E 中选取 h_k 最小的 m 个 E_k，组成特征字典 E′，E′ =（E_f，...，E_g），其中 E′ 中有 m 个向量，f 为大于等于 1 的自然数，g 为大于等于 1 的自然数，g＞f；

利用 $C_{ib} = (E' - 1F_{ib}^{T})(E' - 1F_{ib}^{T})^{T}$ 计算 F_{ib} 与特征字典 E′ 的平方差矩阵 C_{ib}；

计算 F_{ib} 在特征字典 E′ 中的稀疏编码 c_{ib}，$c_{ib} = \tilde{c}_{ib} / 1^{T}\tilde{c}_{ib}$，$\tilde{c}_{ib} = (C_{ib} + diag(h'))\backslash 1$，其中 $h_m' \in h' =（h_1',h_2',...,）$ 为特征向量 F_{ib} 分别到 E′ 中视觉词的欧氏距离，$\mathrm{diag}(h')$ 表示将向量 h′ 的元素作为矩阵的主对角线；

提取 c_{ib} 中的 k 个最大值，得到图像 I_i 的图像稀疏编码 g_i，其中 $g_{ik} \in g_i$；

根据 $g_i' = g_i \cdot idf / \sum_k g_{ik} \cdot idf_k$ 对图像稀疏编码 g_i 进行逆向视觉词频加权归一化，得到 IDF 加权稀疏编码 g′。

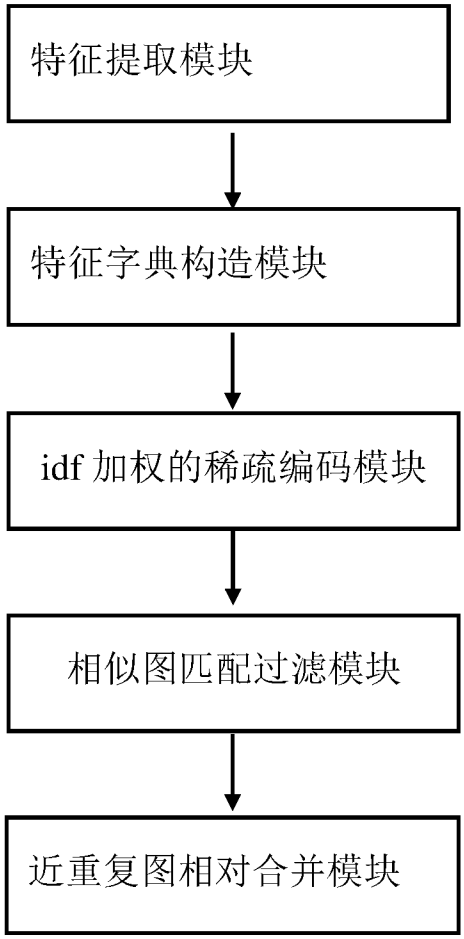


图 1

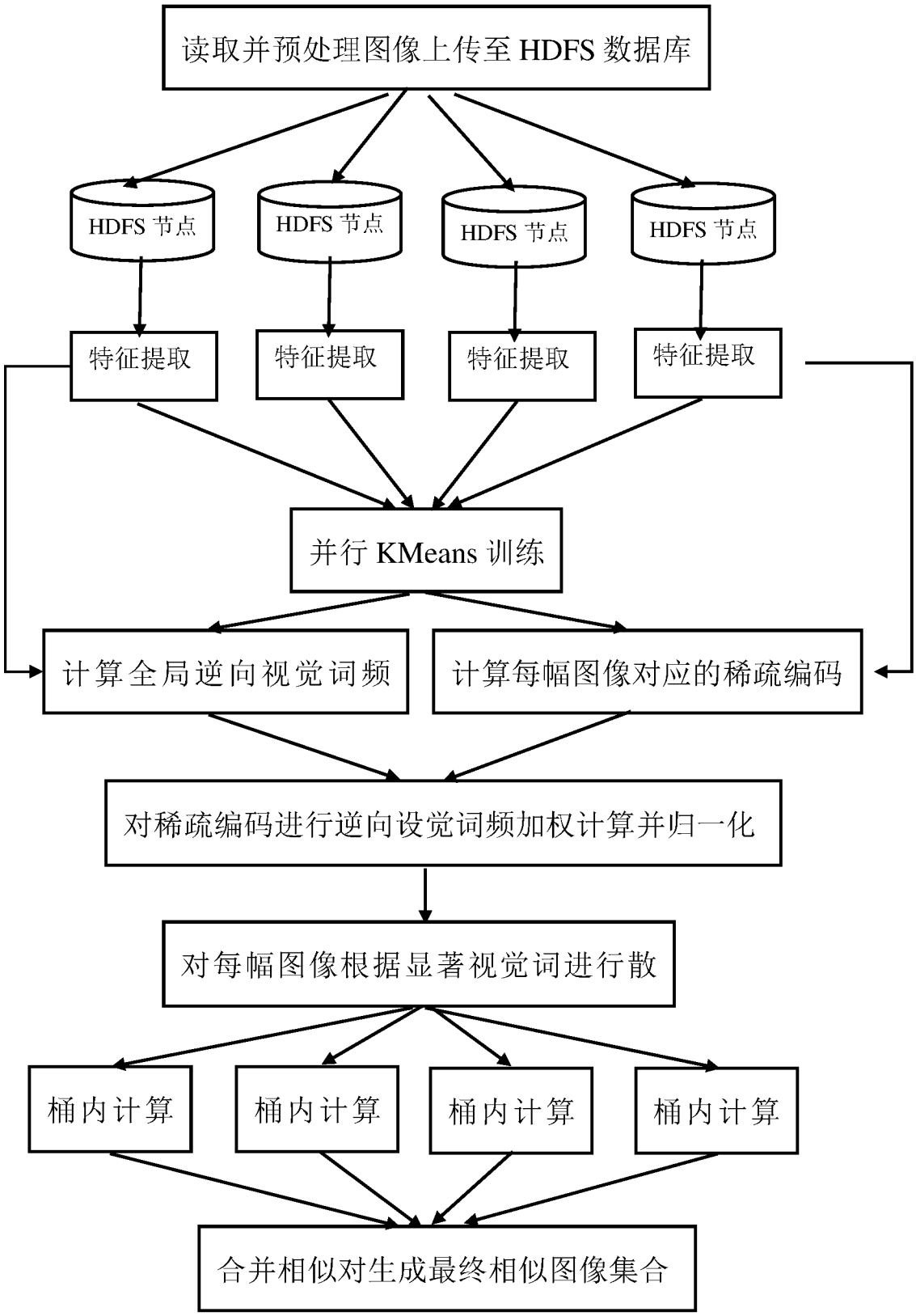

读取并预处理图像上传至 HDFS 数据库
HDFS 节点
HDFS 节点
HDFS 节点
HDFS 节点
特征提取
特征提取
特征提取
特征提取
并行 KMeans 训练
计算全局逆向视觉词频
计算每幅图像对应的稀疏编码
对稀疏编码进行逆向设觉词频加权计算并归一化
对每幅图像根据显著视觉词进行散
桶内计算
桶内计算
桶内计算
桶内计算
合并相似对生成最终相似图像集合

图 2

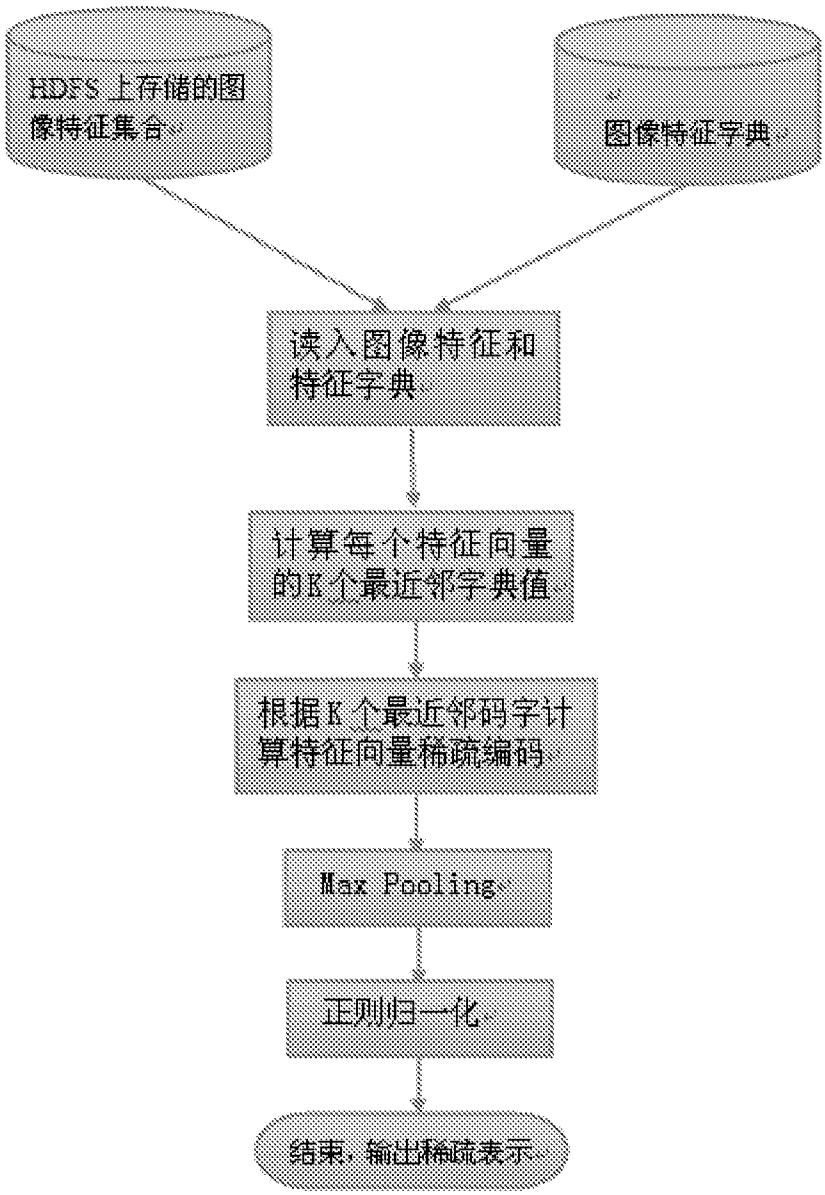

HDFS 上存储的图像特征集合
图像特征字典
读入图像特征和特征字典
计算每个特征向量的K个最近邻字典值
根据K个最近邻码字计算特征向量稀疏编码
Max Pooling
正则归一化
结束，输出稀疏表示

图 3

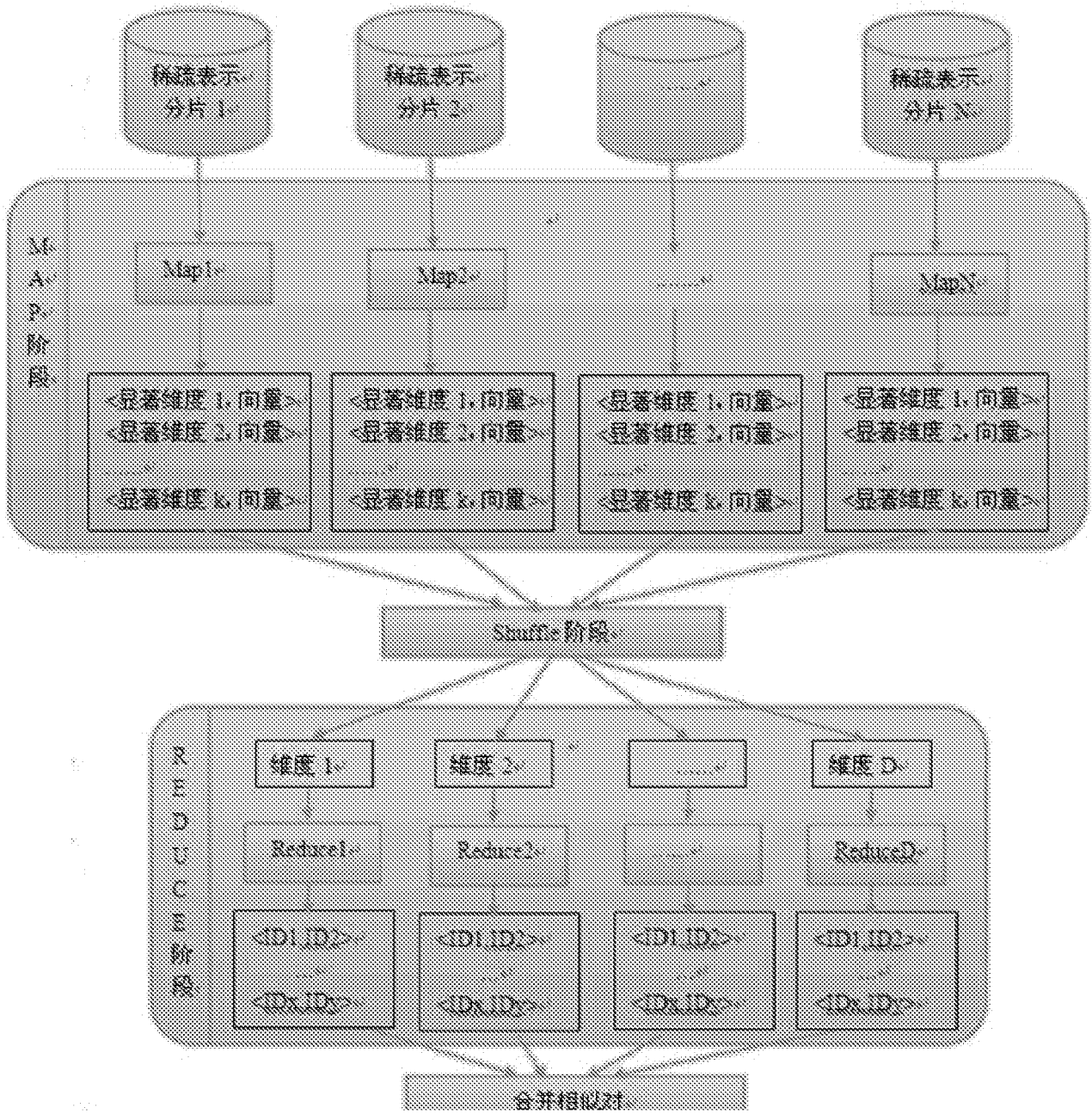

特征表示
分片 1
特征表示
分片 2
特征表示
分片 N
MAP阶段
Map1
Map2
MapN
<显著维度 1,向量>
<显著维度 2,向量>
<显著维度 k,向量>
Shuffle阶段
REDUCE阶段
维度 1
维度 2
维度 D
Reduce1
Reduce2
Reduce3
<ID1,ID2>
<IDx,IDy>
合并相似对

图 4

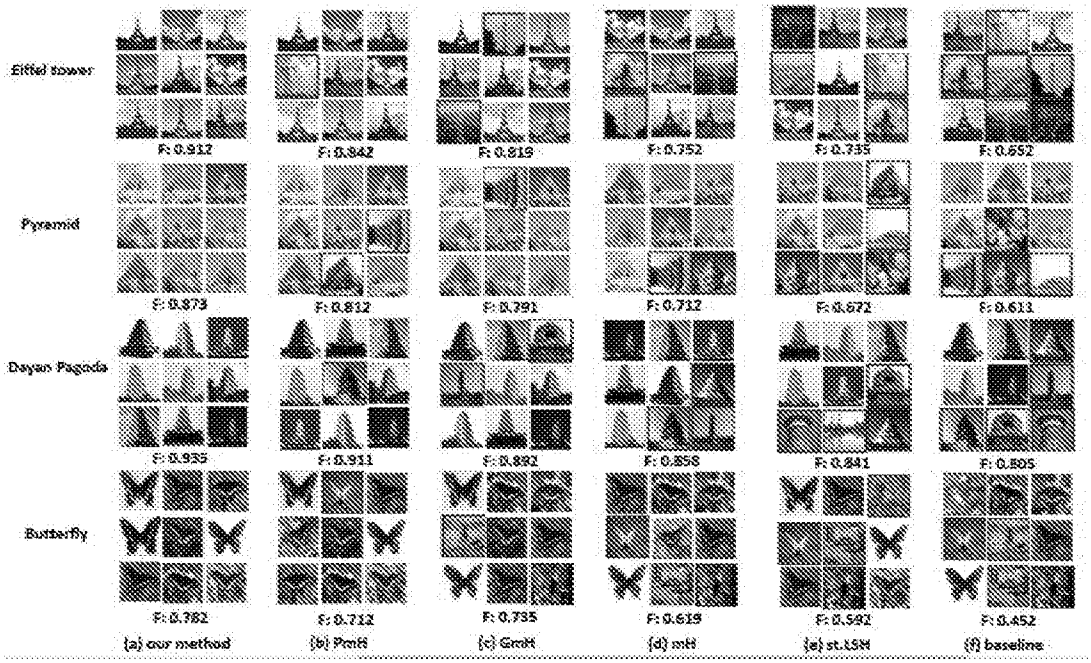

Eiffel tower
F: 0.912
F: 0.842
F: 0.819
F: 0.752
F: 0.735
F: 0.652
Pyramid
F: 0.873
F: 0.812
F: 0.791
F: 0.712
F: 0.672
F: 0.611
Dayan Pagoda
F: 0.935
F: 0.911
F: 0.892
F: 0.858
F: 0.841
F: 0.805
Butterfly
F: 0.782
F: 0.712
F: 0.735
F: 0.619
F: 0.582
F: 0.452
(a) our method
(b) PmH
(c) GmH
(d) mH
(e) st.kSH
(f) baseline

图 5

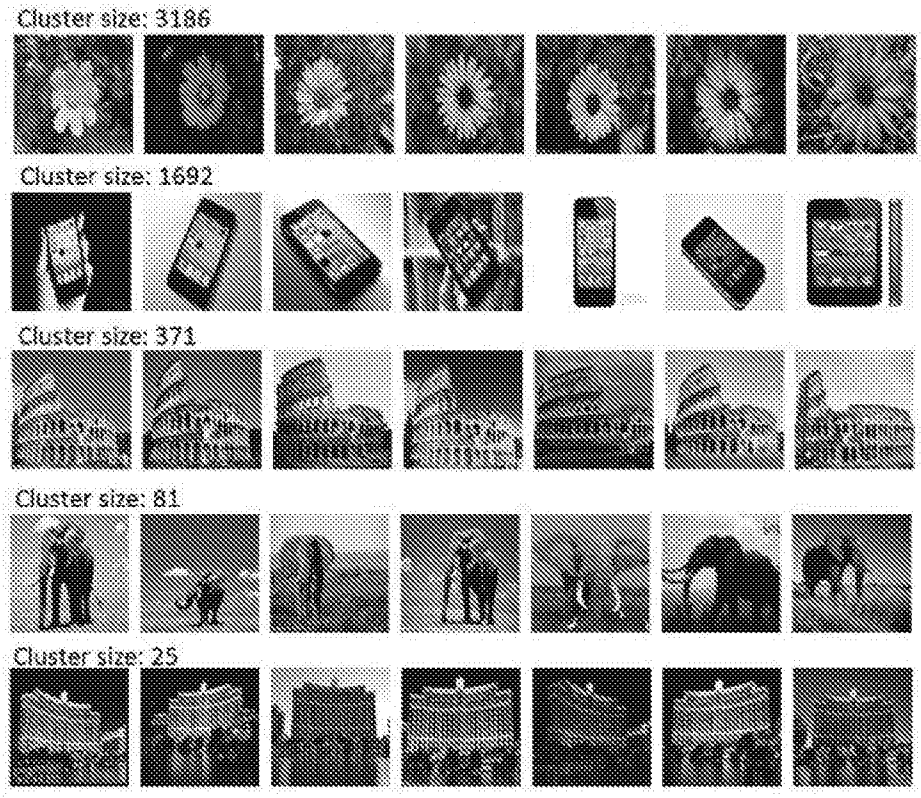

Cluster size: 3186
Cluster size: 1692
Cluster size: 371
Cluster size: 81
Cluster size: 25

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/070197

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K, G06T, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT; CNKI; WPI; EPODOC: 稀疏, 图像, 检测, 匹配, 编码, 相似, 特征, 分类, repeating, image, match+, characteristic, code, similar+, network, mapreduce, classif+, divid+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104504406 A (CHANGAN COMMUNICATION TECHNOLOGY CO., LTD.), 08 April 2015 (08.04.2015), description, paragraphs [0036]-[0054]	1-7
A	CN 104462199 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES), 25 March 2015 (25.03.2015), entire document	1-7
A	CN 104392250 A (INSPUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 04 March 2015 (04.03.2015), entire document	1-7

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 August 2017	07 September 2017

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
LIU, Man
Telephone No. (86-10) 62414025

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/070197

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104504406 A	08 April 2015	None	
CN 104462199 A	25 March 2015	None	
CN 104392250 A	04 March 2015	None	

国际检索报告

国际申请号
PCT/CN2017/070197

A. 主题的分类

G06K 9/62(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06K G06T G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT;CNKI;WPI;EPODOC:稀疏，图像，检测，匹配，编码，相似，特征，分类，repeating, image, match+, characteristic, code, similar+, network, mapreduce, classif+, divid+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 104504406 A (长安通信科技有限责任公司) 2015年 4月 8日 (2015-04-08) 说明书第[0036]-[0054]段	1-7
A	CN 104462199 A (中国科学院自动化研究所) 2015年 3月 25日 (2015-03-25) 全文	1-7
A	CN 104392250 A (浪潮电子信息产业股份有限公司) 2015年 3月 4日 (2015-03-04) 全文	1-7

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2017年 8月 23日	2017年 9月 7日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 刘曼 电话号码 (86-10)62414025

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/070197

检索报告引用的专利文件	公布日（年/月/日）	同族专利	公布日（年/月/日）
CN 104504406 A	2015年 4月 8日	无	
CN 104462199 A	2015年 3月 25日	无	
CN 104392250 A	2015年 3月 4日	无	