

基于聚类分析的无监督视觉词典构造方法研究

徐望明, 郑超兵

(武汉科技大学 信息科学与工程学院, 湖北 武汉 430081)

摘 要: 视觉词典是运用视觉词袋模型有效表示图像的基础和关键. 研究了基于聚类分析的无监督视觉词典构造方法, 针对 K-Means 和层次 K-Means (HKM) 聚类算法在构造视觉词典时对初始值敏感和容易陷入局部最优的问题进行改进, 提出了一种基于谱聚类的视觉词典构造方法, 采用对观测数据的相似性矩阵进行特征值分解的方法来实现聚类, 可在任意分布的特征空间收敛到全局最优. 通过图像检索实验评估了随机采样、K-Means、HKM 和改进的谱聚类视觉词典的性能, 验证了聚类分析尤其是改进的谱聚类方法构造视觉词典的有效性.

关键词: 视觉词典; 谱聚类; 图像检索; 视觉词袋

中图分类号: TP391.4

文献标识码: A

文章编号: 1000-7180(2016)03-0152-05

Unsupervised Visual Dictionary Construction Methods Via Clustering Analysis

XU Wang-ming, ZHENG Chao-bing

(College of Information Science and Engineering, Wuhan University of Science and Technology,
Wuhan 430081, China)

Abstract: Visual dictionary is the basic and key to efficiently represent image with Bag of Visual Words (BoVW) Model. The unsupervised visual dictionary construction methods based on clustering analysis are researched in this paper. Aiming at the problems that K-Means based clustering methods (K-Means and HKM) are sensitive to the initialization and easy to converge to the local optimum, a modified spectral clustering method is proposed to construct visual dictionary, which works in a way to cluster observations via eigenvalue decomposition according to their similarity matrix and can converge to the global optimal regardless of the distribution of feature space. The performance of visual dictionaries constructed by Random Sampling, K-Means, HKM and the improved Spectral Clustering is assessed through image retrieval experiments, and the effectiveness of clustering analysis especially the improved Spectral Clustering method to construct visual dictionary is validated.

Key words: visual dictionary; spectral clustering image retrieval; bag of visual words

1 引言

基于内容的图像检索 (Content-Based Image Retrieval, CBIR) 是计算机视觉和信息检索领域的热门研究课题. CBIR 的目的在于从图像库中找出与给定图像视觉内容最为相似的图像. 如何表示图像是 CBIR 技术研究中需要解决的首要问题. 图像原始的“像素级”表示数据量大且对图像变换和干扰敏感,

因此, 图像表示的有效途径是提取图像特征来描述图像的视觉内容.

视觉词典是 BoVW 模型的基础和关键组成部分, 本文主要研究基于聚类分析的无监督视觉词典构造方法, 针对 K-Means 和 HKM 聚类方法在构造视觉词典时对初始值敏感和容易陷入局部最优的问题进行改进, 提出一种基于谱聚类 (Spectral Clustering, SC) 算法^[1-2]的视觉词典构造方法. 谱聚类在建

立任意维观测数据的相似性矩阵后采用特征值分解的方法在低维空间来实现聚类,可在任意分布的样本空间收敛到全局最优。

2 BoVW 模型表示图像的基本流程

视觉词袋(BoVW)模型是词袋(Bag of Words, BoW)模型从文本检索领域向图像检索领域的一次自然推广^[3],可有效应对大规模图像数据库给CBIR等计算机视觉任务带来的巨大挑战。利用BoVW模型表示图像,获得图像的全局直方图表示形式,其基本流程如下所示。

(1)图像局部特征提取。根据具体应用场合选择合适的局部特征提取算法,从图像中提取局部特征。

(2)视觉词典构造。从图像库中选取部分来自不同场景或类别的图像来组成训练图像集,提取局部特征构成训练特征集,通过适当的去冗余处理得到一些有代表性的特征向量,将其定义为视觉单词。通常所采用的处理方法是聚类分析,将聚类中心定义为视觉单词,所有视觉单词组成视觉词典,视觉单词的数量为视觉词典大小。

(3)特征编码。BoVW模型一般采用向量量化(VQ)技术将图像的局部特征向量量化为视觉单词中与其距离最近的视觉单词。每个局部特征的编码结果是一个维数与视觉词典大小相同的向量,对应最近邻视觉单词位置的编码为1,其余位置编码为0。为了减少量化误差,近年来出现了多种改进的特征编码方法,如Soft-VQ、LLC等^[4-7],此时编码结果中各元素的值是加权系数或重构系数等。

(4)特征集成。一幅图像的所有局部特征向量被量化后,得到一个编码矩阵,统计视觉词典中每个视觉单词出现的频数,得到一个关于视觉单词的直方图,其本质是上一步所得量化编码的全局统计结果,是按视觉单词索引顺序组成的一个数值向量。若在上一步采用了改进的特征编码时,该向量各元素值可选用Max Pooling等不同策略得到,不再是简单的统计计数。根据需要进行进一步做归一化处理后的向量即为图像的最终全局特征表示形式。

3 基于聚类分析的视觉词典构造方法

BoVW模型中,图像的最终表示是依赖于视觉词典的。BoVW模型采用对训练图像的局部特征进行无监督学习——聚类分析的方式来构造视觉词典。聚类过程实际上是相似局部特征逐渐合并的过

程,以聚类中心定义视觉单词,组成视觉词典。常用的聚类分析方法是K-Means和HKM聚类,本文针对其不足引入谱聚类方法来构造视觉词典。

3.1 基于K-Means聚类的视觉词典构造

K-Means算法将训练图像的局部特征看作高维空间中的离散点,以误差平方和最小化为聚类准则,自动寻找聚类中心,再将这些聚类中心作为视觉单词构成视觉词典。其算法步骤如下所示。

算法1:基于K-Means聚类构造视觉词典

设: $X = \{x_1, \dots, x_n\} \in R^d$ 是 d 维空间中的 n 个特征向量所构成的图像局部特征训练数据集; $S = \{S_1, S_2, \dots, S_K\}$ 是数据集 X 上的 K 个非空类; $B = \{b_1, \dots, b_K\}$ 是类 S 的聚类中心向量集,即视觉词典; $U(X) = U = [u_{ki}]$ 表示 X 中所对应的 n 个特征向量排成的 $K \times n$ 类属关系矩阵, u_{ki} 表示 x_i 在 S_k 类里面的类属关系,属于该类为1,不属于为0。

步骤1:输入训练数据矩阵 X 和聚类数 K ,初始化聚类中心 $B = \{b_1, \dots, b_K\}$,并设定最大迭代次数 T ,聚类中心收敛误差 ε ;

步骤2:逐一计算特征向量 $x_i (i = 1, \dots, n)$ 与各个聚类中心 $b_k (k = 1, \dots, K)$ 的距离 d_{ik} ,据此可将 x_i 分配到离其最近的聚类中心所表示的类,即:若 $d_{ik^*} = \min_k \{d_{ik}\}$, 则 $x_i \in S_{k^*}$, 且 $u_{ki} = \begin{cases} 1, & \text{当 } k = k^* \\ 0, & \text{当 } k \neq k^* \end{cases}$;

步骤3:根据最新类属关系矩阵 U 和样本训练数据 X 重新计算聚类中心(视觉词典) $B = \{b_k \mid b_k$

$$= \frac{\sum_{i=1}^n (u_{ki}) x_i}{\sum_{i=1}^n u_{ki}}, k = 1, \dots, K\}, \text{并计算当前聚类中心}$$

与前一次迭代的聚类中心的距离偏差 $E_t = \|B_t - B_{t-1}\|_2 = \sqrt{(B_t - B_{t-1})^T (B_t - B_{t-1})}$, 其中 t 为迭代次数;

步骤4:反复执行步骤2和步骤3,直到条件 $E_t < \varepsilon$ 或 $t = T$ 成立时终止,输出最终的 U (可选) 和 B (即所构造的视觉词典)。

3.2 基于HKM聚类的视觉词典构造

K-Means聚类的算法复杂度与特征数据维度、个数及视觉词典大小均成正比。对于给定的较大规模的训练特征数据,一种有效的提升聚类效率的方法是采用分层策略进行K-Means聚类,即HKM聚类。其算法步骤如下所示。

算法2:基于HKM聚类构造视觉词典

设: $X = \{x_1, \dots, x_n\} \in R^d$ 是 d 维空间中的 n 个特征向量所构成的图像局部特征训练数据集; L 和 K 分别是指定的聚类层数和每一层 K-Means 聚类数; $B = \{b_1, \dots, b_{KL}\}$ 是最后一层聚类中心向量集(叶子节点), 即视觉词典。

步骤 1: 输入训练数据(特征向量)矩阵 X 、层数 L 和每一层 K-Means 聚类数 K ;

步骤 2: 初始化 K-Means 聚类, 将特征空间划分为 K 个子区域, 每个子区域中包含一系列落入该区域的特征向量;

步骤 3: 在每个子区域中再分别进行 K-Means 聚类;

步骤 4: 反复执行步骤 3, 不断划分特征空间, 即可得到一个树状的聚类结构, 从而可输出最终 B (即所构造的视觉词典)。

通过上述方法构建的聚类树中, 树的高度由聚类层数 L 决定, 树的根节点为全部训练数据(特征向量)的集合, 树中每个父节点均有 K 个子节点, 每个子节点是该层中每个子类的聚类中心, 将每个叶子节点看作“视觉单词”, 所有叶子节点便组成了视觉词典。

可见, 使用 HKM 聚类方法, 一棵 L 层的 K 分支树生成的视觉词典大小为 K^L 。若要从 n 个特征向量中构造具有相同大小的视觉词典 B , 则使用 K-Means 算法执行 1 次聚类的算法复杂度是 $O(nKL)$, 而使用 HKM 算法执行 1 次聚类的复杂度是 $O(nKL)$ 。一般 $KL \gg KL$, 故 HKM 聚类效率明显高于 K-Means, 不过是以牺牲部分聚类精度为代价的。

3.3 基于改进谱聚类的视觉词典构造

K-Means 和 HKM 聚类适合于构造视觉词典, 因其本身已被赋予了基于“中心”收敛的算法定义, 对观测数据的聚类结果即为聚类中心, 直接构成视觉词典, 但这两个算法均对聚类中心的初始值比较敏感, 且通常收敛于局部最优。若初始聚类中心选择不当, 可能无法得到有效结果。许多研究工作已证实谱聚类比 K-Means 在聚类效果上更具优势, 它利用观测数据的相似性矩阵(Similarity Matrix)的谱(特征值)进行降维, 采用特征值分解的方法在低维空间中完成聚类, 可在任意分布的特征空间收敛于全局最优。谱聚类算法可描述如下:

给定包含 N 个样本的数据集 $\{x_i\}_{i=1}^N$, 其中 $x_i \in R^d$, 定义任一数据点 x_i 和 x_j 的相似度为 $a_{ij} \geq 0$ 。这些训练样本可表示为图 $G = (V, E)$, V 是顶点集合, E 是边集合。顶点集合 V 表示训练样本数据集, 边 $e_{ij} \in E$ 则刻画了所关联的两个顶点 v_i 和 v_j 的相似性, 权

重为 a_{ij} 。若该相似性测度是对称的, 则该图为无向图。本文基于欧式距离采用互 K 近邻策略^[8] 构建这种无向图。记图 G 的相似性矩阵 A 的第 i 行、第 j 列元素为 $A_{ij} = a_{ij}$ 。顶点 v_i 的度(degree) d_i 定义为与 v_i 相连的所有顶点构成的边的权重和:

$$d_i = \sum_{j=1}^N a_{ij} \quad (1)$$

设度矩阵 D 是 $d_1, d_2, \dots, d_N$ 组成的对角阵, 即:

$$D = \text{diag}(d_1, d_2, \dots, d_N) \quad (2)$$

将拉普拉斯矩阵定义为 $L = D - A$, 则规范化拉普拉斯矩阵定义为

$$L_{NCut} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - A_N \quad (3)$$

则可定义规范化相似性矩阵为

$$A_N = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \quad (4)$$

因此, 谱聚类可表示为如下优化问题:

$$\max_U \|U^T L_{NCut} U\|_F^2, \text{ s. t. } U^T U = I \quad (5)$$

根据 A_N 和 L_{NCut} 之间的关系, 谱聚类还可表示为如下优化问题:

$$\max_U \|U^T A_N U\|_F^2, \text{ s. t. } U^T U = I \quad (6)$$

这里 U 是松弛指示矩阵, 在其上使用 K-Means 聚类便可获得最终的数据划分, 每个样本数据都将被赋予一个“聚类标签”。谱聚类问题可通过对 A_N 或 L_{NCut} 进行特征值分解而求解。由 Rayleigh-Ritz 理论可知, 谱聚类算法得到的拉普拉斯矩阵的若干特征向量就是最优划分问题松弛后的近似解。谱聚类有不同的实现方法, 其中 NJW 算法因其良好的稳定性成为最常用而又经典的谱聚类实现算法之一。

本文提出运用和改造谱聚类算法来构造视觉词典。上述谱聚类中最终进行数据划分所使用的 K-Means 是在一个低维空间而非在原始特征空间中执行, 故其聚类中心不能直接定义为视觉单词; 但谱聚类能返回训练样本的“聚类标签”, 这样就可按“聚类标签”对训练样本分组, 并计算每组样本的中心(即为训练样本在原始特征空间的聚类中心), 这些中心随即被定义为视觉单词, 从而构成视觉词典。

下面给出使用谱聚类(NJW 算法)构造视觉词典的步骤。

设 $X = \{x_1, \dots, x_n\} \in R^d$ 是 d 维空间中的 n 个特征向量所构成的图像局部特征训练数据集, 要将其聚成 K 类, 并将聚类中心定义为视觉单词构成视觉词典 $B = \{b_1, \dots, b_K\}$ 。则图 G 可描述为: 以每个样本数据 x_i 作为顶点, 以顶点间的相似度连边并赋以权值。该聚类问题被化为对图 G 进行划分的问题, 可

用如下谱聚类算法解决。

算法3:基于改进谱聚类(NJW 算法)构造视觉词典

输入:特征数据集 X , 最大聚类数 K , 相似性度量函数参数 σ , K-Means 算法迭代终止条件 ε 以及最大迭代次数 T

输出:视觉词典 B

步骤1:构造相似性矩阵 A . 根据 $A_{ij} = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ 构造相似性矩阵($i \neq j$); $i = j$ 时, 令 $A_{ij} = 0$; (注:本文中限定 x_i 与 x_j 互为 K 近邻时计算 A_{ij} , 否则 $A_{ij} = 0$.)

步骤2:构造规范化拉普拉斯矩阵 L_{NCut} .

(1) 计算度 $d_i = \sum_{j=1}^N A_{ij}, i = 1, 2, \dots, N$;

(2) 构造度矩阵 $D = \text{diag}(d_1, d_2, \dots, d_N)$;

(3) 计算规范化拉普拉斯矩阵 $L_{NCut} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

步骤3:求规范化拉普拉斯矩阵 L_{NCut} 的特征值和特征向量。

(1) 计算 L_{NCut} 的前 K 个最大特征值及其对应的特征向量 $u_1, \dots, u_K$;

(2) 将得到的这 K 个特征向量排列成一个新矩阵 $U = [u_1, \dots, u_K] \in R^{n \times K}$.

步骤4:将矩阵 U 的行向量转换为单位向量, 得矩阵 Y , 即 $Y_{ij} = \frac{U_{ij}}{\sqrt{U^2_{ij}}}$.

步骤5:对矩阵 Y 进行聚类. Y 的每一行代表 R^K 空间中的一个点, 可用 K-Means 对 Y 进行聚类, 得到 K 个类. 于是, 存在如下映射关系: 当且仅当矩阵 Y 的第 i 行被划分到第 j 类时, 原始特征数据 x_i 也被划分到第 j 类。

步骤6:求特征数据在原始特征空间的聚类中心. 将所有被划分到第 j 类中的特征数据 x_i 的集合记为 C_j , 则该类特征数据的聚类中心为 $b_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i, (j = 1, 2, \dots, K)$. 将其定义为视觉单词, 即可构成视觉词典 $B = \{b_1, \dots, b_K\}$.

4 实验及结果分析

本节以图像检索为例对以上聚类分析方法构造的视觉词典的性能进行评估. 采用图像库为 ZuBud^[10], 该库是 CBIR 技术研究中常用的标准数据库之一, 包含 201 栋建筑物共 1 005 幅图像, 每个

建筑物各 5 幅, 原始分辨率 640×480 (本文实验中将其缩放到 320×240 以减少数据量), 是在不同季节和天气条件下从不同视角由两个不同的相机所拍摄, 部分图像如图 1 所示。

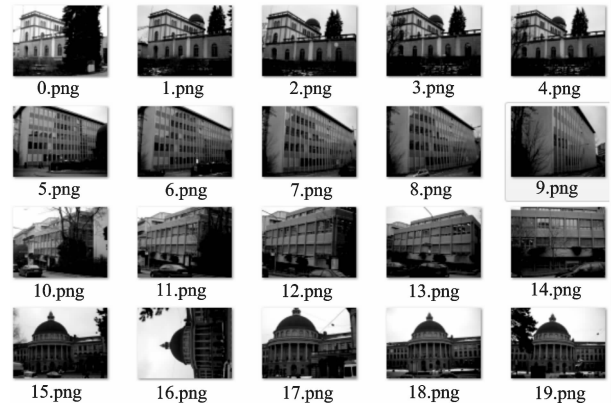


图1 ZuBud 库中的部分图像

实验参数设置如下:

局部特征提取: 对每幅图像提取 128 维的 SIFT 特征。

视觉词典构造: 从每建筑物对应的 5 幅图像中各取 1 幅构成训练图像集, 将这 201 幅图像的 SIFT 特征组成训练特征集, 设置不同的词典大小 K (本文实验取为 $256 = 4^4, 512 = 8^3, 1\ 000 = 10^3$), 运用随机采样 (即不使用聚类分析而随机选取视觉单词)、K-Means、HKM 和谱聚类方法分别构造视觉词典。

特征编码: 使用 LLC 编码方法^[8-9], 用 5 近邻视觉单词对查询图像的局部特征进行稀疏重构。

特征集成: 使用 Max Pooling 方法, 集成后进行 L2-正则化处理, 从而形成图像的最终全局表示, 为 K 维单位向量。

图像检索过程和性能评价: 为了便于统计和评价检索性能, 取每幅库图像作为查询图像, 图像检索过程简化为用两个稀疏单位向量的内积 (夹角余弦) 来衡量两幅图像的相似度 (内积越大越相似), 并按降序排列返回前 T 幅图像作为检索结果. 根据 ZuBud 库特点, 取 $T=5$, 此时查准率 (一次查询过程中系统返回的查询结果中相关图像数占所有返回图像数的比例) 与查全率 (一次查询过程中系统返回的查询结果中相关图像数占图像库中所有相关图像数的比例) 相同. 这里采用平均查准率 (Average Precision, AP) 作为性能评价标准, 即:

$$AP = \frac{1}{N} \sum_{i=1}^N \text{Precision}_i = \frac{1}{N} \sum_{i=1}^N \frac{n_i}{T} \quad (7)$$

式中, n_i 是第 i 幅查询图像检索出的相关图像数目,

$N = 1\ 005$.

实验结果及分析:实验在 MATLAB 软件平台完成,表 1 所示为不同实验参数组合情况下的图像检索实验结果.

表 1 ZuBud 库上的图像检索实验结果				
序号	词典构造方法	词典大小	构造时间/ s	AP/%
1	Random Sampling	256	0.005 3	45.61
2	K-Means Clustering	256	1 261.6	56.10
3	HKM Clustering	256	36.1	53.13
4	Spectral Clustering	256	235.6	58.65
5	Random Sampling	512	0.005 5	52.76
6	K-Means Clustering	512	1 394.6	64.34
7	HKM Clustering	512	42.3	61.81
8	Spectral Clustering	512	1 261.6	66.55
9	Random Sampling	1 000	0.023 8	65.63
10	K-Means Clustering	1000	2 308.3	72.44
11	HKM Clustering	1 000	61.2	71.88
12	Spectral Clustering	1 000	3 683.3	73.57

为了更加直观,将以上结果显示于图 2 中.

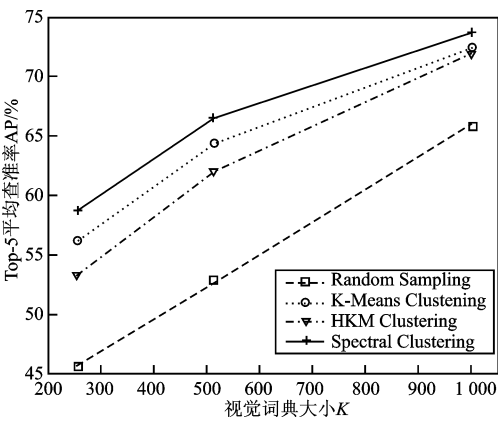


图 2 使用不同视觉词典时的图像检索性能比较

可见,相比于随机采样,聚类分析所构造的视觉词典对于改善图像检索性能的作用是比较明显的.而在使用聚类分析方法构造视觉词典时,HKM 聚类相比于 K-Means 聚类提高了效率但降低了部分精度,谱聚类则相比于 K-Means 聚类和 HKM 聚类都取得了更好的效果.

5 结束语

本文对 BoVW 模型中基于聚类分析的无监督视觉词典构造方法进行研究分析,针对 K-Means 和 HKM 聚类方法在构造视觉词典时对初始值较敏感和容易陷入局部最优的问题进行改进,提出了一种基于谱聚类的视觉词典构造方法.相比之下,谱聚类的主要优势在于它仅与特征数据数量有关,而与特征维度无关,可避免高维特征数据造成的奇异性问题,同时它能直接通过求解规范化拉普拉斯矩阵的特征向量进行数据划分,

适用于任意分布的特征数据聚类,并能收敛于全局最优.在 ZuBud 图像库上开展的图像检索实验评估了随机采样、K-Means、HKM 和改进谱聚类视觉词典的性能,验证了聚类分析尤其是谱聚类方法构造视觉词典的有效性.不过,谱聚类算法在特征数据量大和需要构造较大的视觉词典时计算复杂度仍然较高,这是本研究工作下一步改进的方向.

参考文献:

[1] David G Lowe. Distinctive image features from scale invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[3] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos [C]//Proceeding of the 9th IEEE International Conference on Compute Vision. Nice, France;IEEE, 2003:1470-1477.

[4] DavidNister, Henrik Stewenius, Scalable recognition with a vocabulary tree [C] // Proceedings of the 2006 IEEE Computer society Conference on Computer Vistion and Pattern Recognition. Washington, DC USA: ACM, 2006: 2161-2168.

[2] UlrikeLuxburg. A tutorial on spectral clustering [J]. Statistics and Computing, 2007,17(4): 395-416.

[5] Chatfield K, Lempitsky V, Vedaldi A, et al. The devil is in the details: an evaluation of recent feature encoding methods [C] // Proc BMVC. [s. l.]: IEEE, 2011: 1 - 12.

[6] Divya Sardana, and Raj Bhatnagar. Graph clustering using mutual K-nearest neighbors [C] // Proceedings of the 10the Intertional Conference AMT. Warsaw, Ploand: Springer,2014: 35-48.

[7] Ng A Y, Jordan M L, Weiss Y. On spectral clustering: analysis and an algorithm [C]// Advances in Neural Information Processing Systems. San Ffrancisco: Morgan Kaufmann, 2002: 849-856.

[8] Shao H, Svoboda T, Van Gool L. ZUBUD-Zurich buildings database for image based recognition [R]. Swiss Federal Institute of Technology, 2003.

[9] Jinjun Wang, Jianchao Yang, Kai Yu, et al. Locality-constrained linear coding for image classification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisso, CA;IEEE, 2010: 3360-3367.

作者简介:

徐望明 男,(1979-),博士,高工.研究方向为图像处理、机器学习、模式识别. E-mail:xuwangming@wust.edu.cn.

郑超兵 男,(1988-),硕士生. 研究方向为图像处理与计算机视觉.