文章编号：2095-4980(2014)02-276-08

# 一种基于随机化视觉词汇和聚类集成目标分类

朱道广，李弼程

(信息工程大学 信息系统工程学院，河南 郑州 450002)

**摘 要**：针对传统的视觉词典法存在的时间复杂度高，视觉单词同义性、歧义性和高维局部特征聚类不稳定问题，提出了一种基于随机化视觉词汇和聚类集成的目标分类方法。采用精确欧式位置敏感哈希(E2LSH)对训练图像库的局部特征点进行哈希映射，生成一组随机化视觉词汇；然后，聚类集成这组随机化视觉词汇，构建随机化视觉词汇集成词典(RVVAD)；最后，基于该词典构建图像的视觉单词直方图并使用支持向量机(SVM)分类器完成目标分类。实验结果表明，本文方法有效增强了词典的表达能力，提高了目标分类的准确率。

**关键词**：目标分类；聚类集成；精确欧式位置敏感哈希；随机化视觉词汇

**中图分类号**：TN702 **文献标识码**：A **doi**：10.11805/TKYDA201402.0276

# An object categorization approach based on randomized visual vocabulary and clustering aggregation

ZHU Dao-guang, LI Bi-cheng

(Institute of Information System Engineering, Information Engineering University, Zhengzhou Henan 450002, China)

**Abstract**: Considering the problems with the conventional Bag-of-Visual-Words approaches, such as great time consumption, the synonymy and ambiguity of visual word, and instability of clustering high-dimensionality image local features, this paper presents a novel object categorization approach based on randomized visual vocabulary and clustering aggregation. Firstly, Exact Euclidean Locality Sensitive Hashing (E2LSH) is used to cluster local features of the training dataset, and a group of randomized visual vocabularies is constructed. Then, the randomized visual vocabularies are aggregated by using clustering aggregation technique, resulting in Randomized Visual Vocabularies Aggregating Dictionary(RVVAD). Finally, the visual words histogram is generated according to the dictionary, and the Support Vector Machines(SVM) are adopted to accomplish image object categorization. Experimental results indicate that the expression ability of the dictionary is effectively improved, and the object categorization precision is increased dramatically.

**Key words**: object categorization; clustering aggregation; Exact Euclidean Locality Sensitive Hashing; randomized visual vocabulary

在图像目标分类领域，视觉词典法[1-2]以其显著的性能已成为目标分类的主流方法。该方法借鉴了文本处理领域中的“词袋法”，其基本流程为：首先，从训练图像库中提取大量的局部特征，并应用聚类算法将这些特征点聚类得到一个视觉词典，其中每个聚类中心代表一个视觉单词；然后，对于每幅图像，将其每个局部特征点与视觉单词进行匹配，得到一个视觉单词直方图作为该图像的特征向量；最后，利用机器学习方法对这些特征向量进行训练，建立目标分类器，实现目标分类。虽然视觉词典法性能突出，但仍然存在 3 个方面的问题。

一是算法时间复杂度高。K-Means 方法对局部特征点的聚类过程涉及到大量的高维数据点最近邻查找操作，其时间复杂度与训练数据的个数、高维点的维度以及聚类的个数成正比，严重制约了视觉词典的生成速度。为此，Philbin[3]采用近似 K-Means 算法(Approximate K-Means，AKM)加快聚类收敛速度，并引入倒排文档结构进一步提高检索效率；Wang 等人[4]采用快速近似 K-Means 算法(Fast Approximate K-Means，F-AKM)，减少每轮迭代中被重新分配候选簇的数据点个数，进一步加快聚类收敛速度，提高了构建视觉词典的效率；Wu 等人[5]提出采用

收稿日期：2013-02-18；修回日期：2013-04-08

分层自适应模糊 K-Means 方法将树形结构植入到聚类过程当中，构造出了视觉词汇树(Visual Vocabulary Tree)，提高了视觉单词的索引效率。然而，常用的图像局部特征具有维数高、数量大的特点，如一幅图像通常包含 1 000 至 2 000 个原始的 128 维 SIFT[6]特征点,由于 K-Means 聚类方法在构建视觉词典过程中涉及大量的高维数据最近邻查找操作，“维数灾难”问题严重限制了 K-Means 聚类方法的效率。

二是视觉单词的同义性和歧义性问题。视觉单词的同义性是指，多个视觉单词所描述视觉内容具有很强的相似性；而歧义性是指多个视觉差异明显的图像块对应同一个视觉单词。视觉单词的同义性和歧义性问题严重制约了视觉词典法的性能，为减弱同义性和歧义性的影响，研究人员进行了诸多尝试。Gemert 等人[7]提出视觉单词不确定性(Visual Word Uncertainty)模型，并验证了其对于减弱视觉单词同义性和歧义性的有效性。López-Sastre 等人[8]提出了一种新的聚类质量评价准则，并在聚类生成视觉词典过程中引入监督机制，提高了视觉单词的区分能力。Wang 等人[9]在进行局部特征点与视觉单词的匹配时，通过使用 QP 规划(Quadratic Programming)策略提高了匹配精度。Yu 等人[10]通过引入语义上下文(Semantic Context)信息，提出了上下文嵌入视觉词汇(Context-embedded Visual Vocabulary)模型，在一定程度上减弱了视觉单词的歧义性。Zhang 等人[11]提出视觉单词语义纯度(Visual Word Semantic Purity)概念，并为视觉单词重新分配权重，提高了视觉单词的语义代表性。Yang 等人[12]提出将局部特征点的空间上下文信息作为边信息(Side Information)，指导视觉词典的构建，构造出更精确的视觉词典，提高了图像标注和分类的准确度。然而上述方法都是应用 K-Means 及其改进聚类算法进行视觉词典的构造，不能解决 BoVW 方法存在的效率问题。

三是高维局部特征点聚类的不稳定性问题。Beyer 等人[13]的研究结果表明，高维空间中诸如 K-Means,Mean Shift 等聚类方法在算法效率急剧下降的同时，其聚类结果的稳定性也将恶化。为此，Moosmann 等人[14]借鉴随机决策树和随机森林算法思想，提出了随机聚类森林(Randomized Clustering Forests)方法，并将其用于生成一组视觉词典，构建视觉词典森林，该方法减弱了视觉词典生成过程中的随机性，但其时间复杂度和内存占用率过高。Mu 等人[15]将位置敏感哈希(Locality Sensitive Hashing，LSH)[16]引入视觉词典法中，构建了多个随机化位置敏感词典(Randomized Locality Sensitive Vocabularies，RLSV)，该方法减弱了视觉单词的歧义性和视觉词典不稳定问题，但该方法只适用于汉明空间。Kontschieder 等人[17]提出将训练图像库中图像对象标签的分布信息整合到随机森林的训练当中，进一步提高了视觉词典森林的可区分能力。López-Sastre 等人[18]提出使用聚类集成(Clustering Aggregation)方法[19]联合优化多个粗糙的 K-Means 聚类结果，生成最终的视觉词典，该方法虽然能够在一定程度上减弱视觉单词的同义性和歧义性，但多次使用 K-Means 聚类方法增加了复杂度。

针对上述问题，考虑到精确欧式位置敏感哈希(E2LSH)能够高效地处理高维数据，并且能够将高维空间中距离相近的点以较大概率哈希到同一个桶中，将相距较远的点以较大概率哈希到不同的桶中，所以可将其替代传统的 K-Means 方法对局部特征点进行哈希映射聚类生成视觉词汇，降低方法的复杂度并减弱视觉单词的同义性和歧义性。此外，为减小由 E2LSH 对局部特征点哈希聚类生成的视觉词汇的随机性，可将多个 LSH 函数生成的词汇进行聚类集成生成最终的视觉词典，提高视觉词典稳定性的同时，进一步减弱视觉单词的同义性和歧义性，并且词典的大小也能够在聚类集成过程中自动确定。综上所述，本文提出一种基于随机化视觉词汇和聚类集成的目标分类方法。首先，该方法采用 E2LSH 对训练图像库的 SIFT 特征点进行哈希映射聚类，生成一组随机化视觉词汇；然后，对这组随机化视觉词汇进行聚类集成，生成随机化视觉词汇集成词典(RVVAD)；最后，基于该词典构建图像的视觉单词直方图并使用 SVM 分类器完成目标分类。

本文剩余部分组织如下。第 2 节简要介绍了 E2LSH 和聚类集成的定义和基本原理；第 3 节给出了基于随机化视觉词汇和聚类集成的目标分类方法涉及的关键技术，其中着重介绍了基于 E2LSH 的随机化视觉词汇生成算法和随机化视觉词汇集成算法；第 4 节对本文方法进行了实验验证和性能分析；第 5 节是结束语。

## 1　相关知识

### 1.1 E2LSH 哈希原理

E2LSH 的基本思想是利用基于 $p$-稳定分布的位置敏感函数对高维数据进行降维映射，将距离相近的点以较大的概率哈希到同一个桶中，将距离较远的点以较大的概率哈希到不同的桶中。位置敏感函数以及 $p$-稳定分布的定义见文献[16]。E2LSH 中定义位置敏感函数为：

$$h_{\alpha,\beta}(\boldsymbol{v})=\left\lfloor \frac{\boldsymbol{\alpha}\cdot\boldsymbol{v}+\beta}{\varpi} \right\rfloor \tag{1}$$

式中：$\lfloor\cdot\rfloor$为向下取整操作；$\boldsymbol{\alpha}$是从 $p$-稳定分布函数中随机抽样得到的 $d$ 维向量；$\beta$ 是在$[0,\varpi]$上均匀分布的随机

变量。该函数满足位置敏感特性，并且将 $d$ 维向量 $\boldsymbol{v}$ 映射为一个整数，实现了降维功能。将多个独立的位置敏感函数联合使用，可进一步拉大距离近的点映射后碰撞概率与距离远的点映射后碰撞概率之间的差距。定义函数集：

$$\mathcal{G}=\{g:R^d\to Z^k\} \tag{2}$$

式中 $g(\boldsymbol{v})=(h_1(\boldsymbol{v}),h_2(\boldsymbol{v}),\cdots,h_k(\boldsymbol{v}))$。则函数 $g(\boldsymbol{v})$ 可将 $R^d$ 空间中的任意数据点 $\boldsymbol{v}$ 哈希映射为一个 $k$ 维向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_k)$，其每一维元素均为整数值。最后，使用主哈希函数 $h_1$ 和次哈希函数 $h_2$ 对 $\boldsymbol{a}$ 进行哈希，即可建立一个哈希表，$h_1$ 和 $h_2$ 分别为：

$$h_1(\boldsymbol{a})=\left(\left(\sum_{i=1}^{k}r_i'a_i\right)\bmod prime\right)\bmod\ tablesize \tag{3}$$

$$h_2(\boldsymbol{a})=\left(\sum_{i=1}^{k}r_i''a_i\right)\bmod prime \tag{4}$$

式中：$r_i'$ 和 $r_i''$ 是 2 个随机整数；$tablesize$ 是哈希表的大小；$prime$ 是一个大的素数，通常取值 $2^{32}-5$。将主哈希值 $h_1$ 和次哈希值 $h_2$ 都相同的数据点存储到哈希表的同一个桶中，即可实现对数据点的聚类划分。

由上述分析可知，对于函数族 $\mathcal{G}$ 中的每个函数 $g(\boldsymbol{v})$，都可以通过上述桶哈希机制建立一个哈希表。将哈希表中每个桶看成是一个视觉单词，那么哈希表中所有的桶就可以构成一个视觉词汇。由于函数 $g(\boldsymbol{v})$ 的选取具有较强的随机性，因此通过其建立的视觉词汇也带有较强的随机性，称之为随机化视觉词汇。单个随机化视觉词汇由于其内在的随机性不能作为视觉词典使用，为此，本文从 $\mathcal{G}$ 中选取 $L$ 个独立的函数 $g_1,g_2,\cdots,g_L$，建立 $L$ 个哈希表，得到 $L$ 个相互独立的随机化视觉词汇，然后聚类集成这 $L$ 个词汇，生成最终的视觉词典。

### 1.2 聚类集成原理

聚类集成[19]是指，给定一个初始聚类结果集合，寻找一个聚类使其对于所有的初始聚类，尽可能多地符合或一致。聚类集成作为一个最优化问题，是利用多个初始聚类结果找到一个全新的聚类，使该聚类能够在最大程度上综合利用所有初始聚类结果提供的对数据集的聚类信息，提高聚类结果的质量和聚类的鲁棒性。同时，聚类集成也可自动检测确定最优的簇个数，并能够检测出孤立点。

对于聚类集成问题，Gionis 等[19]给出了一种基于相关聚类(Correlation Clustering)[20]的解决方案。对于数据集 $X=\{x_1,x_2,\cdots,x_n\}$，给定 $m$ 个初始聚类结果 $\{C_1,C_2,\cdots,C_m\}$，聚类集成的目标是寻找一个最优的聚类结果 $C$ 使其与 $m$ 个初始聚类尽可能地符合或一致。对于数据集 $X$ 中的任意 2 个数据 $u$ 和 $v$，可以定义如下一种距离度量：

$$d(u,v)=\frac{1}{m}\cdot\left|\{i\,|\,C_i(u)\neq C_i(v),1\leqslant i\leqslant m\}\right| \tag{5}$$

式中：$C_i(u)$ 为数据 $u$ 在聚类 $C_i$ 所属的簇序号，$C_i(u)\neq C_i(v)$ 指数据 $u$ 和 $v$ 在聚类 $C_i$ 中被分到不同的簇中；$|\bullet|$ 表示集合中元素的个数，即集合的势。$d(u,v)$ 的含义为，将数据 $u$ 和 $v$ 划分到不同簇中的初始聚类在 $m$ 个初始聚类中所占的比重。聚类集成的目标是寻找聚类 $C$，使得下面的函数取值最小：

$$d(C)=\sum_{C(u)=C(v)}d(u,v)+\sum_{C(u)\neq C(v)}(1-d(u,v)) \tag{6}$$

由上述分析可知，聚类集成通过综合利用多个初始聚类结果，能够降低聚类过程中的随机性，提高聚类质量。因此，将聚类集成技术应用到随机化视觉词汇集成，减弱视觉词汇的随机性，不但能够提高构建视觉词典稳定性，而且能进一步减弱视觉单词的同义性和歧义性，同时在词汇集成的过程中视觉词典的规模也能够被自动确定为合适大小。

## 2 基于随机化视觉词汇和聚类集成的目标分类

针对传统聚类算法效率低，视觉单词同义性、歧义性和聚类结果不稳定问题，本文引入 E2LSH 哈希映射聚类和聚类集成方法，构建性能更为突出的视觉词典，并将其用于图像目标分类，如图 1 所示。

该方法流程为：首先，提取训练图像数据库中所有图像的 SIFT 特征，利用 E2LSH 对 SIFT 点进行哈希映射，得到一组随机化视觉词汇，对这组随机化视觉词汇进行聚类集成，生成随机化视觉词汇集成词典；然后，基于该词典采用软分配的方法构建图像的视觉单词直方图；最后，利用该视觉单词直方图数据训练 SVM 分类器。当有待分类的图像时，提取该图像的 SIFT 点并与视觉单词匹配，构建该图像的视觉单词直方图，并输入 SVM 分类器完成对该图像目标的分类。

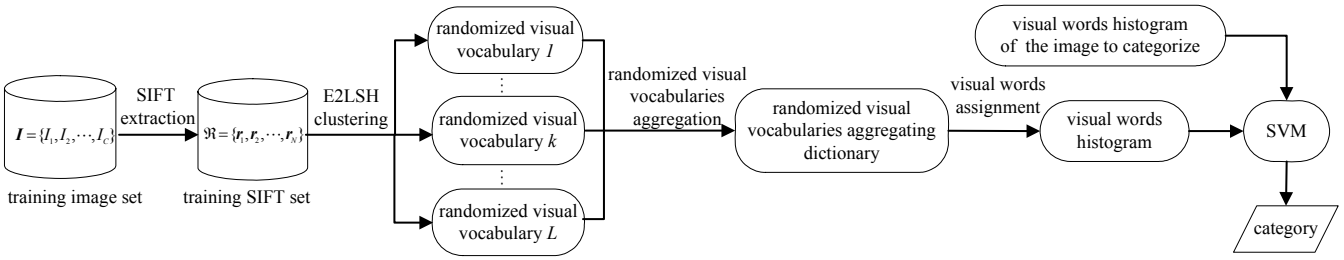



Fig.1 System diagram
图 1 系统框图

### 2.1 基于 E2LSH 的随机化视觉词汇生成

由 1.1 节的分析可知，为减弱通过聚类生成视觉词典中视觉单词的同义性和歧义性，本文采用 E2LSH 替代传统的 K-Means 聚类方法对训练 SIFT 特征集进行哈希聚类，生成一个随机化视觉词汇。算法的具体流程如下：

1) 提取训练图像库中所有图像的 SIFT 特征，构成训练特征库 $\mathfrak{R}=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_i,\cdots,\boldsymbol{r}_{N-1},\boldsymbol{r}_N\}$，其中，点 $\boldsymbol{r}_i$ 是 SIFT 特征，$N$ 为训练库中 SIFT 特征的总数；

2) 由式(2)从函数集 $\mathcal{G}$ 中随机选取一个位置敏感函数 $g$，作为哈希函数对 SIFT 特征集进行哈希映射划分；

3) 利用步骤 2)中选取的哈希函数 $g$ 对特征库 $\mathfrak{R}$ 进行哈希映射，得到 $\mathfrak{R}$ 中的每个 SIFT 特征 $\boldsymbol{r}_i$ 对应的 $k$ 维向量 $g(\boldsymbol{r}_i)$；

4) 计算 $\boldsymbol{r}$ 的主哈希值 $h_1(g(\boldsymbol{r}_i))$ 和次哈希值 $h_2(g(\boldsymbol{r}_i))$；

5) 对特征库 $\mathfrak{R}$ 中的所有 SIFT 向量，将主、次哈希值都相同的 SIFT 特征放入同一个桶，即可构成一个哈希表 $T_g=\{b_1,b_2,\cdots,b_k,\cdots,b_{Z-1},b_Z\}$，其中 $b_k$ 指哈希表中第 $k$ 个桶，$Z$ 表示哈希表中桶的总个数。则哈希表 $T_g$ 是对 SIFT 特征库 $\mathfrak{R}$ 的一个划分，将表中的每个桶中心看作一个视觉单词，那么所有桶就构成了视觉词汇。

由上述过程可知，由于函数 $g$ 的选取是随机的，生成的视觉词汇也具有较强的随机性，称为随机化视觉词汇，记为 $V=\{w_1,w_2,\cdots,w_k,\cdots,w_{Z-1},w_Z\}$，其中 $w_k$ 为 $b_k$ 对应的视觉单词。利用 E2LSH 哈希映射对 SIFT 特征库进行聚类生成的视觉词汇，能够较好地抑制视觉单词的同义性和歧义性，具体实验分析见第 3 节。

### 2.2 随机化视觉词汇集成

不同的聚类结果从不同的角度反映了数据集合的结构，聚类集成能够综合利用多个聚类结果，提高聚类结果的质量和聚类的鲁棒性，并且能够自动检测最合适的簇个数。由于 E2LSH 哈希映射聚类生成视觉词汇过程中，位置敏感函数的选择具有很强的随机性，这种随机性会降低视觉词典的性能；另一方面，不同的词汇又能够从不同的侧面反映特征库中 SIFT 特征的空间分布结构，可以考虑综合利用多个随机化视觉词汇，提高词典的表达能力。因此，本文从函数族 $\mathcal{G}$ 中随机选取 $L$ 个函数 $g_1,g_2,\cdots,g_L$，采用 2.1 节中方法生成 $L$ 个独立的随机化视觉词汇 $V_1,V_1,\cdots,V_L$。对该组词汇进行聚类集成，即可构建最终的视觉词典，具体过程如下：

1) 构建特征库的无向连通图。对 2.1 节中的图像特征库 $\mathfrak{R}=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_i,\cdots,\boldsymbol{r}_{N-1},\boldsymbol{r}_N\}$ 中的任意 2 个 SIFT 点 $\boldsymbol{r}_i$ 和 $\boldsymbol{r}_j$，由式(5)计算它们之间的距离 $d(\boldsymbol{r}_i,\boldsymbol{r}_j)$，具体为：

$$d(\boldsymbol{r}_i,\boldsymbol{r}_j)=\frac{1}{L}\left|\left\{k\,|\,V_k(\boldsymbol{r}_i)\neq V_k(\boldsymbol{r}_j),1\leqslant k\leqslant L\right\}\right| \tag{7}$$

其中 $V_k(\boldsymbol{r}_i)\neq V_k(\boldsymbol{r}_j)$ 的含义为 SIFT 点 $\boldsymbol{r}_i$ 和 $\boldsymbol{r}_j$ 在词汇 $V_k$ 中被划分到 2 个不同的视觉单词。则可以构建一个以 SIFT 特征点为顶点，以两两 SIFT 特征点之间相似性距离 $d(\boldsymbol{r}_i,\boldsymbol{r}_j)$ 为边权重的无向连通图 $\boldsymbol{G}$；

2) SIFT 特征点排序。针对每个 SIFT 特征点 $\boldsymbol{r}_i$，计算所有与其相连的边的权重之和 $\omega_i$，并以该权重从小到大的顺序排列 $\mathcal{R}$ 中的所有 SIFT 特征点，记该排列为 $\pi$；

3) 选取 $\pi$ 中第一个未被聚类的特征点 $\boldsymbol{u}$，找到所有满足 $d(\boldsymbol{u},\boldsymbol{r})\leqslant 1/2$ 且未被聚类的 SIFT 点，记为集合 $\boldsymbol{B}$；

4) 计算 $\boldsymbol{u}$ 与 $\boldsymbol{B}$ 中所有 SIFT 点的平均距离 $d(\boldsymbol{u},\boldsymbol{B})$：

$$d(\boldsymbol{u},\boldsymbol{B})=\frac{1}{|\boldsymbol{B}|}d(\boldsymbol{u},\boldsymbol{r}),\ \boldsymbol{r}\in\boldsymbol{B},\ |\boldsymbol{B}|\text{为集合的势} \tag{8}$$

5) 若 $d(\boldsymbol{u},\boldsymbol{B})\leqslant\alpha$，则 $\boldsymbol{B}\cup\{\boldsymbol{u}\}$ 构成一个簇，并标记 $\boldsymbol{B}$ 中所有的 SIFT 点为已被聚类，否则 $\{\boldsymbol{u}\}$ 构成一个簇；

6) 更新无向连通图 $\boldsymbol{G}$ 和排列 $\pi$。删除 $\boldsymbol{G}$ 中所有已被聚类的 SIFT 点对应的顶点及与其相连接的边，删除 $\pi$ 中所有已被聚类的 SIFT 点；

7) 重复步骤 3)~5)，直到所有 SIFT 点都被聚类，得到对特征库 $\mathfrak{R}$ 的聚类结果 $C=\{c_1,c_2,\cdots,c_k,\cdots,c_{K-1},c_K\}$，其中 $K$ 为簇的个数，$c_k$ 为第 $k$ 个簇。计算每个簇中心，即可得到视觉词典 $W=\{w_1,w_2,\cdots,w_k,\cdots,w_{K-1},w_K\}$，其中 $w_k$ 为第 $k$ 个视觉单词；

8) 视觉单词过滤。包含数据点太少或太多的视觉单词所携带信息的区分性往往不大，在信息损失很小的前提下，可以将这些视觉单词滤除掉，最终得到包含 $M$ 个视觉单词的视觉词典 $W^*=\{w_1,w_2,\cdots,w_k,\cdots,w_{M-1},w_M\}$。

上述通过聚类集成 $L$ 个独立的随机化视觉词汇生成的词典，称为随机化视觉词汇集成词典(RVVAD)，该词典相比传统的通过 K-Means 聚类得到的词典，有 4 个方面的优势：1) 构建视觉词典的时间复杂度低；2) 运用 E2LSH 降维映射生成的随机化视觉词汇有效地降低了视觉单词的同义性和歧义性；3) 通过引入聚类集成技术，综合利用多个随机化视觉词汇的特性，减弱了视觉词典生成过程中的随机性，提高了词典的表达能力；4) 视觉词汇集成过程中，视觉词典的规模大小能够自动确定。

### 2.3 视觉单词直方图构建

将每幅图像中的 SIFT 特征与视觉词典中单词进行匹配，统计每个单词的频次，即可生成 $M$ 维的直方图 $\boldsymbol{H}=[h_1,h_2,\cdots,h_M]$ 作为该图像的特征向量。传统的视觉词典法采用硬分配(Hard Assignment)的方法，即每个 SIFT 特征点只分配给一个视觉单词，然而由于视觉单词并非有明确语义含义的实体，视觉单词之间没有严格的语义区分，硬分配方法不能充分发挥视觉词典的性能。虽然本文采用 E2LSH 降维映射和聚类集成构建视觉词典，能够有效降低视觉单词的同义性和歧义性，但是视觉单词语义模糊性的问题仍然存在，所以，这里采用基于视觉单词不确定性(Visual Word Uncertainty)的软分配方法生成图像的视觉单词直方图。

该方法将高斯核植入到 SIFT 点空间，并假设 SIFT 点和视觉单词之间的距离变量服从高斯分布，即：

$$K_\sigma(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{9}$$

式中 $\sigma$ 为高斯核的大小。

对于视觉词典中的任意单词 $\omega_i$，其在图像 $I$ 中的分布可表示为：

$$h_i=h(\omega_i)=\frac{1}{n}\sum_{k=1}^{n}\frac{K_\sigma(D(\omega_i,\boldsymbol{r}_k))}{\sum_{j=1}^{M}K_\sigma(D(\omega_j,\boldsymbol{r}_k))} \tag{10}$$

式中：$D(\omega,\boldsymbol{r})$ 为视觉单词 $\omega$ 与 SIFT 点 $\boldsymbol{r}_k$ 之间的欧式距离；$n$ 为图像 $I$ 中的 SIFT 点个数；$M$ 为视觉词典的大小。则图像 $I$ 视觉单词直方图可表示为 $\boldsymbol{H}=[h_1,h_2,\cdots,h_M]$。

## 3 实验验证

本文实验数据采用 Caltech-101 数据集[21]和 PASCAL VOC 2007 数据集[22]。Caltech-101 数据集总共包含 8 677 幅图像，分为 101 个目标类别，每个类别包含 31 至 800 幅图像。选取其中的 10 个目标进行实验，分别为 Brain, Butterfly,Ewer,Grand piano,Helicopter,Kangaroo,Laptop,Menorah,Starfish,Sunflower，每个目标选取 50 幅图像构成训练集，30 幅图像用作测试集。PASCAL VOC 2007 数据集共包含 9 963 幅图像，分为 20 个类别，被广泛应用于图像目标识别、目标检测实验。分别选取 trainval 子集和 test 子集作为训练集和测试集。实验采用 SVM 分类器，采用“一对多”方法进行多类别分类。实验硬件配置为一台 Core 3.1G×4 CPU,4G 内存的台式机。为获取可靠的实验结果，每个实验重复 10 次，每次从图像库中随机选取不同的训练集和测试集。

分类性能评价指标为各类别的平均准确率(Average Precision，AP)和所有类别的平均准确率均值(Mean Average Precision，MAP)，定义如下：

$$\text{平均准确率}=\frac{\text{各次实验分类准确率之和}}{\text{实验次数}}\times 100\% \tag{11}$$

$$\text{平均准确率均值}=\frac{\text{各类别平均准确率之和}}{\text{类别个数}}\times 100\% \tag{12}$$

本文首先在二维空间上对比了 E2LSH 和 K-Means 聚类方法对数据点的聚类效果，实验结果如图 2 所示。从图中可以看出，使用 K-Means 聚类算法得到的聚类中心的分布特点为，在数据点密集的区域聚类中心多，而在稀疏区域的聚类中心少，使用 E2LSH 聚类得到的聚类中心的分布较为均匀。同理，在 SIFT 特征空间，SIFT 点密集的区域其代表的视觉内容相似性大，用较少的视觉单词表达该视觉内容即可，而 SIFT 点稀疏区域其代表的

视觉内容差异性大，需要用多个视觉单词分别表达对应的视觉内容。而 E2LSH 聚类方法相较于 K-Means 方法，其聚类中心分布能够更好地满足上述要求，因此也能够更好地抑制视觉单词的同义性和歧义性问题，提高视觉单词的代表能力。

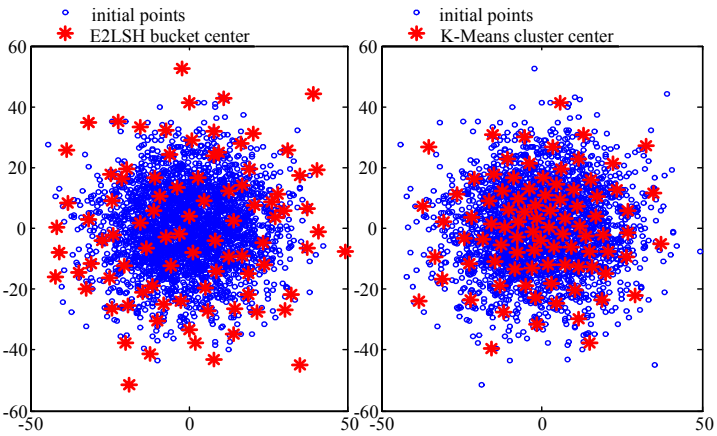
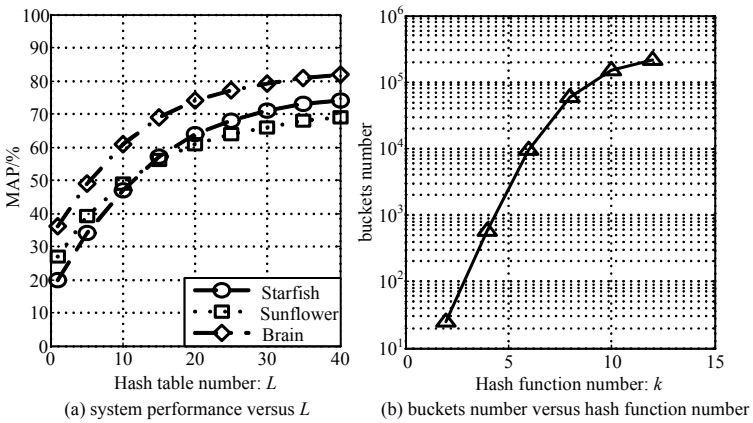



Fig.2 Clustering results by E2LSH and K-Means
图 2 E2LSH 和 K-Means 的聚类结果

为分析 E2LSH 聚类算法中参数 $L$ 和 $k$ 对算法性能的影响，本文针对 Caltech-101 的图像库，分别对不同的 $L$ 和 $k$ 值下对应的视觉词典进行性能测试，分析其对目标分类准确率的影响，实验结果如图 3 所示。从图 3(a)可以看出，目标分类的准确率随 $L$ 值的增大而提高，这是因为 $L$ 值对应随机化视觉词汇的个数，$L$ 值越大，生成的随机化视觉词汇个数越多，越能从多个方面挖掘特征空间中 SIFT 点的空间分布结构，集成得到的视觉词典随机性就越小。但另一方面，若 $L$ 值过大，算法的效率会降低。$k$ 值对应哈希降维后特征向量的维数，其对哈希表中桶的个数具有较大的影响，$k$ 值越大，哈希表中桶的个数越多，哈希聚类花费的时间也越多。综合考虑到算法的精确度和效率，本文取 $L=20$，$k=8$。




(a) system performance versus $L$　(b) buckets number versus hash function number

Fig.3 System performance versus E2LSH parameters
图 3 E2LSH 参数对性能的影响

然后，本文进一步对比了本文方法和 AKM 算法[3]在构建视觉词典的时间效率。从 Caltech-101 图像库中随机选取 1 000 幅图像，提取出约 1 530 000 个 SIFT 特征点构成 SIFT 特征库，分别采用本文方法和 AKM 算法进行聚类，生成视觉词典，2 种算法的时间效率如图 4 所示。2 种方法的时间消耗都随着视觉词典规模增大而增加，而 AKM 的时间复杂度与特征点数量成正比[3]，而本文方法受特征点数量的影响较小。因此，本文方法在图像规模增大的情况下，依然可以高效地构建大规模的视觉词典，具有更强的可扩展性。

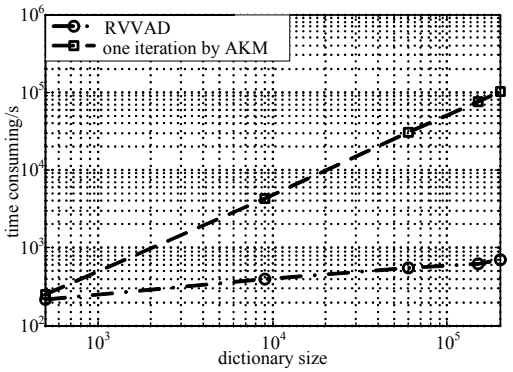



Fig.4 Time-consuming performance of RVVAD and AKM
图 4 本文方法与 AKM 方法的时间效率对比图

为验证本文方法(记为 RVVAD)的目标分类性能，将其与经典的基于 AKM 和软分配的方法[7](记为 AKM+SA)，基于互最近邻聚类(Reciprocal Nearest Neighbor Clustering)和相关聚类优化的方法[8](记为 RNNC+CC)，随机化位置敏感词典方法[15](记为 RLSV)及基于 AKM 和视觉单词集成的软分配方法[18](记为 AKM+VWA+SA)，在 Caltech-101 数据库上进行目标分类的性能对比实验，实验结果如表 1 所示。

从表 1 可以看出，AKM+SA 方法的平均准确率均值(MAP)均低于其他 4 种方法。RNNC+CC 方法利用训练图像的类别标签在生成视觉词典的过程中加入了监督机制，提高了视觉单词的区分能力，其 MAP 值相比 AKM+SA 方法有了较大的提高。AKM+VWA+SA 方法由于聚类集成多个粗糙的 K-Means 聚类结果，减小了初始聚类中心对聚类结果的影响，提高视觉词典的质量，其 MAP 值相比 AKM+SA 方法也有较大提高，但由于该方法未能有效处理视觉单词同义性和歧义性的问题，其 MAP 值相比 RVVAD 还有较大差距。RLSV 方法利用 LSH 生成位置敏感词典，抑制了视觉单词的同义性和歧义性问题，其 MAP 值高于前 3 种方法，但该方法构建的视觉词典存在较强的随机性，与加入了聚类集成的 RVVAD 方法相比还有差距。本文方法在利用 E2LSH 生成随机化视觉词汇的基础上又对其进行聚类集成，生成随机化视觉词汇集成词典，既降低了视觉单词的同义性和歧义性，又综合利用多种视觉词汇更好地挖掘 SIFT 点空间分布特性，降低了词典的随机性，提高了词典的表达能力，其 MAP 值高于其他几种方法。

因此，采用本文方法能够构建具有较强表达能力的视觉词典，提高目标分类的准确率。

表 1 Caltech-101 数据集实验结果

Table1 Experimental results on Caltech-101

| object | AKM+SA/% | RNNC+CC/% | AKM+VWA+SA/% | RLSV/% | RVVAD/% |
|---|---|---|---|---|---|
| Brain | 53.60 | 62.00 | 59.10 | 67.80 | 74.30 |
| Butterfly | 40.30 | 38.20 | 60.50 | 49.00 | 63.40 |
| Ewer | 54.80 | 43.40 | 43.50 | 52.60 | 56.80 |
| Grand piano | 68.50 | 73.30 | 61.40 | 56.50 | 72.30 |
| Helicopter | 46.70 | 64.50 | 63.70 | 74.20 | 75.60 |
| Kangaroo | 52.10 | 70.70 | 71.90 | 75.80 | 78.50 |
| Laptop | 65.60 | 59.80 | 68.80 | 66.40 | 72.10 |
| Menorah | 36.10 | 49.60 | 56.40 | 53.50 | 62.40 |
| Starfish | 47.80 | 60.50 | 58.00 | 58.00 | 63.10 |
| Sunflower | 55.40 | 63.20 | 60.30 | 68.50 | 70.80 |
| MAP | 52.09 | 58.52 | 60.36 | 62.23 | 68.93 |

此外，为更进一步验证本文方法的有效性，将其与其他方法在 PASCAL VOC 2007 图像库进行目标分类实验，结果如表 2 所示。由表 2 可知，本文提出的 RVVAD 方法的平均准确率达到了 62.10%，高于其他 4 种方法。图 5 给出了不同方法对各目标分类的平均准确率结果。对于 20 个目标类别，本文方法在其中 14 个类别当中的平均准确率最高，进一步验证了本文方法在降低视觉单词同义性、歧义性，挖掘特征点空间分布结构并提高视觉词典表达能力方面的优越性。

表 2 PASCAL VOC 2007 数据集实验结果

Table2 Experimental results on PASCAL VOC 2007

| | AKM+SA/% | AKM+VWA+SA/% | RLSV/% | INRIA_Genetic[22]/% | RVVAD/% |
|---|---|---|---|---|---|
| MAP | 58.22 | 58.90 | 59.07 | 59.40 | 62.10 |

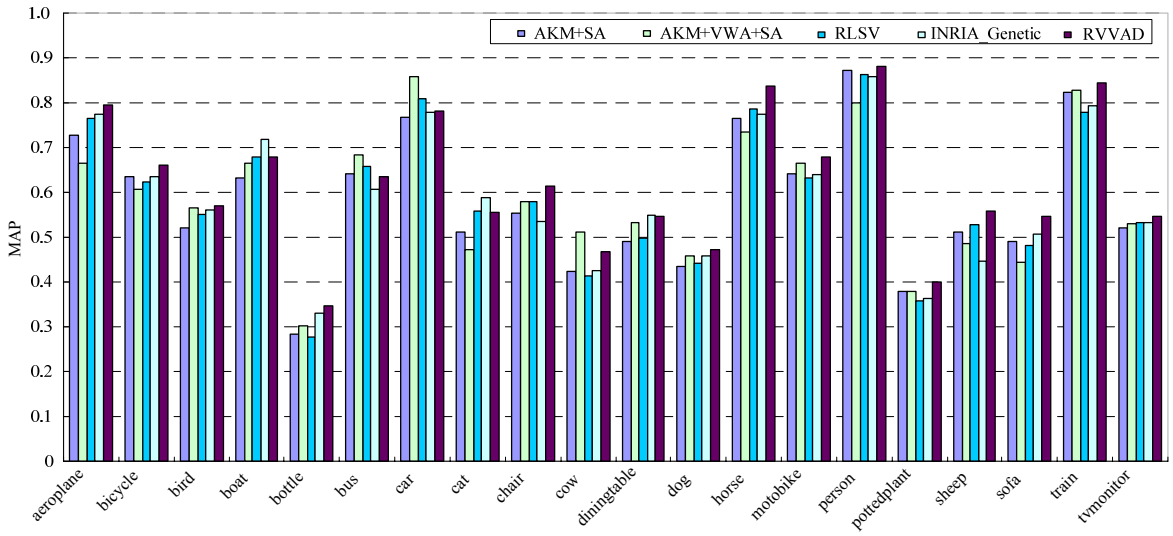



Fig.5 Object categorization performance analysis

图 5 不同方法的目标分类平均准确率

## 4 结论

针对传统视觉词典法中 3 个方面的问题，本文采用 E2LSH 算法和聚类集成技术构建随机化视觉词汇集成词典，并将其应用到图像目标分类。与传统方法相比，采用 E2LSH 进行哈希映射聚类生成随机化视觉词汇，有效地降低了时间和内存开销，并且能够在一定程度上克服视觉单词的同义性和歧义性。同时，引入聚类集成技术进行随机化视觉词汇的聚类集成，有效地降低了视觉词典的随机性，提高了词典的表达能力。下一步将考虑在生成视觉词典的过程中引入监督机制，以期望能构建区分性和语义代表性更强的词典。

**参考文献：**

[1] Sivic J,Zisserman A. Video Google: a text retrieval approach to object matching in videos[C]// Proceedings of 9th IEEE International Conference on Computer Vision. Nice:IEEE, 2003:1470-1477.

[2] Kersorn K. An Enhanced bag-of-visual word vector space model to represent visual content in athletics images[J]. IEEE Transactions on Multimedia, 2012,14(1):211-222.
[3] Philbin J. Scalable object retrieval in very large image collections[D]. Oxford:University of Oxford, 2010.
[4] Wang J,Wang J D,Ke Q F,et al. Fast approximate K-means via cluster closures[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington,DC,USA:IEEE Computer Society, 2012:3037-3044.
[5] WU J,CUI Z M,ZHAO P P,et al. Visual vocabulary tree construction research using adaptive fuzzy k-means clustering[J]. Advanced Science Letters, 2012,11(1):258-262.
[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2):91-110.
[7] Van Gemert J C,Veenman C J,Smeulders A W M,et al. Visual word ambiguity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010,7(32):1271-1283.
[8] López-Sastre R J,Tuytelaars T,Acevedo-Rodríguez F J,et al. Towards a more discriminative and semantic visual vocabulary[J]. Computer Vision and Image Understanding, 2011,115(3):415-425.
[9] WANG J Y,LI Y P,ZHANG Y,et al. Bag-of-features based medical image retrieval via multiple assignment and visual words weighting[J]. IEEE Transactions on Medical Imaging, 2011,30(11):1996-2011.
[10] YU S,Jurie F. Visual word disambiguation by semantic contexts[C]// Proceedings of 13th IEEE International Conference on Computer Vision. Barcelona:IEEE, 2011:311-318.
[11] ZHANG C J,LIU J,WANG J Q,et al. Image classification using spatial pyramid coding and visual word reweighting[J]. Lecture Notes in Computer Science, 2011,6494:239-249.
[12] YANG Z G,PENGY X,XIAO J G. Visual vocabulary optimization with spatial context for image annotation and classification[C]// Proceedings of 18th International Conference on Advances in Multimedia Modeling. Klagenfurt: Springer-Verlag, 2012:89-102.
[13] Beyer K S,Goldstein J,Ramakrishnan R,et al. When is "nearest neighbor" meaningful?[C]// Proceedings of 7th International Conference on Database Theory. Jerusalem:Springer-Verlag, 1999:217-235.
[14] Moosmann F,Nowak E,Jurie F. Randomized clustering forests for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008,30(8):1632-1646.
[15] MUY D,SUN J,HAN T Y,et al. Randomized locality sensitive vocabularies for bag-of-features model[C]// Proceedings of 11th European Conference on Computer Vision. Berlin:Springer-Verlag, 2010:748-751.
[16] Datar M,Immorlica N,Indyk P,et al. Locality-sensitive hashing scheme based on p-stable distributions[C]// Proceedings of the 20th Annual ACM Symposium on Computational Geometry. New York:ACM, 2004:253-262.
[17] Kontschieder P,Bulo S R,Bischof H,et al. Structured class-labels in random forests for semantic image labeling[C]// Proceedings of 13th IEEE International Conference on Computer Vision. Barcelona:IEEE, 2011:2190-2197.
[18] López-Sastre R J,Renes-Olalla J,Gil-Jiménez P,et al. Visual word aggregation[C]// Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis. Berlin:Springer-Verlag, 2011:676-683.
[19] Gionis A,Mannila H,Tsaparas P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1):1-30.
[20] Bansal N,Blum A,Chawla S. Correlation clustering[J]. Machine Learning, 2004,56:89-113.
[21] LI Fei-fei,Fergus R,Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision. Washington D C:IEEE, 2004.
[22] Everingham M,Van Gool L,Williams C K I,et al. The PASCAL Visual Object Classes Challenge 2007(VOC 2007) Results[J]. International Journal of Computer Vision, 2010,88(2):303-338.

**作者简介:**

**朱道广**(1988-),男,河南省许昌市人,在读硕士研究生,主要研究方向为图像分类、图像检索.email:zhudg06@163.com.

**李弼程**(1970-),男,湖南省衡阳市人,博士,教授,博士生导师,研究方向为智能信息处理.