基于弱监督 E2LSH 和显著图加权的目标分类方法

赵永威* 李弼程 柯圣财
(解放军信息工程大学信息系统工程学院 郑州 450002)

摘 要：在目标分类领域，当前主流的目标分类方法是基于视觉词典模型，而时间效率低、视觉单词同义性和歧义性及单词空间信息的缺失等问题严重制约了其分类性能。针对这些问题，该文提出一种基于弱监督的精确位置敏感哈希(E2LSH)和显著图加权的目标分类方法。首先，引入 E2LSH 算法对训练图像集的特征点聚类生成一组视觉词典，并提出一种弱监督策略对 E2LSH 中哈希函数的选取进行监督，以降低其随机性，提高视觉词典的区分性。然后，利用 GBVS(Graph-Based Visual Saliency)显著度检测算法对图像进行显著度检测，并依据单词所处区域的显著度值为其分配权重；最后，利用显著图加权的视觉语言模型完成目标分类。在数据集 Caltech-256 和 Pascal VOC 2007 上的实验结果表明，所提方法能够较好地提高词典生成效率，提高目标表达的分辨能力，其目标分类性能优于当前主流方法。

关键词：目标分类；视觉词典模型；精确位置敏感哈希；视觉显著图；视觉语言模型
中图分类号： TP391 **文献标识码**： A **文章编号**：1009-5896(2016)01-0038-09
DOI: 10.11999/JEIT150337

Object Classification Method Based on Weakly Supervised E2LSH and Saliency Map Weighting

ZHAO Yongwei LI Bicheng KE Shengcai
(*Institute of Information System Engineering*, *Information Engineering University*, *Zhengzhou* 450002, *China*)

Abstract: The most popular approach in object classification is based on the bag of visual-words model. However, there are several fundamental problems that restricts the performance of this method, such as low time efficiency, the synonym and polysemy of visual words, and the lack of spatial information between visual words. In view of this, an object classification method based on weakly supervised Exact Euclidean Locality Sensitive Hashing (E2LSH) and saliency map weighting is proposed. Firstly, E2LSH is employed to generate a group of visual dictionary by clustering SIFT features of the training dataset, and the selecting process of hash functions is effectively supervised inspired by the random forest ideas to reduce the randomcity of E2LSH. Secondly, Graph-Based Visual Saliency (GBVS) algorithm is applied to detect the saliency map of different images and visual words are weighted according to the saliency prior. Finally, saliency map weighted visual language model is carried out to accomplish object classification. Experimental results on datasets of Caltech-256 and Pascal 2007 indicate that the distinguishability of objects is effectively improved and the proposed method is superior to the state-of-the-art object classification methods.

Key words: Object classification; Bag of visual words model; Exact Euclidean Locality Sensitive Hashing (E2LSH); Visual saliency map; Visual language model

1 引言

视觉词典模型(Bag of Visual Words Model, BoVWM)[1–4]的出现迈出了由图像底层特征向高层视觉语义过渡的第 1 步。由于其性能优越，在图像分类[5]等领域的应用十分广泛，然而，以下几个关键问题的存在却极大地限制了其性能。首先是词典生成效率低，当前主要的词典生成算法(如 K-Means[1])在对特征点[6]聚类时都需要多次迭代高维近似近邻计算，随着数据量的增大时间效率会急剧下降。其次是传统聚类算法的初始聚类中心大都是随机生成的，导致聚类结果对噪声的鲁棒性较差且容易引起视觉单词同义性和歧义性问题[7]。此外，传统的视觉词典模型都面临视觉单词空间信息缺失的问题，极大地降低了该模型的语义表达能力。

收稿日期：2015-03-23；改回日期：2015-09-09；网络出版：2015-11-17
*通信作者：赵永威 zhaoyongwei369@163.com
基金项目：国家自然科学基金(60872142, 61301232)
Foundation Iterms: The National Natural Science Foundation of China (60872142, 61301232)

为了提高词典生成效率，文献[8]提出了一种分层 K-Means 聚类算法(Hierarchical K-Means, HKM)，可以在一定程度上降低算法时间复杂度，而引入 KD 树的近似 K-Means 算法[9](Approximate K-Means, AKM)也能提高词典生成效率。MU 等人[10]尝试将位置敏感哈希(LSH)用于视觉聚类，提出了一种利用位置敏感哈希函数的随机映射来代替传统聚类生成视觉词典的方法。CAO 等人[11]将 LSH 用在大量图像数据的聚类当中，极大地改善了检索效率。2012 年，XIA 等人[12]将多核学习与 LSH 相结合，并将其用于图像检索，有效地改善了系统的检索性能。张瑞杰等人[13]则在此基础上通过引入多个非线性核函数，提出了一种 E2LSH-MKL 词典生成方法，增强算法的稳定性。文献[10–13]能够有效地提高词典生成效率，然而，它们忽略了 LSH 中哈希函数本身的随机性，很难保证生成视觉词典的质量。

针对视觉单词同义性和歧义性问题带来的不利影响，诸多研究人员都对此展开研究，比如文献[14, 15]通过将语义相近的视觉单词组合为“视觉词组”的方法，然而，这种方法只是简单统计了视觉单词间的共发频率，忽略了单词间的空间信息。Philbin 等人[16]则通过将一个特征点分配至多个视觉单词并对不同的单词赋予不同权重的方式提出了软分配方法。GEMERT 等人[7]利用多个核函数实现了特征点与视觉单词的软映射，有效降低了视觉单词同义性和歧义性问题带来的量化误差严重的问题。WEINSHALL 等人[17]则将软分配策略与潜在狄里克雷分布模型相结合(Latent Dirichlet Allocation, LDA)，提出了一种软分配的 LDA 模型。

此外，研究人员为解决视觉单词空间信息缺失的问题同样做了大量工作，如空间金字塔匹配(Spatial Pyramid Matching, SPM)[18,19]就是一种简单却行之有效的方法；赵春晖等人[20]则在结合感兴趣区域提取与空间金字塔匹配原理提出了一种优化方法，能够在训练数据较少的情况下取得良好效果。文献[21]针对空间信息缺失及量化误差严重的问题，结合空间金字塔和潜在概率语义分析模型(Probabilistic Latent Semantic Analysis, PLSA)提出了一种稀疏编码多尺度空间的图像分类方法，在弥补空间信息的同时，提高了图像内容表达的鲁棒性。XIE 等人[22]则提出了一种几何短语联合(Geometric Phrase Pooling, GPP)策略，利用互补的单词建立几何视觉词组，捕捉单词的空间信息。此外，视觉语言模型(Visual Language Model, VLM)[23,24]也越来越多地被用来记录单词之间的空间信息。但是，该模型在估计视觉单词的概率分布时，没有考虑图像背景区域的视觉单词给目标内容表达带来的不利影响，因此，在有效利用单词空间关系的同时，引入了新的背景噪声。

针对上述问题，本文提出了基于弱监督 E2LSH 和显著图加权的目标分类方法。该方法首先利用 E2LSH[25]对 SIFT 特征进行聚类，并借鉴随机森林思想对哈希函数的选取进行有效地监督以降低 E2LSH 算法本身的随机性，生成弱随机化视觉词典组，提高词典生成效率，增强视觉词典的区分性和语义表达能力。然后，利用 GBVS(Graph-Based Visual Saliency)算法[26]对图像进行显著性检测，并根据图像区域显著度的不同为视觉单词分配相应的权重，弱化图像背景噪声的影响，最后，引入视觉语言模型完成目标分类。

2　精确位置敏感哈希(E2LSH)

位置敏感哈希(Locality Sensitive Hashing, LSH)被广泛用于解决大规模快速图像近邻搜索问题[27,28]，其基本思想是利用若干个哈希函数对高维特征进行降维映射，并使得原始空间中距离较近的点以较大的概率哈希到同一桶中，相距较远的点则概率较小，这里用桶的概念来代表距离较近的两个矢量哈希后的哈希值相同，其值表征为桶，其具体意义可以类比为传统聚类算法对特征聚类后形成视觉词典中的视觉单词。E2LSH中的哈希函数都是基于 p-稳态分布的，这里我们选取的是基于2-稳态分布的哈希函数，其定义为

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

其中，$\lfloor\cdot\rfloor$ 为向下取整操作，$\boldsymbol{\alpha}$ 是一个随机抽样得到的 d 维向量，β 为在区间 $[0,\varpi]$ 中均匀分布的随机变量。然而，一个哈希函数往往分辨力不强，因此，E2LSH 常选取 k 个哈希函数联合起来使用。定义函数族 $\mathcal{G}=\{g:S\to U^k\}$。其中，$g(\boldsymbol{v})=(h_1(\boldsymbol{v}),h_2(\boldsymbol{v}),\cdots,h_k(\boldsymbol{v}))$，对任一数据点 $\boldsymbol{v}\in\mathbb{R}^d$，经过 $g(\boldsymbol{v})\in\mathcal{G}$ 降维映射就能得到一个 k 维向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_k)$。然后，再利用主哈希函数 h_1 和次哈希函数 h_2 对向量 $\boldsymbol{a}$ 进行哈希，主次哈希函数 h_1,h_2 定义为

$$h_1(\boldsymbol{a})=\left(\left(\sum_{i=1}^{k}r_i^{'}a_i\right)\bmod m\right)\bmod s \tag{2}$$

$$h_2(\boldsymbol{a})=\left(\sum_{i=1}^{k}r_i^{''}a_i\right)\bmod m \tag{3}$$

其中，$r_i^{'}$ 和 $r_i^{''}$ 为随机整数，s 为图像特征点的总数，m 取值为 $2^{32}-5$。E2LSH 可以将主哈希值 h_1 和次哈希值 h_2 都相同的数据点哈希到同一个桶中，并以此实现图像特征点的聚类。然而，由于函数 $g(\boldsymbol{v})$ 中 h 函

数具有相当的随机性，因此，由其构建的词典随机性也较强。为了降低这种随机性，本文借鉴随机森林算法[29]思想提出一种弱监督策略。

3 基于弱监督 E2LSH 和显著图加权的目标分类

对训练图像集 $\mathcal{I}=\{\mathcal{I}_1,\mathcal{I}_2,\cdots,\mathcal{I}_k\}$ 而言，基于弱监督 E2LSH 和显著图加权的目标分类流程可由图 1 描述。具体步骤可描述如下：

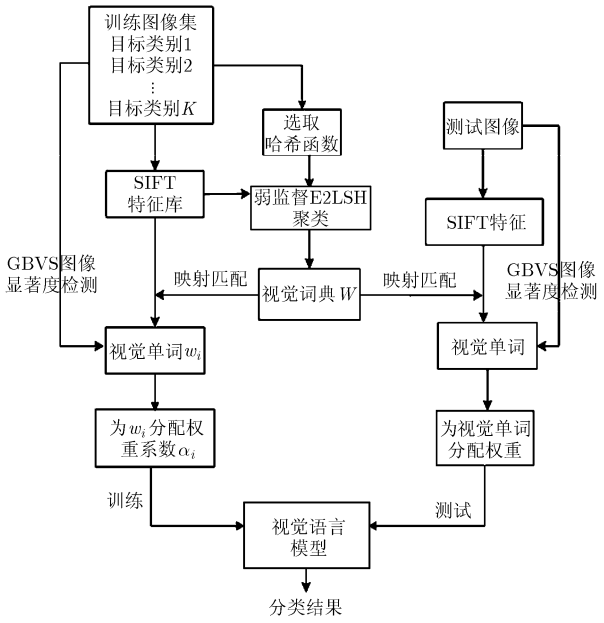


图 1 基于弱监督 E2LSH 和显著图加权的目标分类方法流程

步骤 1 提取训练图像库的 SIFT 特征，得训练图像特征库，记为 $\mathfrak{R}=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_N\}$，其中 $\boldsymbol{r}$ 代表一个 128 的 SIFT 特征，N 为特征总数；

步骤 2 利用本文算法，有监督地选取适用于该数据集的哈希函数，然后利用弱监督 E2LSH 对图像特征库 $\mathfrak{R}$ 进行哈希聚类，得视觉词典 W；

步骤 3 利用 GBVS 算法对每幅图像进行显著度检测，得各图像显著图，并给出不同显著区域的 SIFT 特征点以标识值；

步骤 4 根据视觉单词中各 SIFT 点所处区域显著度的不同，为单词分配相应的权重系数；

步骤 5 利用显著度加权视觉语言模型对训练图像类别和测试图像分别建模，然后计算得出测试图像类别。

3.1 弱监督 E2LSH 聚类

假设函数 g_i 中已经选取了 j 个哈希函数 $h_1,h_2,\cdots,h_j, 1\le j<k$，则对第 $j+1$ 哈希函数选取进行弱监督策略如下：

首先，按式(2)和式(3)计算 SIFT 特征点 $\boldsymbol{r}$ 的主、次哈希值 $h_1(g_i(\boldsymbol{r})),h_2(g_i(\boldsymbol{r}))$，并将主、次哈希值都相同的点哈希到同一桶中，得到初始视觉词典 $W(w_1,w_2,\cdots,w_{N_j})$；然后，计算词典中各单词 w_i 的香农熵如式(4)所示。

$$H_C(w_i)=-\sum_{l\in K}\frac{n_l}{n}\log_{w_{ij}}\frac{n_l}{n} \tag{4}$$

香农熵代表了由初始 j 个哈希函数聚类结果的信息量大小，它在一定意义上是一种相对熵增益，相对于绝对熵而言具有更好的稳定性与适应性。其中，代表了 n 为单词 w_i 所在哈希桶中的特征点总数，n_l 是属于目标类别 l 的特征点数目，然后，选取一个 h 函数作为候选第 $j+1$ 个哈希函数 $\hat{h}$，并计算其对 w_i 所在哈希桶的分裂熵为

$$H_S(\hat{h},w_i)=-\sum_{j=1}^{w_{ij}}\frac{n_j}{n}\log_{w_{ij}}\frac{n_j}{n} \tag{5}$$

这里，分裂熵代表了新选取的 h 函数对哈希桶也即是视觉单词分裂结果的信息量大小，值越大说明分裂结果的不确定性越大。假设 $\hat{h}$ 将 w_i 所在的哈希桶分裂为 w_{ij} 个，n_j 为分裂桶 w_{ij} 中的特征点数。通过计算上述分裂结果的互信息 $I_i(\hat{h})=H_C(w_i)-\sum_{j=1}^{w_{ij}}\frac{n_j}{n_i}H_C(w_{ij})$，这里互信息是指上述分裂结果的绝对熵增益，其值越大说明选取的 h 函数与前 j 个哈希函数之间的区分性更强。如此，就能得到候选哈希函数 $\hat{h}$ 与前 j 个哈希函数 $h_1,h_2,\cdots,h_j$ 的差别得分。

$$S_{j+1}(\hat{h})=\frac{1}{N_j}\sum_{i=1}^{N_j}\frac{2I_i(\hat{h})}{H_C(w_i)+H_S(\hat{h},w_i)} \tag{6}$$

在得到若干个候选哈希函数 $\hat{h}$ 之后，根据准则：$h^*=\arg\max_{\hat{h}}S_{j+1}(\hat{h})$ 选出与 j 个哈希函数区分性最大的第 $j+1$ 个哈希函数。重复上述过程就可以选出 k 个代表性和区分性强的哈希函数，并以此减弱 E2LSH 聚类结果的随机性，增强各视觉单词的语义表达能力。那么，利用选取的 L 个函数 $g_1,g_2,\cdots,g_L$ 就能够生成 L 个视觉词典以构建弱随机化视觉词典组，具体流程如图 2 所示。

完成聚类之后，为了进一步提高词典质量，我们根据式(7)得出各视觉单词对目标内容表达的权重值，并以从小到大的方式去除一定数量的分辨力较弱的噪声单词，使得每个词典的规模为 M，即

$$W_i=\{w_1^{(i)},w_2^{(i)},\cdots,w_k^{(i)},\cdots,w_{M-1}^{(i)},w_M^{(i)}\},\ i=1,2,\cdots,L$$

3.2 显著图加权的视觉语言模型

文献[30]提出的算法是一种经典的显著性检测算法，文献[26]在文献[30]算法的基础上引入了图论的知识，用马尔科夫链生成显著图，提取的显著区域比文献[30]算法更加准确。图 3 给出采用两种不同检测算法对同一幅图像进行检测得到的显著图。

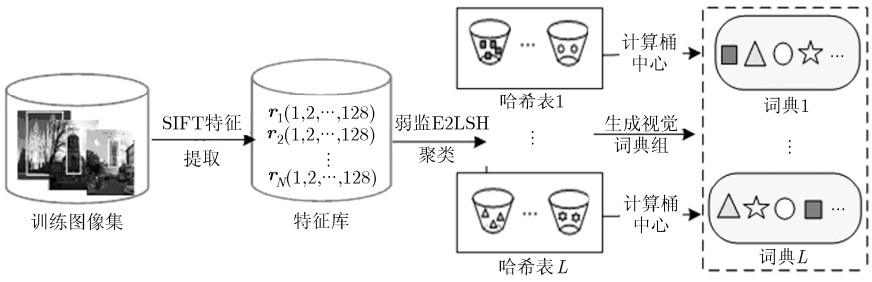
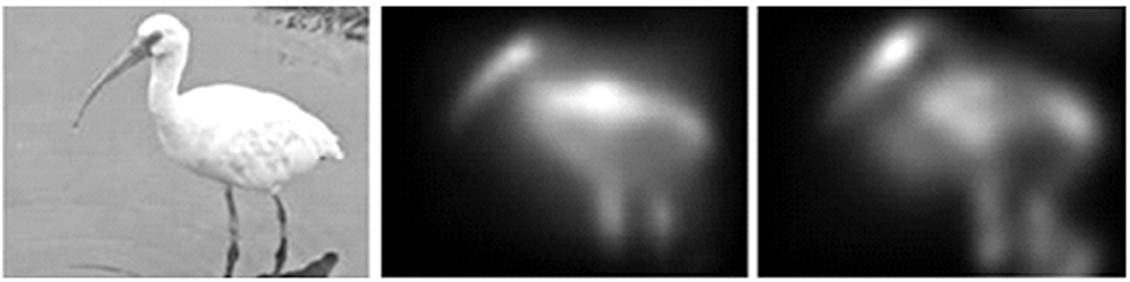


图 2 弱监督 E2LSH 生成视觉词典组



(a)原始图像　(b) GBVS算法得到的显著图　(c)文献[30]算法得到的显著图

图 3 利用不同显著度检测算法得到的显著图

从图 3 中不难看出，GBVS 算法能够检测到更为准确的目标前景区域。因此，本文采用一种基于 GBVS 显著图加权的视觉语言模型，进而在利用视觉语言模型建模时更好地为视觉单词分配相应的权重，提高目标分类准确率。

图 4 给出了代表 3 种显著度的图像特征点，方块点是属于目标前景的，三角点是属于目标背景区域的，而圆点则是介于目标前景和背景之间的。那么，由弱监督聚类生成得到的视觉单词就可以分配权值如下：

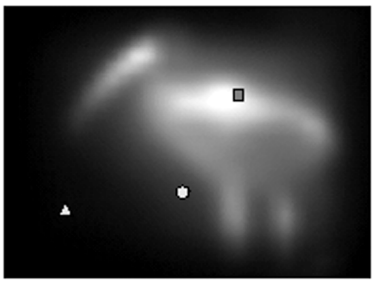


图 4 位于不同显著度位置的特征点示例

$$\alpha_k^{(i)} = \frac{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right) + 1/2 \underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right)}{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j\right)} \tag{7}$$

$$\beta_k^{(i)} = 1 - \alpha_k^{(i)} \tag{8}$$

其中，$w_k^{(i)}$ 对应第 i 个视觉词典 W_i 中的第 k 个视觉单词，$\alpha_k^{(i)}, \beta_k^{(i)}$ 分别为该单词的显著度权值和非显著度权值，$\boldsymbol{r}_j \in w_k^{(i)}$ 表示 $\boldsymbol{r}_j$ 为哈希到单词 $w_k^{(i)}$ 所在桶的 SIFT 特征点，y_j 为 SIFT 特征点 $\boldsymbol{r}_j$ 所处区域的标识值，且 $1 \le i \le L; 1 \le k \le M$。$y_j = 1$ 表示 $\boldsymbol{r}_j$ 处于显著图的目标前景区域，$\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right)$ 表示哈希到单词 $w_k^{(i)}$ 且处于显著图前景区域的 SIFT 特征数目；$y_j = 0$ 表示 $\boldsymbol{r}_j$ 处于显著图的前景与背景交界区域，$\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right)$ 表示哈希到单词 $w_k^{(i)}$ 且处于显著图的前景与背景交界区域的 SIFT 特征数目；$y_j = -1$ 则表示 $\boldsymbol{r}_j$ 处于背景区域。而为了避免显著权值 $\alpha_k^{(i)}$ 的情况出现，这里对式(7)作简单平滑如式(9)：

$$\alpha_k^{(i)} = \frac{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 1\right) + 1/2 \underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j \mid y_j = 0\right) + 1}{\underset{\boldsymbol{r}_j \in w_k^{(i)}}{\text{count}}\left(\boldsymbol{r}_j\right) + 1} \tag{9}$$

然后，就可采用视觉语言模型对图像进行建模，这里我们采用一元视觉模型，如式(10)所示。

$$p\left(w_k^{(i)} \middle| w_0^{(i)} w_1^{(i)} \cdots w_M^{(i)}\right) = p\left(w_k^{(i)}\right) \tag{10}$$

则每个不同权值的视觉单词 $w_k^{(i)}$ 的条件概率分布可由最大似然估计计算：

$$p\left(w_k^{(i)} \middle| C_t\right) = \frac{\alpha_k^{(i)} \cdot F_N\left(w_k^{(i)} \middle| C_t\right)}{\sum\limits_{w_k^{(i)} \in W_i} F_N\left(w_k^{(i)} \middle| C_t\right)} \tag{11}$$

其中 C_t 代表数据集中第 t 个目标类别，$F_N(w_k^{(i)} \mid C_t)$ 表示视觉单词 $w_k^{(i)}$ 在目标类别 C_t 中出现的频次。需要注意的是，由最大似然估计引起的零概率偏移问

题。通常采用数据平滑的方法解决零概率问题，使得所有的单词都有一个大于 0 的概率。这里，采用 J-M 平滑算法[23]，计算如式(12)：

$$p_\lambda\left(w_k^{(i)}\middle|C_t\right)=(1-\lambda)p\left(w_k^{(i)}\middle|C_t\right)+\lambda p\left(w_k^{(i)}\middle|\mathbb{C}\right) \quad (12)$$

其中，$p(w_k^{(i)}\,|\,\mathbb{C})$ 为视觉单词 $w_k^{(i)}$ 在训练图像集 $\mathbb{C}$ 中的概率分布，λ 为一个与图像无关的平滑参数，$\lambda\in[0,1]$。那么，对于一幅测试图像 I 而言，利用该图像在 L 个视觉词典下的最大似然概率均值就可以判断其归属于哪个类别，具体如式(13)：

$$C^*=\arg\max_{C_t}\frac{1}{L}\sum_{1}^{L}\prod_{w_k^{(i)}\in I}p\left(w_k^{(i)}\middle|C_t\right)p(C_t),$$
$$i=1,2,\cdots,L;\ \ t=1,2,\cdots,N_t \quad (13)$$

4 实验设置与性能评价

4.1 实验设置

本文分别在目标分类常用的 Caltech-256 图像集[31]和 Pascal Voc 2007 图像集[32]对本文方法性能进行评估。首先，随机选取 Caltech-256 图像集中的 airplanes breadmaker butrlerfly 等 8 个类别进行实验以验证文中各方法的有效性。其中，从每个类别中随机选取 80 幅图像构成训练图像集，其余作测试集，视觉词典规模为 1000。而为了获取可靠的实验结果，所有结果都是进行 10 次独立的图像分类实验平均得来。实验硬件配置为一台 Core 3.1G×4 CPU，内存为 4G 的台式机。目标分类性能评价指标为平均准确率(Average Precision, AP)、召回率(Recall)和以召回率为基础的混淆矩阵(Confusion Matrix)，相关定义为

$$\text{召回率}=\frac{\text{被正确分类的图像数}}{\text{该类图像总数}}\times100\% \quad (14)$$

$$\text{准确率}=\frac{\text{被正确分类的图像数}}{\text{被分类的图像总数}}\times100\% \quad (15)$$

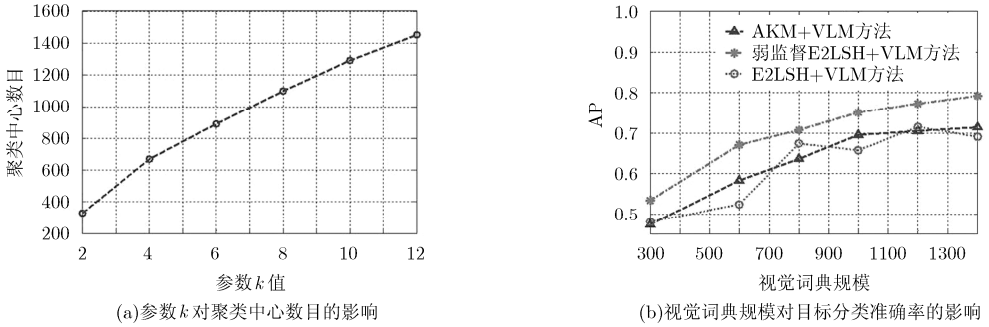
$$\text{平均准确率}=\frac{\text{各图像类别分类准确率之和}}{\text{图像类别总数}}\times100\% \quad (16)$$

4.2 实验结果

为了更好地分析弱监督 E2LSH 生成的视觉词典相较于初始 E2LSH 以及传统的近似 K-Means 算法的优越性，实验首先分别采用弱监督 E2LSH 及 E2LSH 对训练图像集聚类生成词典，得视觉单词数目随 k 变化情况如图 5(a)所示。然后，利用视觉语言模型(VLM)在这些视觉词典对 Caltech-256 图像集中的 8 个目标类别进行分类实验，得目标分类 AP 值随视觉词典规模变化情况如图 5(b)所示。从图 5(a)中可以看出，随着 k 值的增大，E2LSH 和弱监督 E2LSH 的聚类中心数目会随之增加；而从图 5(b)中不难看出针对不同数目的视觉词典，由弱监督 E2LSH 生成的视觉词典的目标分类性能均明显优于 AKM 算法生成的词典，而由初始 E2LSH 算法生成的视觉词典性能变化较大，随机性较强。综合考虑算法效率和性能，这里选取 k=8 。此外，图 6 则给出了参数 k=8 时，L 值对目标分类准确率的影响。从图 6 中可以看出参数 L 取值越大，目标分类准确性随之增加，然而，过大的 L 值会影响算法的效率，所以，本文中选取 L=5 。

然后，实验将弱监督 E2LSH 和 AKM 算法在对选取的 8 类 Caltech-256 图像集构建视觉词典时的时间消耗作了对比，实验结果如图 7 所示。从图 7 可以看出，AKM 算法的时间消耗要远高于弱监督 E2LSH，由此说明弱监督 E2LSH 更适用于大规模数据环境。

为了验证显著图加权视觉语言模型对目标分类的有效性，在利用弱监督 E2LSH 聚类生成视觉词典时令参数 $k=8$，$L=5$，且使得过滤之后词典规模为 1000。然后，分别采用传统的一元视觉语言模型和本文中基于显著图加权的视觉语言模型(Saliency Map Weighted Visual Language Model, SMW-VLM)对随机选取的 8 个目标类别测试集进行实验，得分类结果的混淆矩阵如图 8(a)和图 8(b)所示。从图 8(a)和图 8(b)中可以看出，采用本文方法(WS-



(a)参数 k 对聚类中心数目的影响　　(b)视觉词典规模对目标分类准确率的影响

图 5 不同参数对目标分类的影响

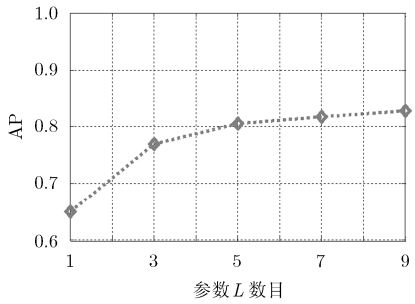


图 6 参数 L 对目标分类结果的影响

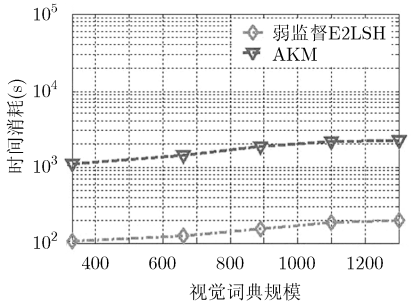
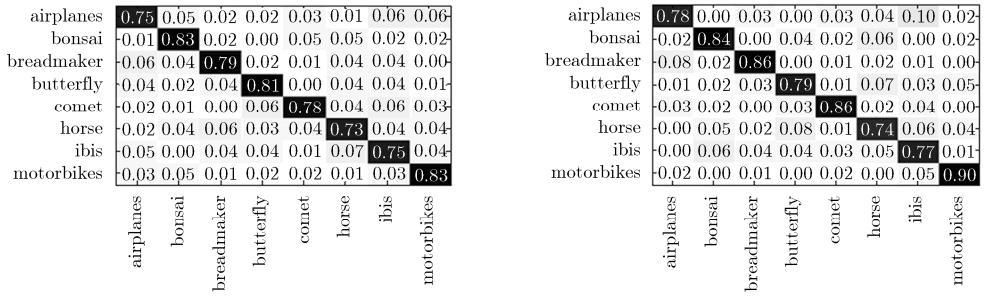


图 7 不同算法的时间效率对比



(a)传统一元视觉语言模型　　(b)显著图加权视觉语言模型

图 8 利用不同视觉语言模型进行分类的混淆矩阵对比

E2LSH+SMW-VLM)进行目标分类时，多个目标分类的召回率均保持较高水平。且采用显著图加权的视觉语言模型相较于传统的一元视觉语言模型可以使目标分类的召回率均有一定的提升，说明显著图加权视觉语言模型能够在利用视觉单词空间信息的同时有效地克服图像背景噪声的不利影响。

为了进一步验证本文方法的有效性，又在 Pascal Voc2007 图像集上进行实验，过滤后的词典规模为 10K。将本文方法(WS-E2LSH+SM-VLM，文中参数分别为 k=8 , L=5)与文献[16]的软分配方法(AKM+SA)、文献[23]基于语言模型的方法(AKM+Language Model, AKM+VLM)，文献[21]中基于稀疏编码多尺度空间潜在语义分析的图像分类方法(AKM+SPM+PLSA)以及文献[22]中的几何短语联合方法(AKM+Geometric Phrase Pooling, AKM+GPP)进行实验对比，得各目标分类平均准确率如表 1 所示。从表 1 中不难看出，本文方法中由于弱监督 E2LSH 生成的视觉词典能够更好地克服视觉单词同义性和歧义性问题，且显著图加权视觉语言模型能够根据视觉单词的显著值为其分配权重，在利用视觉单词空间信息的同时更好地避免了图像背景噪声的不利影响。因此，本文方法的目标分类准确率明显高于其他方法。此外，我们又对比了不同方法在这两个数据集上的平均召回率结果，具体如表 2 所示。从表 2 中不难看出，在两个数据集中，本文方法的目标分类平均召回率均要高于其他方法。但是，需要指出的是，对于不同类别的目标而言，其性能改善程度还与数据本身特性及训练数据是否充分等因素有关。

最后，我们在数据库 Pascal Voc2007 进行实验，将本文方法(WS-E2LSH+SMW-VLM)与其他方法之间的目标分类时间效率进行分析对比，得平均训练时间和平均测试时间如表 3 所示。从表 3 可以看出，本文方法的训练时间要明显低于其他方法，而文献[21]中的 AKM+SPM+PLSA 方法和文献[22]中的 AKM+GPP 方法由于做了诸多空间信息算法优化，因此其训练时间要高于文献[16]中的 AKM+SA 方法。而本文方法由于在视觉语言模型的基础上进行了显著图加权，所以本文方法的测试时间要略高于文献[23]中的 AKM+VLM 方法，然而却明显低于其他方法。

5　结束语

本文提出了一种基于弱监督 E2LSH 和显著图加权的目标分类方法。首先，针对传统视觉词典生成方法造成的效率低、视觉单词同义性及歧义性问题，引入 E2LSH 算法聚类生成词典，并提出一种弱监督策略对哈希函数的选取进行监督以增强其代表

表 1 不同方法对 Pascal Voc2007 中 20 类目标分类准确率对比(%)

目标类别	AKM+SA	AKM+VLM	AKM+SPM+PLSA	AKM+GPP	WS-E2LSH+SMW-VLM
airplanes	65.5	71.6	70.8	76.7	83.1
bicycle	64.8	65.1	69.6	74.5	82.7
bird	65.1	66.4	71.4	72.1	75.8
boat	71.4	67.2	73.8	77.5	82.4
bottle	58.7	55.1	65.7	63.4	68.5
bus	65.9	72.8	76.4	78.4	83.4
car	70.6	64.4	75.9	73.8	80.2
cat	68.6	64.7	72.1	76.4	81.3
chair	65.8	74.2	75.4	77.1	80.7
cow	71.3	74.6	77.2	80.4	85.1
diningtable	67.4	75.3	79.3	76.8	83.6
dog	63.5	59.1	70.1	74.2	79.4
horse	78.2	73.4	75.4	81.3	85.1
motorbike	71.4	65.2	69.3	79.5	80.5
person	74.8	73.1	71.0	86.4	91.4
pottedplant	62.2	66.8	74.8	75.4	79.3
sheep	63.1	67.2	72.7	72.1	80.2
sofa	60.7	69.6	67.4	74.2	77.6
train	76.5	78.4	81.4	84.6	90.5
tvmontor	62.3	65.3	73.5	70.4	76.7
average	67.4	68.5	73.2	76.3	81.4

表 2 两种不同数据库上几种分类方法的平均召回率对比(%)

目标分类方法	图像库	
	Caltech-256	Pascal Voc2007
AKM+SA[19]	61.8	56.9
AKM+VLM[26]	66.4	60.3
AKM+SPM+PLSA[24]	71.5	67.4
AKM+GPP[25]	78.4	69.2
WS-E2LSH+SMW-VLM	82.6	75.1

性，进而使得视觉词典的语义表达能力更强；此外，在引入视觉语言模型克服单词空间信息缺失问题的同时，利用图像显著图分析视觉单词与各目标前景的相关性，为单词分配相应的显著度权值，降低图像背景噪声带来的不利影响。实验结果也能有效地验证本文方法的目标分类性能优于当前主流方法。然而，需要指出的是如何在特征提取和距离度量阶段更好地反映图像语义本质将是我们下一步研究的重点。

表 3 不同方法在数据集 Pascal Voc2007 上的时间效率对比

分类方法	AKM+SA[19]	AKM+VLM[26]	AKM+SPM+PLSA[24]	AKM+GPP[25]	WS-E2SLH+SMW-VLM
训练时间(s)	67.6	54.5	86.7	69.3	7.8
测试时间(ms)	42.3	14.8	131.2	56.5	15.7

参 考 文 献

[1] SIVIC J and ZISSERMAN A. Video Google: a text retrieval approach to object matching in videos[C]. Proceedings of 9th IEEE International Conference on Computer Vision, Nice, France, 2003: 1470–1477.

[2] CHEN Y Z, Dick A, LI X, *et al*. Spatially aware feature selection and weighting for object retrieval[J]. *Image and Vision Computing*, 2013, 31(6): 935–948.

[3] WANG J Y, Bensmail H, and GAO X. Joint learning and weighting of visual vocabulary for bag-of-feature based tissue

classification[J]. *Pattern Recognition*, 2013, 46(3): 3249–3255.

[4] OTÁVIO A, PENATTI B, FERNANDA B S, *et al*. Visual word spatial arrangement for image retrieval and classification[J]. *Pattern Recognition*, 2014, 47(1): 705–720.

[5] 宋相法, 焦李成. 基于稀疏编码和集成学习的多示例多标记图像分类方法[J]. 电子与信息学报, 2013, 35(3): 622–626. doi: 10.3724/SP.J.1146.2012.01218.
SONG Xiangfa and JIAO Licheng. A multi-instance multi-label image classification method based on sparse coding and ensemble learning[J]. *Jounal of Electronics & Information Technology*, 2013, 35(3): 622–626. doi: 10.3724/SP.J.1146.2012.01218.

[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, 32(7): 1271–1283.

[8] NISTER D and STEWENIUS H. Scalable recognition with a vocabulary tree[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 2161–2168.

[9] PHILBIN J, CHUM O, ISARD M, *et al*. Object retrieval with large vocabularies and fast spatial matching[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007: 1–8.

[10] MU Y D, SUN J, and YAN S C. Randomized locality sensitive vocabularies for bag-of-features model[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 1–14.

[11] CAO Yiqun, JIANG Tao, and THOMAS G. Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing[J]. *Bioinformatics*, 2010, 26(7): 953–959.

[12] XIA Hao, WU Pengcheng, and STEVEN C H. Boosting multi-kernel locality-sensitive hashing for scalable image retrieval[C]. Proceedings of 35th ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, Oregon, USA, 2012: 55–64.

[13] 张瑞杰, 郭志刚, 李弼程. 基于 E2LSH-MKL 的视觉语义概念检测[J]. 自动化学报, 2012, 38(10): 1671–1678.
ZHANG Ruijie, GUO Zhigang, and LI Bicheng. A visual semantic concept detection algorithm based on E2LSH-MKL[J]. *Acta Automatica Sinica*, 2012, 38(10): 1671–1678.

[14] ZHENG Q and GAO W. Constructing visual phrases for effective and efficient object-based image retrieval[J]. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2008, 5(1): 1–19.

[15] CHEN T, YAP K H, and ZHANG D J. Discriminative soft bag-of-visual phrase for mobile landmark recognition[J]. *IEEE Transactions on Multimedia*, 2014, 16(3): 612–622.

[16] PHILBIN J, CHUM O, ISARD M, *et al*. Lost in quantization: improving particular object retrieval in large scale image databases[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, 2009: 278–286.

[17] WEINSHALL D, LEVI G, and HANUKAEV D. LDA topic model with soft assignment of descriptors to words[C]. Proceedings of the 30th International Conference on Machine Learning, Atlanta, USA, 2013: 711–719.

[18] LAZEBNIK S, SCHMID C, and PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 2169–2178.

[19] SHARMA G and JURIE F. Learning discriminative spatial representation for image classification[C]. Proceedings of the 22nd British Machine Vision Conference, Dundee, Britain, 2011: 1–11.

[20] 赵春晖, 王莹, KANEKO M. 一种基于词典模型的图像优化分类方法[J]. 电子与信息学报, 2012, 34(9): 2064–2070. doi: 10.3724/SP.J.1146.2012.00047.
ZHAO Chunhui, WANG Ying, and KANEKO M. An optimized method for image classification based on bag of words model[J]. *Journal of Electronics & Information Technology,* 2012, 34(9): 2064–2070. doi: 10.3724/SP.J.1146.2012.00047.

[21] 赵仲秋, 季海峰, 高隽, 等. 基于稀疏编码多尺度空间潜在语义分析的图像分类[J]. 计算机学报, 2014, 37(6): 1251–1260.
ZHAO Zhongqiu, JI Haifeng, GAO Jun, *et al*. Sparse coding based on multi-scale spatial latent semantic analysis for image classification[J]. *Chinese Journal of Computers*, 2014, 37(6): 1251–1260.

[22] XIE L, TIAN Q, and ZHANG B. Spatial pooling of heterogeneous features for image classification[J]. *IEEE Transactions on Image Processing*, 2014, 23(5): 1994–2008.

[23] GENG B, YANG L, and XU C. A study of language model for image retrieval[C]. Proceedings of IEEE International Conference on Data Mining Workshops, Washington, DC, USA, 2009: 158–163.

[24] 吴磊. 视觉语言分析: 从底层视觉特征表达到语义距离学习[D]. [博士论文], 中国科学技术大学, 2010.
WU Lei. Visual language analysis: from low level feature representation to semantic metric learning[D]. [Ph.D. dissertation], University of Science and Technology of China, 2010.

[25] DATAR M, IMMORLICA N, and INDYK P.

Locality-sensitive hashing scheme based on p-stable distributions[C]. Proceedings of the 20th Annual Symposium on Computational Geometry, New York, USA, 2004: 253–262.

[26] HAREL J, KOCH C, and PERONA P. Graph-based visual saliency [C]. Proceedings of Advances in Neural Information Processing Systems, NewYork, USA, 2007: 545–552.

[27] SLANEY M and CASEY M. Locality-sensitive hashing for finding nearest neighbors[J]. *IEEE Signal Processing Magazine*, 2008, 25(2): 128–131.

[28] 高毫林, 彭天强, 李弼程. 基于多表频繁项投票和桶映射链的快速检索方法[J]. 电子与信息学报, 2012, 34(11): 2574–2581. doi: 10.3724/ SP.J.1146.2012.00548.
GAO Haolin, PENG Tianqiang, and LI Bicheng. A fast retrieval method based on frequent items voting of multi table and bucket map chain[J]. *Journal of Electronics & Information Technology*, 2012, 34(11): 2574–2581. doi: 10.3724/SP.J.1146.2012.00548.

[29] BREIMAN L. Random forests [OL]. http://www.stat.-berkeley.edu/RandomForests/.2014–07.

[30] ITTI L, KOCH C, and NIEBUR E. A model of saliency-based visual attention for rapid scene analysis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(4): 1254–1259.

[31] LI F F, FERGUS R, and PERONA P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[J]. *Computer Vision and Image Understanding*, 2007, 106(1): 59–70.

[32] EVERINGHAM M, VAN Gool L, WILLIAMS C K I, *et al*. The pascal visual object classes challenge 2007 (VOC 2007) results [OL]. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/results/index.shtml, 2014.

赵永威： 男，1988 年生，博士生，研究方向为视频/图像分析及处理.

李弼程： 男，1970 年生，博士，教授，博士生导师，主要研究方向为智能信息处理.

柯圣财： 男，1991 年生，硕士生，研究方向为图像检索与分类.