计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150070
53(5):1043-1052,2016

基于词典优化与空间一致性度量的目标检索

赵永威[1] 周 苑[2] 李弼程[3]
[1](武警工程大学电子技术系 西安 710000)
[2](河南工程学院计算机学院 郑州 451191)
[3](解放军信息工程大学信息系统工程学院 郑州 450002)
(zhaoyongwei369@163.com)

Object Retrieval Based on Enhanced Dictionary and Spatially-Constrained Similarity Measurement

Zhao Yongwei[1], Zhou Yuan[2], and Li Bicheng[3]
[1](*Department of Electronic Technology, CAPF Engineering University, Xi'an* 710000)
[2](*School of Computer Science, Henan University of Engineering, Zhengzhou* 451191)
[3](*Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou* 450002)

Abstract Bag of visual words model based object retrieval methods have several problems, such as low time efficiency, the low distinction of visual words and the weakly visual semantic resolution because of missing spatial information and quantization error. In this article, an object retrieval method based on enhanced dictionary and spatially-constrained similarity measurement is proposed aiming at the above problems. Firstly, E^{2}LSH (exact Euclidean locality sensitive hashing) is used to identify and eliminate the noise key points and similar key points, consequently, the efficiency and quality of visual words are improved; Then, the stop words of dictionary are eliminated by chi-square model (CSM) to improve the distinguish ability of visual dictionary; Finally, the spatially-constrained similarity measurement is introduced to accomplish object retrieval, furthermore, a robust re-ranking method with the K-nearest neighbors of the query for automatically refining the initial search results is introduced. Experimental results indicate that the quality of visual dictionary is enhanced, and the distinguish ability of visual semantic expression is effectively improved and the object retrieval performance is substantially boosted compared with the traditional methods.

Key words object retrieval; bag of visual words model; exact Euclidean locality sensitive hashing (E^{2}LSH); spatially-constrained similarity measure; chi-square model (CSM)

摘 要 基于视觉词典模型(bag of visual words model, BoVWM)的目标检索存在时间效率低、词典区分性不强的问题,以及由于空间信息的缺失及量化误差等导致的视觉语义分辨力不强的问题.针对这些问题,提出了基于词典优化与空间一致性度量的目标检索方法.首先,该方法引入 E^{2}LSH(exact Euclidean locality sensitive hashing)过滤图像中的噪声和相似关键点,提高词典生成效率和质量;然后,引入卡方模型(chi-square model, CSM)移除词典中的视觉停用词增强视觉词典的区分性;最后,采用空间一致性度量准则进行目标检索并对初始结果进行 K-近邻(K-nearest neighbors, K-NN)重排序.实验结果表明:新方法在一定程度上改善了视觉词典的质量,增强了视觉语义分辨能力,进而有效地提高目标检索性能.

收稿日期:2015-01-20;**修回日期**:2015-07-07
基金项目:国家自然科学基金项目(60872142,61301232)
This work was supported by the National Natural Science Foundation of China (60872142,61301232).

关键词 目标检索;视觉词典模型;精确欧氏位置敏感哈希;空间一致性度量;卡方模型

中图法分类号 TP391

近年来,随着图像数据规模的增大,使得图像处理面临的环境更加复杂.虽然 SIFT 等[1]局部特征在图像处理领域表现出了良好的性能,但是,其特征维数较高,若采用 VA-File,*K*-d 树等一些传统的索引结构进行检索就会导致"维数灾难"现象.视觉词典模型(bag of visual words model, BoVWM)[2-3]由于其突出性能,已成为当前图像标注[4]、图像检索与分类[5-8]等领域的主要解决方法.但是,以下 3 个关键性问题的存在极大地限制了 BoVWM 模型的性能:1)关键点检测算子会产生大量的噪声点无疑会增加计算消耗、降低词典生成效率;2)当前聚类算法的局限性[9-10]和图像背景噪声的存在,使得聚类生成的词典中包含一些类似于文本信息中的"的"、"和"、"是"等"停用词",这里称其为"视觉停用词",严重影响了视觉词典的质量;3)传统的 BoVWM 模型中视觉单词间空间信息的缺失和量化误差严重等导致视觉语义表达分辨力不强.

近年来,研究人员针对这些问题做了许多探索性研究,如在过滤噪声关键点方面:Rudinac 等人[11]将相互距离小于 1 个像素值的特征点看作相似的近邻点,然后计算其中心值作为代表性特征点,这种方法最大的缺点是计算开销大,因为它需要遍历图像的每个像素点.Jamshy 等人[12]通过学习特征点对某一特定应用的先验知识来过滤大部分特征点,然而这种方法却降低了图像分类性能.而针对"视觉停用词"去除问题,Sivic 等人[2]考虑到单词的信息量大小与其出现的频率有一定的关系,从而提出了一种基于词频的"停用词"过滤方法,然而,这种方法却忽略了视觉单词和目标语义概念间的相互关系.Tirilly 等人[13]则根据关键点的几何性和概率隐语义分析模型淘汰无用的视觉单词,Yuan 等人[14]试图以统计视觉单词组合也即"停用词组"出现的概率来滤除一些无用信息,但是却忽略视觉词组内部各单词之间的空间关系.

针对视觉单词间空间信息的缺失和量化误差严重的问题,刘硕研等人[15]采用一种基于上下文语义信息的图像块视觉单词生成算法,利用 PLSA 模型和 Markov 随机场共同挖掘单词的上下文信息.张瑞杰等人[16]考虑到图像多尺度空间与单词上下文语义共生关系,在不同的图像尺度空间挖掘单词的上下文语义信息,进一步弥补了传统 BoVWM 模型的空间信息不足问题.Chen 等人[10]则提出了一种基于软分配的视觉词组(visual phrase)构建方法,在弥补视觉单词空间信息的同时,有效克服了传统视觉词组构建方法[17]导致的特征信息丢失问题.而为了减小量化误差,Gemert 等人[18]提出了视觉单词不确定性(visual word uncertainty)模型,该模型同样是采用软分配策略对 SIFT 特征编码,进一步验证了软分配方法对于减弱视觉单词同义性和歧义性影响的有效性.Otávio 等人[19]则提出一种基于视觉单词空间分布的图像检索和分类方法,该方法将视觉单词的空间信息嵌入到向量空间中,并对单词在图像中的相对位置关系进行编码,从而得到更为紧致的视觉表达方式.Yang 等人[20]则利用视觉语言模型结合目标区域周围的视觉单元构建了包含上下文语义信息的目标语言模型,进一步改善了目标检索性能.此外,文献[21]在利用上下文近义词构建视觉词汇直方图的同时,结合查询扩展方法解决目标视角变化较大、目标遮挡严重的情况问题.但是,查询扩展方法都依赖于较高的初始查全率,在初始查全率较低时反而会带来一些负面影响.

针对上述问题,本文提出一种基于视觉词典优化与空间一致性度量的目标检索方法.1)引入精确欧氏位置敏感哈希算法[22](exact Euclidean locality sensitive hashing, E^2LSH),利用该算法的位置敏感性和处理高维数据的高效性对图像初始关键点进行过滤,降低噪声点的影响,降低计算消耗;2)根据引入卡方模型(chi-square model, CSM)分析视觉单词与目标类别的相关性大小并结合单词词频滤除一定数量的视觉停用词,增强视觉词典的区分性;3)采用一种包含特征点角度、方向等空间一致性信息的度量方法完成目标检索,并引入 *K*-近邻重排序方法,进一步改善目标检索在复杂环境下的性能.

1 视觉词典优化

1.1 关键点过滤

假设在视觉位置相近的关键点是相似的,将其捆绑在一块计算其质心,并将其作为一个有代表性

的关键点.每个关键点 $p_i=\{\boldsymbol{u}_i,s_i,\theta_i,\boldsymbol{r}_i\}$ 由 4 部分组成,分别为:特征点在图像中的位置坐标 $\boldsymbol{u}_i$、特征的尺度 s_i、主方向 θ_i 及 128 维 SIFT 描述向量 $\boldsymbol{r}_i$.为了提高过滤效果,本文选取 $k(k=6)$ 个位置敏感函数联合起来以拉大关键点碰撞概率之间的差距,定义函数族:

$$G=\{g:S\rightarrow U^k\}, \tag{1}$$

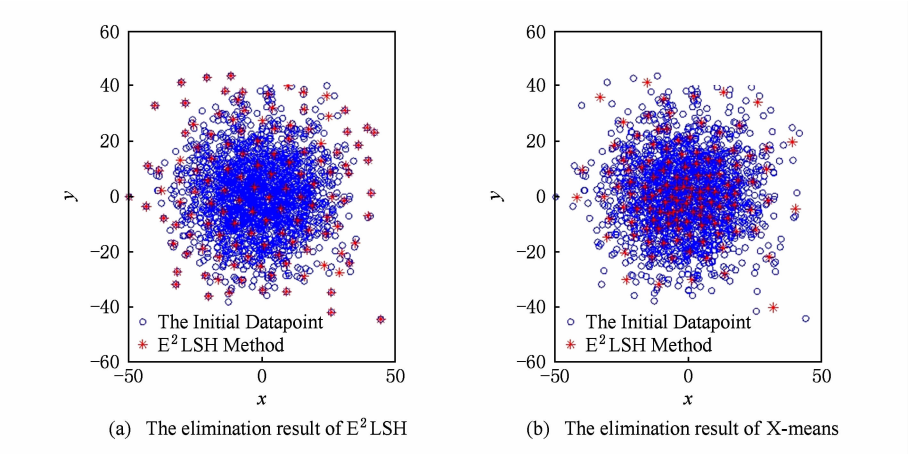
其中,$g(p)=(h_1(p),h_2(p),\cdots,h_k(p))$,可知,经函数 $g(p)\in G$ 降维映射后,关键点 p 都会变为一个 k 维向量 $\boldsymbol{a}=(a_1,a_2,\cdots,a_k)$,然后,再采用主次 Hash 函数 h_1,h_2 对向量 $\boldsymbol{a}$ 进行 Hash,构建 Hash 表并存储关键点.主次 Hash 函数的定义如下:

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

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

其中,r_i' 和 r_i'' 均为随机整数;s 是图像关键点总数目;m 为一个大的素数,通常取 $2^{32}-5$.

由文献[8]的研究表明,X-means 算法是当前关键点过滤方法中较为有效的主流过滤方法,为此,本文分别采用 E^{2}LSH 和 X-means 算法对随机产生的数据点进行过滤以验证 E^{2}LSH 算法的有效性.如图 1 所示.从图 1 不难看出,X-means 方法过滤得到的代表性关键点较为不均,而由 E^{2}LSH 过滤得到的代表性关键点更为均匀.因此,基于 E^{2}LSH 的过滤方法在一定程度上能够避免关键点密集区域描述同一语义概念的关键点被分别捆绑到多个类别的现象,同时也能避免关键点稀疏区域描述不同语义概念的关键点被错误地捆绑到一个类别的现象,进而,提高过滤后各关键点的代表性和区分能力.图 2 进一步给出了 E^{2}LSH 对 Oxford5K 数据库中 all_souls81 图片关键点过滤的效果图,其中,圆圈代表初始关键点,星形点则表示经 E^{2}LSH 过滤后的关键点.由图 2 不难看出,E^{2}LSH 算法能有效地对关键点进行过滤,提高关键点的代表性.



(a) The elimination result of E^{2}LSH (b) The elimination result of X-means

Fig. 1 The sketch map of different methods for eliminating key points.

图 1 不同方法对关键点过滤的示意图

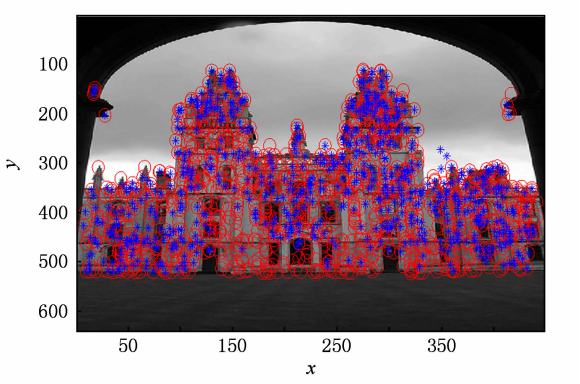


Fig. 2 The map of E^{2}LSH to eliminate the key points.

图 2 E^{2}LSH 对 all_souls81 图片关键点过滤效果图

1.2 “视觉停用词”去除

卡方模型是一种医学上常用的测量 2 个随机变量相关性的方法,受此启发,可以采用卡方模型统计视觉单词与各目标图像类别之间的相关性,卡方值越小表示该视觉单词与各图像类别的相关性越小,区分性也就弱,反之亦然.因此,可以结合单词词频以更好地滤除“视觉停用词”.假设视觉单词 w 的出现频次独立于目标类别 C_j,$C_j\in C$,$1\leqslant j\leqslant k$,图像集 $C=\{C_1,C_2,\cdots,C_k\}$,而视觉单词 w 与图像集 C 中目标类别的相互关系可以由表 1 来描述.

表 1 中,n_{1j} 表示目标类别 C_j 包含单词 w 的图像数目,n_{2j} 表示目标类别 C_j 不包含单词 w 的图像数目,n_{+j} 则表示目标类别 C_j 中的图像总数,并用 n_{i+},$i=1,2$ 分别表示图像集 C 中包含单词 w 的图像总数和不包含 w 的图像总数.如此,表 1 中视觉

Table 1 The Relationship Between Visual Word and Object Categories

表 1 视觉单词与各目标类别关系

Image Number	C_1	C_2	…	C_k	Total
The Image Number Including w	n_{11}	n_{12}	…	n_{1k}	n_{1+}
The Image Number not Including w	n_{21}	n_{22}	…	n_{2k}	n_{2+}
Total	n_{+1}	n_{+2}	…	n_{+k}	N

单词 w 与各图像类别的卡方值可计算如下：

$$x^2 = \delta = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(N\times n_{ij} - n_{i+}\times n_{+j})}{N\times n_{i+}\times n_{+j}}, \quad (4)$$

卡方值 x^2 就代表了 w 与各目标类别间统计相关性的大小，同时考虑到单词 w 词频的影响，对卡方值赋予权重如下：

$$\tilde{x}^2 = \frac{x^2}{tf(w)}, \quad (5)$$

其中，$tf(w)$表示单词 w 词频. 由此，就能够按照式(5)对每个单词的卡方值进行排序，然后去除一定数量 S 的“视觉停用词”即可.

2 相似性度量准则

2.1 空间一致性度量方法

假设一幅由矩形框界定好的查询目标图像，其空间信息可以表示如下：$B=\{x_c, y_c, w, h, \theta\}$，如图 3 所示，其中，$(x_c, y_c)$是界定目标的矩形框中心坐标；$w, h$ 分别代表矩形框的宽和高；θ 表示矩形框的旋转角度. 通过相似变换 T 就能检索到图像库中与之最匹配的图像，$T=\{\boldsymbol{R}(a), s, \boldsymbol{t}\}$，$a$ 指目标的旋转角度，$\boldsymbol{R}(a)=\begin{bmatrix}\cos a, & -\sin a\\ \sin a, & \cos a\end{bmatrix}$，$s$ 代表尺度变化，$\boldsymbol{t}=(x_t, y_t)$代表位置变化，那么经过变换后的目标图像即为：$B'=T(B)=\{x_c+x_t, y_c+y_t, s\times w, s\times h, \theta=a\}$.

这里，用 Q 表示查询图像，D 表示图像库中任一幅图像，其 SIFT 特征点分别表示为$\{f_1, f_2, \cdots, f_m\}$，$\{g_1, g_2, \cdots, g_n\}$，那么，2 幅图像之间的空间一致性度量可计算如下：

$$S(Q, D|T)=\sum_{k=1}^{n}\sum_{\substack{(f_i, g_j)\\ f_i\in Q, g_j\in D\\ w(f_i)=w(g_j)=w_k\\ \|T(\boldsymbol{L}(f_i))-\boldsymbol{L}(g_j))\|<\varepsilon}}\frac{idf^2(w_k)}{tf_Q(w_k)\times tf_D(w_k)}, \quad (6)$$

其中，w_k 表示视觉词典里的第 k 个单词；n 为词典规模；$w(f_i)=w(g_j)=w_k$ 表示特征点 f_i, g_j 都被映射至单词 w_k 上；$\boldsymbol{L}(f)=(x_f, y_f)$表示特征点的位置；$\|T(\boldsymbol{L}(f_i)-\boldsymbol{L}(g_j))\|<\varepsilon$ 为 2 个相互匹配的特征点之间的约束条件，保证它们经过变换之后位置依然较近；$idf(w_k)$表示单词 w_k 的逆文档频率；$tf_Q(w_k)$表示单词 w_k 在图像 Q 中的词频；$tf_D(w_k)$表示在图像 D 中的词频，这样就能降低那些在图像中经常出现的单词的权重，有助于提高相似性度量的准确性. 因此，对于图像库中的每一幅图像而言，就是要找到最优变换 T^* 使得度量值最大. 也即：

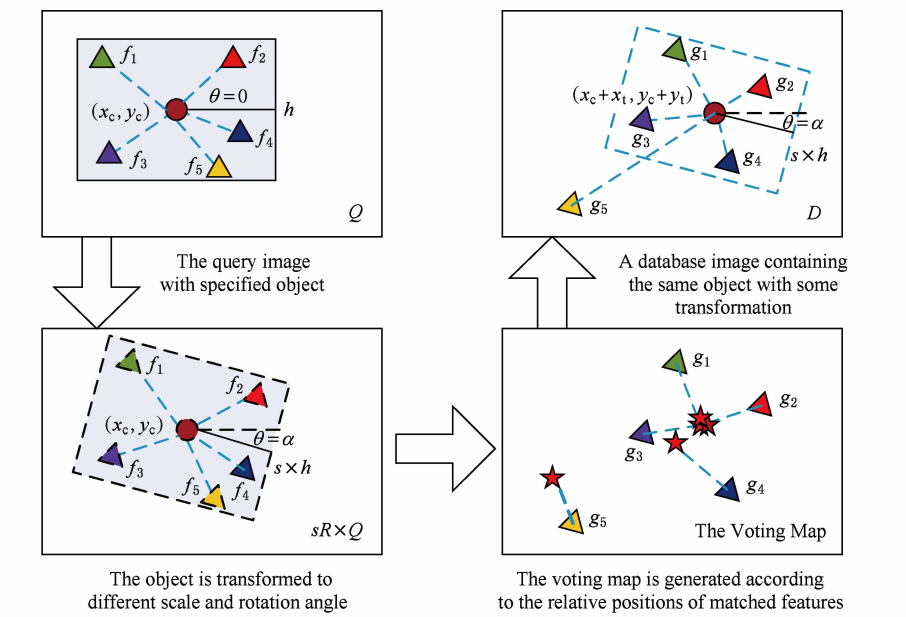


Fig. 3 The illustration of spatially-constrained similarity measurement.

图 3 空间一致性度量示意图

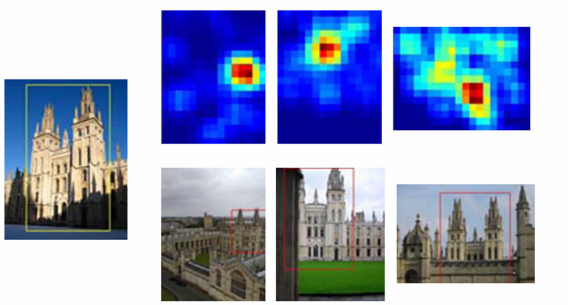
$$T^*=\{\boldsymbol{R}(a^*),s^*,\boldsymbol{t}^*\}=\arg\max_T S(Q,D|T). \quad (7)$$

故而,就有 $S^*(Q,D)=S(Q,D|T^*)$可以用来衡量图像 Q 和图像 D 之间的相似性,且所有的检索结果也能以此进行排序.由图 3 上面 2 幅图不难看出,2 幅图像中只有特征点(f_i,g_i),$i=1,2,3$ 是满足空间一致性条件的.(f_5,g_5)是一个错误匹配点对,(f_4,g_4)的取舍则决定于式(6)中参数 ε 的大小.

为了计算 $S^*(Q,D)$,需要找到最优变换 T^*,这里可将对 T 进行分解处理,首先将 360 角度空间划分 n_R 部分,(一般 $n_R=4$ 或 8),同样地,尺度空间被划分为 n_S 部分,通常 $n_S=8$,变化范围为 1/2 到 2 之间.令 $\boldsymbol{V}(f)$表示特征点与查询图像中的矩形框中心 c_Q 之间的相对位置关系向量,那么由匹配的特征点对(f,g)的位置及 $\boldsymbol{V}(f)$就能定位图像 D 中的矩形框中心,$\boldsymbol{L}(c_Q)=\boldsymbol{L}(g)-\boldsymbol{V}(f)$,如果 $w(f)=w(g)=w_k$,特征点对(f,g)的投票得分为

$$Score(w_k)=\frac{idf^2(w_k)}{tf_Q(w_k)\times tf_D(w_k)}, \quad (8)$$

不难看出,若相互匹配的特征点对符合空间一致性条件,那么尤其投票得出的矩形框中心位置也是相近的,如图 3 所示.每次投票得出的目标位置中心就代表了一个变换 T,那么利用式(8)投票所得分数就等同于利用式(6)进行相似性度量.可以看出,这种机制可以同时进行目标检索和定位而不需要子图检索和后处理,极大地提高了目标检索系统的实用性和方便性.在实际应用中,可将投票得分图归一化为 $n_x\times n_y$ 个图像块大小,同时为了避免投票时的量化误差及弱化目标遮挡等情况的影响,本文对所估计的中心块周围的 16×16 像素的窗口块进行投票,而每个块的得分大小为 $Score(w_k)\times e^{-d/\sigma^2}$,$e^{-d/\sigma^2}$ 为权重系数,由每个块与中心块之间的距离 d 和 σ 参数决定,整个过程相当于对投票得分图进行一次高斯平滑.



(a) Query object (b) Retrieval object

Fig. 4 The retrieval result examples of vote map and object location map.

图 4 检索结果的投票得分图和目标定位示意图

图 4 给出了对中心块周围的像素的窗口块进行高斯平滑以及对应的目标定位的示例图,从图 4 可以看出,给出一幅查询图像就能按照上述方法得到相应检索图像的投票得分图,然后依据此对目标进行定位,而每个投票得分图都存在一个极值点,也即是大部分匹配特征点对都将票数投向的位置.

2.2 *K*-近邻重排序

根据上述相似性度量方法对数据库进行检索,那么结果可依据 $S^*(Q,D)$值的大小进行排序,记为 $R(Q,D)$,并令 N_i 表示查询图像的第 i 个检索结果,则有 $R(Q,N_i)=i$,用 $N_q=\{N_i\}$,$i=1,2,\cdots,k$ 表示查询图像的 K-近邻.为了有效地利用 K-近邻图像包含的信息,本文重新利用其中的每一幅图像作为查询图像重新检索,并分别将排序结果记为 $R(N_i,D)$,依据这个排序结果给图像库中的每幅图像分配一个得分 $1/R(N_i,D)$,那么经重排序之后的图像得分可定义为

$$\bar{S}(Q,D)=\frac{\omega_0}{R(Q,D)}+\sum_{i=1}^{k}\frac{\omega_i}{R(N_i,D)}, \quad (9)$$

其中,ω_i 为权重系数,由初始排序决定,这里,令 $\omega_0=1$,$\omega_i=\frac{1}{1+R(N_i,D)}=\frac{1}{1+i}$,若将查询图像本身看作其第 0 近邻,那么式(9)可转化为

$$\bar{S}(Q,D)=\sum_{i=0}^{k}\frac{\omega_i}{R(N_i,D)}=\frac{1}{(i+1)R(N_i,D)}, \quad (10)$$

其中,$\bar{S}(Q,D)$是一个单向性度量,而只有 $R(Q,N_i)$和 $R(N_i,Q)$同时排在前列的时候才能保证两者互为近邻.为此,可将权重系数 w_i 改为 $\omega_i=1/(R(Q,N_i)+R(N_i,Q)+1)=1/(i+R(N_i,Q)+1)$,由此可以得到最终的相似性度量准则为

$$\bar{S}(Q,D)=\sum_{i=0}^{k}\frac{1}{(i+R(N_i,Q)+1)R(N_i,D)}, \quad (11)$$

然后,所有图像即可按照式(11)进行重排序,完成检索.

3 实验设置与性能分析

3.1 实验设置

本文选取 Oxford5K 数据库[23]作为实验数据库,并从每个目标类别中选取 50 幅图像,共 550 幅图像作为训练图像库来生成视觉词典,词典规模为 10 000.此外,引入 Flickr1 数据库[24]作为干扰数据以验证本文方法在复杂环境下的实验性能.实验硬件

配置为 Core 2.6 GHz×4、内存 4 GB 的台式机，软件环境为 MATLAB2012a，性能评价指标采用查准率均值(average precision, *AP*)和平均查准率均值(mean average precision, *MAP*)以及时间效率，相关定义如下：

$$查全率=\frac{检索出的相关图像}{全部相关图像}\times 100\%, \quad (12)$$

$$查准率=\frac{检索出的相关图像}{检索出的全部图像}\times 100\%. \quad (13)$$

3.2 实验性能分析

首先，为了选取合适的 Hash 函数个数值，实验从 550 幅训练图像库中提取约 1 436 634 个特征点，然后利用 E^{2}LSH 对其过滤，并采用 AKM 聚类算法对未过滤关键点和不同 k 值过滤后的特征点进行聚类，生成相同单词数目的词典进行目标检索分析了参数 k 对目标检索结果 *MAP* 值的影响(此时，令 $\sigma^2=0$)，如图 5 所示. 从图 5 不难看出，随着参数 k 值的变化，目标检索的 *MAP* 值也随之变化，且在 $k>3$ 时，经 E^{2}LSH 过滤后的检索 *MAP* 值要高于未过滤的目标检索. 当 $k=6$ 时，目标检索 *MAP* 值最大，这是因为，当 k 值较小时会使得过滤后的关键点数目过少，从而容易丢失图像包含的细节信息，而当 k 值较大时导致过滤后的特征点数目过多，使得算法过滤效果不明显，综合考虑，本文取 $k=6$ 时剩余代表性关键点数目为 1 002 105 个，过滤率为 31.3%. 然后，实验又将本文方法与传统的 AKM 算法在生成视觉词典时的时间消耗作了对比，具体如图 6 所示. 从图 6 可以看出，本文方法在经 E^{2}LSH 算法的过滤以后，视觉词典的生成效率有较为明显的提升.

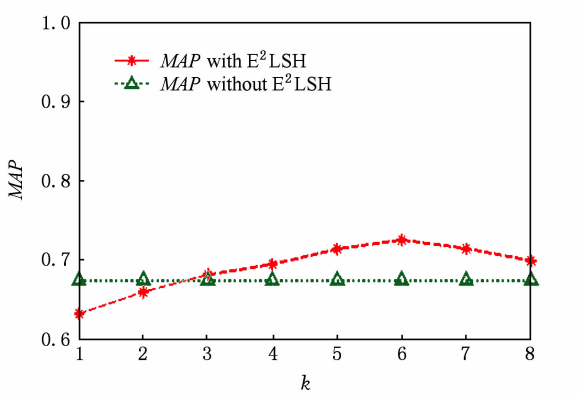


Fig. 5 The influence of parameter k on *MAP*.

图 5 参数 k 对目标检索 *MAP* 值的影响

随后，为了验证卡方模型对滤除“视觉停用词”的有效性，实验在 E^{2}LSH 函数个数 $k=6$ 的情况下对关键点进行过滤，并生成规模为 10 000 的视觉词典，然后利用卡方模型滤除一定数量 S 的视觉停用词，验证过滤不同数目“视觉停用词”对目标检索结果的影响，并与未进行视觉停用词滤除时的目标检索结果进行对比，得其检索 *MAP* 值如图 7 所示. 从图 7 不难看出，采用卡方模型滤除一定数目的“视觉停用词”能够在一定程度上提高目标检索的 *MAP* 值，并且在滤除数目 $S=1\,000$ 时能够达到最高的 *MAP* 值，即为 76.4%. 同时，从图 7 可以看出，当滤除的单词数目过多时，会导致目标检索性能降低，这是因为滤除过多难免使一些代表性强的单词也被错误地滤除.

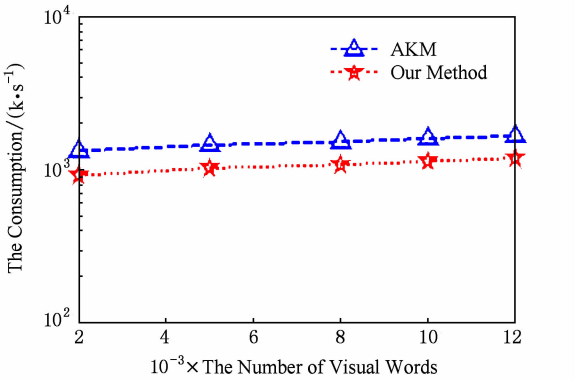


Fig. 6 The efficiency comparison of different methods.

图 6 不同方法构建词典效率对比

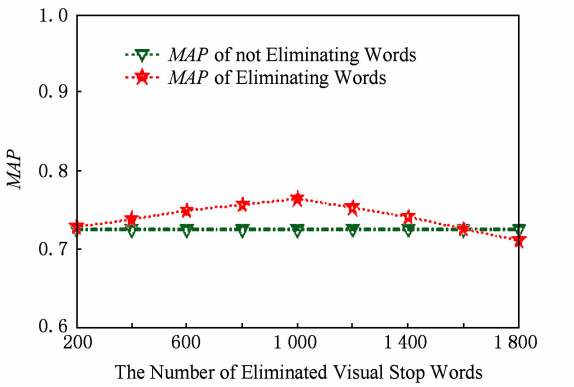


Fig. 7 The influence of the number of eliminated visual stop words on *MAP*.

图 7 去除停用词数目对目标检索 *MAP* 值的影响

然后，在 E^{2}LSH 函数个数 $k=6$、去除视觉停用词数目 $S=1\,000$ 的情况下，实验以 Oxford5K 为实验数据库分析了空间一致性度量准则中参数 σ^2 对目标检索 *MAP* 值的影响，结果如图 8 所示. 其中，当 $\sigma^2=0$ 时表示不对投票结果进行高斯平滑，也即是每个匹配特征对都将票数投向根据式(8)所估计的一个中心块，由图 8 不难看出，当 $\sigma^2>0$ 时，也即表示对所估计的中心块周围 16×16 窗口块进行投

票的 *MAP* 值明显优于未对投票结果进行高斯平滑的情况(即 $\sigma^2=0$),且在 $\sigma^2=2.5$ 时取得最大的 *MAP* 值,因此,本文取 $\sigma^2=2.5$.

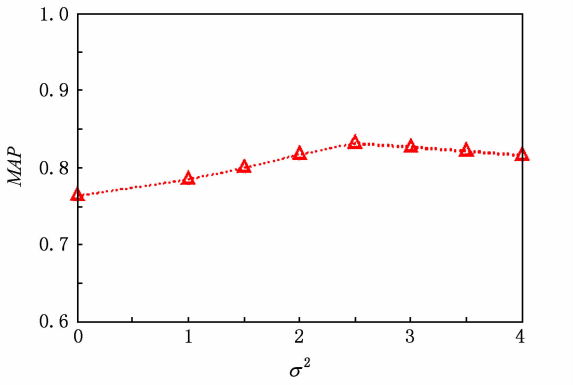


Fig. 8 The influence of parameter σ^2 on *MAP*.

图 8 参数 σ^2 对 *MAP* 值的影响

其次,由于基于上下文语言模型的目标检索方法[20](AKM+language model, AKM+LM)能够很好地记录视觉单词间的空间关系,是当前弥补空间信息不足方面具有代表性的方法,而基于上下文近义词和查询扩展目标检索方法[21](contextual synonymous visual words+query expansion, CSVW+QE)在映射视觉词汇直方图时,很好地利用了视觉单词的上下文近义词,也是当前较为经典的利用单词空间信息的方法,且该方法又引入了查询扩展策略进一步改善检索结果.因此,为了验证本文方法中空间一致性度量准则以及重排序方法对改善目标检索结果的有效性,实验将本文方法(enhanced visual dictionary and spatially-constrained similarity measure, EVD+SCSM)与 AKM+LM 方法、CSVW+QE 方法以及将优化的视觉词典与语言模型相结合的方法(enhanced visual dictionary+language model, EVD+LM)在 Oxford5K 数据库上对 11 个查询目标的检索准确度作了比较,得平均查准率均值 *MAP* 如表 2 所示:

Table 2 The Comparison of Object Retrieval *MAP* Values of Different Methods

表 2 目标检索 *MAP* 值对比 %

Query Object	AKM+LM	CSVW+QE	EVD+LM	EVD+SCSM
Ashmolean	60.4	76.1	74.5	86.8
All Souls	57.6	75.4	73.2	88.7
Balliol	60.5	73.2	70.7	81.5
Bodleian	53.9	65.7	64.3	72.3
Cornmarket	63.3	76.8	77.8	85.2
Christ Church	60.4	71.3	70.3	75.4
Magdalen	33.2	41.6	45.5	59.6
Hertford	79.5	87.3	85.8	94.2
Pitt Rivers	91.2	96.2	93.5	98.1
Keble	75.7	90.3	87.5	91.4
Radcliffe Cam	60.8	79.7	77.1	86.1
Average	63.32	75.78	74.56	83.57

从表 2 可知,对不同的查询目标而言,采用 AKM+LM 方法的 *MAP* 值均低于其他 3 种方法;而 EVD+LM 方法的 *MAP* 值相较于 AKM+LM 方法有一定的改善,足以说明本文提出的词典优化方法能有效降低图像背景噪声点和停用词的影响,提高视觉词典的区分性;同时 CSVW+QE 方法的性能要略好于 EVD+LM 方法,这是因为 CSVW+QE 方法在利用空间信息的基础上又结合查询扩展策略,得到了更多与查询目标相关的图像.但是,本文方法的检索 *MAP* 值要远高于上述 3 类方法,与 EVD+LM 方法对比可以看出,本文中的空间一致性度量准则对单词空间信息的利用优于视觉语言模型.由此也说明改善视觉词典质量能提高视觉词典对图像内容的语义表达能力,而更加准确的度量方法则能更加精确地对图像内容的表达形式进行度量,二者都能在一定程度上提高目标检索精度.与 CSVW+QE 方法相比,可知本文方法在词典优化的基础上,结合空间一致性度量准则和重排序,使得本文方法综合性能优于 CSVW+QE 方法.

然后,又引入 Flickr1 数据库作为干扰数据验证本文方法在复杂数据环境下的性能.实验结果如图 9 和图 10 所示.对比图 9 和图 10 可知,采用本文方法

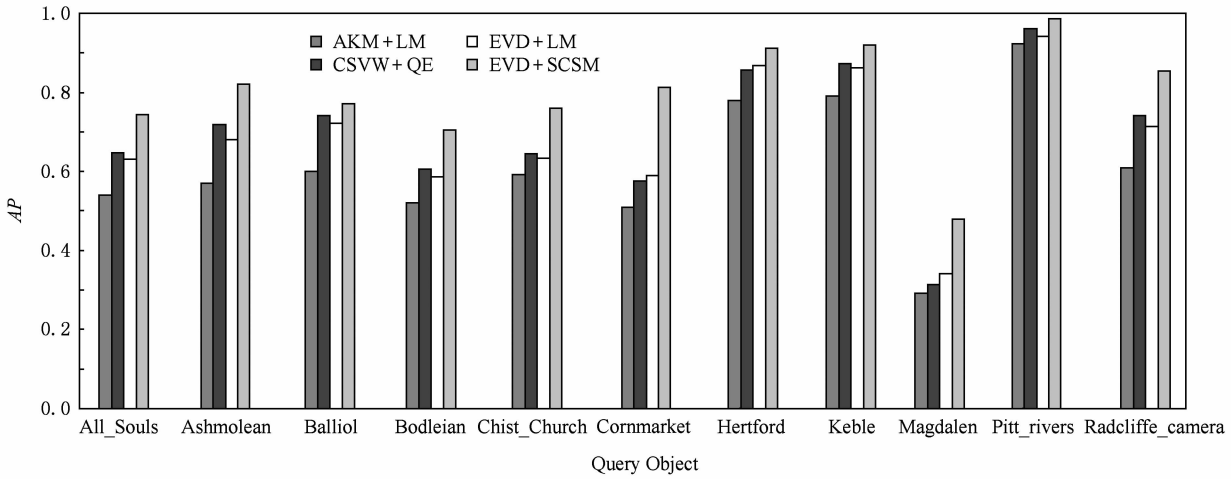


Fig. 9 The *AP* of different methods on Oxford5K.

图 9 在 Oxford5K 数据库上的目标检索 *AP* 值

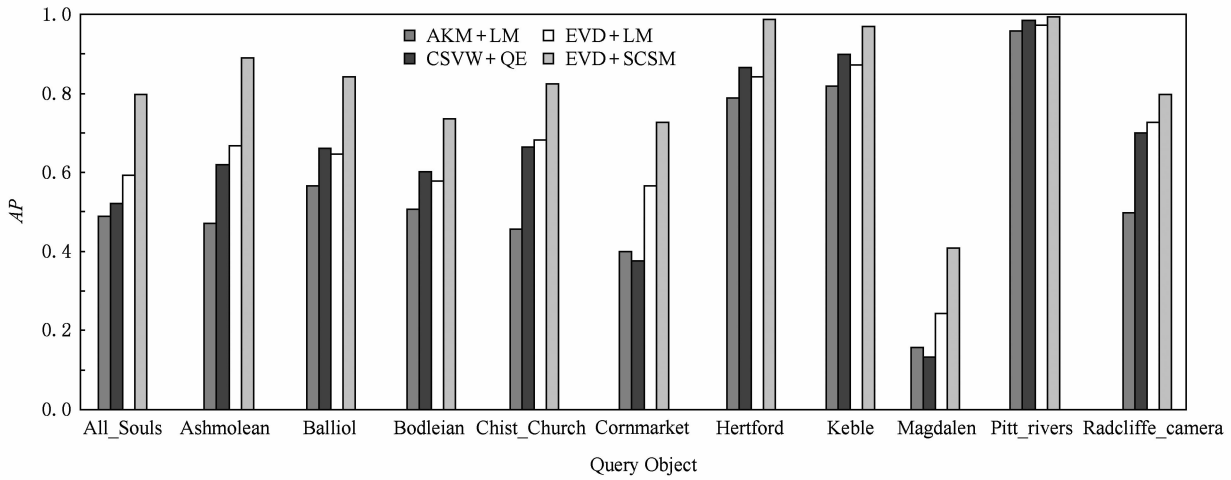


Fig. 10 The *AP* of different methods on Oxford5K+Flickr1.

图 10 在 Oxford5K+Flickr1 数据库上的目标检索 *AP* 值

(EVD+SCSM)进行检索较之其他 3 种方法有更好的表现,且在加入干扰项数据库之后,AKM+LM 方法、EVD+LM 方法因没有对查询目标的信息进行有效扩展,因此其检索性能都有明显的下降. CSVW+QE 方法及本文方法却下降不明显;但是,当加入大规模干扰数据之后,由于 CSVW+QE 方法中的查询扩展策略依赖于较高的初始查全率,所以对于初始查全率较低的 Cornmarket, Magdalen 等目标而言,其检索 *AP* 值反而低于 AKM+LM 方法和 EVD+LM 方法. 而本文方法采用的 *K*-近邻重排序方法是在空间一致性度量准则下进行的,能够自动地舍弃那些不满足空间一致性条件的特征点信息,所以其检索 *AP* 值不受初始查全率影响,由此说明本文方法在大规模干扰数据情况下仍能取得较好的检索结果,实用性更强.

最后,图 11 给出了 *K*-近邻重排序方法的效果示例图. 从图 11 可以看出,第 1 行图像中的第 5 幅

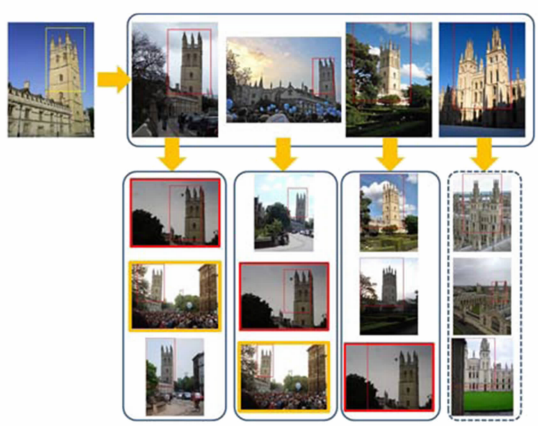


Fig. 11 Example of *K*-NN re-ranking result.

图 11 *K*-近邻重排序结果示意图

最近邻图像与查询目标图像无关,但是由其检索得到的虚线框中的任何一幅图像的最终检索得分不会改变,因为它们与其他的最近邻图像不相关.而用实线框框标识的图像会得到较高的检索得分,因为它们与 K-近邻中的多数图像相关.不难看出,采用 K-近邻重排序方法之后可以得到更多包含查询目标的图像.

4 结 语

为了改善生成视觉词典的质量、提高视觉单词对图像内容的表达能力,本文首先利用 E^2LSH 算法对图像初始关键点进行过滤,降低噪声点的影响;然后,引入卡方模型统计各视觉单词与目标类别的相关性,并结合单词词频信息移除词典中的视觉停用词;最后,为了确保度量的准确性,采用空间一致性度量准则进行相似性度量以弥补传统视觉词典模型中单词空间关系缺失降低量化误差并对初始检索结果进行 K-近邻重排序.实验结果有效地验证了本文方法的有效性.

需要注意的是,在今后需要研究如何降低 E^2LSH 算法的随机性问题来提高过滤效果的鲁棒性.此外,如何通过距离度量的学习使得特征空间的距离更加接近真实的语义距离也是今后亟待解决的问题.

参 考 文 献

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

[2] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos [C] //Proc of the 9th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2003: 1470-1477

[3] Jégou H, Douze M, Schmid C. Improving bag-of-features for large scale image search [J]. Computer Vision, 2010, 87(3): 316-336

[4] Ji Chuanjun, Liu Zuotao, Chan Wen, et al. Context modeling based automatic image annotation system [J]. Journal of Computer Research and Development, 2011, 48(1): 441-445 (in Chinese)
(纪传俊, 刘作涛, 产文, 等. 一个基于语义上下文建模的图像自动标注系统[J]. 计算机研究与发展, 2011, 48(1): 441-445)

[5] 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(12): 935-948

[6] Wang J Y, Bensmail H, Gao X. Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification [J]. Pattern Recognition, 2013, 46(12): 3249-3255

[7] Cao Y, Chang H W, Zhiwei L, et al. Spatial-bag-of-features [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3352-3359

[8] Zhu Jun, Zhao Jieyu, Dong Zhenyu. Image classification using hierarchical feature learing method combined with image silency [J]. Journal of Computer Research and Development, 2014, 51(9): 1919-1928 (in Chinese)
(祝军, 赵杰煜, 董振宇. 融合显著信息的层次特征学习图像分类[J]. 计算机研究与发展, 2014, 51(9): 1919-1928)

[9] Li Dai, Sun Xiaoyan, Wu Feng, et al. Large scale image retrieval with visual groups [C] //Proc of the 20th IEEE Conf on Image Processing. Piscataway, NJ: IEEE, 2013: 2582-2586

[10] Chen Tao, Yap K H, Zhang Dajiang. Discriminative soft bag-of-visual phrase for mobile landmark recognition [J]. IEEE Trans on Multimedia, 2014, 16(3): 612-622

[11] Rudinac M, Lenseigne B, Jonker P. Keypoint extraction and selection for object recognition [C] //Proc of the 8th IEEE Conf on Machine Vision Applications. Piscataway, NJ: IEEE, 2009: 191-194

[12] Jamshy S, Krupka E, Yeshurun Y. Reducing keypoint database size [C] //Proc of the 15th Int Conf on Image Analysis and Processing. Berlin: Springer, 2009: 113-122

[13] Tirilly P, Claveau V, Gros P. Language modeling for bag of visual words image categorization [C] //Proc of the 2008 Int Conf on Content-based Image and Video Retrieval. New York: ACM, 2008: 249-258

[14] Yuan J, Wu Y, Yang M. Discovery of collocation patterns: From visual words to visual phrases [C] //Proc of the 20th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2007: 1-8

[15] Liu Shuoyan, Xu De, Feng Songhe, et al. A novel visual words definition algorithm of image patch based on contextual semantic information [J]. Acta Electronica Sinica, 2010, 38(5): 1156-1161 (in Chinese)
(刘硕研, 须德, 冯松鹤, 等. 一种基于上下文语义信息的图像视觉单词生成算法[J]. 电子学报, 2010, 38(5): 1156-1161)

[16] Zhang Ruijie, Li Bicheng, Wei Fushan. Image scene classification based on multi-scale and contextual semantic information [J]. Acta Electronica Sinica, 2014, 42(4): 646-652 (in Chinese)
(张瑞杰, 李弼程, 魏福山. 基于多尺度上下文语义信息的图像场景分类算法[J]. 电子学报, 2014, 42(4): 646-652)

[17] Yeh J B, Wu C H. Extraction of robust visual phrases using graph mining for image retrieval [C] //Proc of 2010 IEEE Int Conf on Multimedia and Expo (ICME 2010). Piscataway, NJ: IEEE, 2010: 3681-3684

[18] Van Gemert J C, Veenman C J, Smeulders A W M, et al. Visual word ambiguity [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 7(32): 1271-1283

[19] Otávio A B P, Fernanda B S, Eduardo V, et al. Visual word spatial arrangement for image retrieval and classification [J]. Pattern Recognition, 2014, 47(1): 705-720

[20] Yang Linjun, Geng Bo, Cai Yang, et al. Object retrieval using visual query context [J]. IEEE Trans on Multimedia, 2012, 13(6): 1295-1307

[21] Xie Hongtao, Zhang Yongdong, Tan Jianlong, et al. Contextual query expansion for image retrieval [J]. IEEE Trans on Multimedia, 2014, 56(99): 1-32

[22] Slaney M, Casey M. Locality-sensitive hashing for finding nearest neighbors [J]. IEEE Signal Processing Magazine, 2008, 8(3): 128-131

[23] Robotics Research Group. Oxford5K dataset [DB/OL]. [2014-03-26]. http://www.robots.ox.ac.uk/_vgg/data/oxbuildings

[24] Yahoo Company. Flickr1 dataset [DB/OL]. [2014-03-24]. http://www.flickr.com

Zhao Yongwei, born in 1988. PhD, lecturer. His research interests include image analysis and processing.

Zhou Yuan, born in 1978. Master, lecturer. Her research interests include image processing and multimedia technology.

Li Bicheng, born in 1970. PhD, professor. His research interests include data mining and artificial intelligence processing.