一种基于多级空间视觉词典集体的图像分类方法

罗会兰[1],郭敏杰[1],孔繁胜[2]
(1.江西理工大学信息工程学院,江西赣州 341000;2.浙江大学计算机科学技术学院,浙江杭州 310027)

摘 要: 针对单一特征时存在提取的信息量不足,对图像内容描述比较片面,提出将传统的SIFT特征与KDES-G特征进行串行融合,生成一个联合向量作为新的特征向量.针对传统的视觉词典构造方法缺乏考虑视觉词汇在空间的分布特点,本文引入图像空间信息,提出了一种空间视觉词典的构造方法,先对图像进行空间金字塔划分,再把空间各子区域内的特征分别聚类,构建属于对应子空间区域的空间视觉词典.在图像表示阶段,图像各子区域内的特征基于其对应的空间视觉词典进行LLC稀疏编码,根据各子区域对图像贡献程度的不同,把编码后各子区域的特征向量赋予不同的权重加权处理,再连接形成最终的图像描述.最后,利用线性SVM进行图像分类,实验结果表明了本文方法的有效性和鲁棒性.

关键词: 图像分类; 特征融合; 空间视觉词典; LLC编码; 加权处理
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)04-0684-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.04.009

An Image Classification Method Based on Multiple Level Spatial Visual Dictionary Ensemble

LUO Hui-lan[1], GUO Min-jie[1], KONG Fan-sheng[2]
(1. *School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*;
2. *School of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang* 310027, *China*)

Abstract: Using single feature of images to describe the image content is one-sided because of insufficient information. To solve this problem, combing SIFT and KDES-G features to describe images is proposed by generating a joint vector as a new feature vector. Considering images' spatial information, an image classification method based on the spatial visual dictionary is proposed. Images are first divided into sub-regions according to spatial pyramid, and the spatial visual dictionaries are respectively constructed by grouping features of each region into a number of clusters. The features of each region are coded by LLC based on its corresponding dictionary, and then the coded feature vectors are given different weights according to the different contribution of each region. After that, the feature vectors of different regions are concatenated and regarded as the final image description. Finally, a linear SVM is used to classify images. Experimental results show that the proposed method has better performance and robustness compared with some state-of-the-art works.

Key words: image classification; feature fusion; spatial visual dictionary; local constrained linear coding (LLC); weighted processing

1 引言

图像分类是一种模式分类问题,需要根据图像中语义内容判别一幅图像中是否出现某类对象.由于图像内的对象存在光照、尺度及视角的变化、对象之间差异较大、物体变形、遮挡和背景嘈杂等多种因素的影响,使得图像分类一直以来都是计算机视觉领域一个具有挑战性的难题,许多图像特征描述和分类技术因此得到迅速发展.

近年来,基于局部语义概念的图像分类方法成为最主流的方法.这类方法大多使用视觉特征袋(Bag-Of-Features, BOF)[1]和主题分布等对图像内容进行表述,有效弥合底层特征与高层语义之间的语义鸿沟[2].为了更好的利用视觉词典中视觉单词的上下文信息,Lazebnik等[3]提出了空间金字塔匹配(Spatial Pyramid Matching, SPM)方法,该分类方法处理的关键是对图像进行视觉

收稿日期:2013-08-05;修回日期:2014-04-23;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61105042, No.61462035);国家973重点基础研究发展计划(No.2010CB327900);江西省教育厅科技项目(No.GJJ13421)

特征提取、视觉词典的构造以及特征编码的过程.但是,这三方面的图像处理仍存在很多困难,主要体现为:(1)难以提取有效的视觉特征;(2)视觉词典的生成缺乏考虑视觉单词在空间分布的特点;(3)特征编码时,难以选择合适的特征向量组织方式,使之既便于计算又能减少图像的信息丢失.

一些图像分类方法只提取图像单方面特征,得到的特征信息在很多情况下都不足以充分表述图像.在文献[4]中,汪成亮等提出了一种基于主成分分析降维的方向梯度直方图(Histogram of Oriented Gradient,HOG)[5]描述子来提高识别率和分类速度.在文献[6]中,张静等提出利用群体兴趣点形成图像区域描述,再采用 Hough 变换方法来提取出图像中的有效信息.这些方法都只提取了图像单方面的特征,得到的图像分类准确度都不高.近年来,在文献[7]中,Harzallah 等结合了尺度不变特征转换(Scale-Invariant Feature Transform,SIFT)[8]和 HOG 特征,提取得到稠密的特征空间,使得图像的特征表述更完善.在文献[9]中,高常鑫等通过整合局部特征和滤波器特征获得丰富的表征信息.在文献[10]中,程刚等充分利用整体结构特征和局部纹理特征的优势,利用两级分类器融合这两种特征.这些方法都在分类效果上比单一特征提取方法有很大的提高.

传统的视觉词典构造方法缺乏考虑视觉词汇在空间的分布特点,所以,一些研究者开始考虑将图像的空间信息引入视觉词典的构造中.在文献[11]中,赵永威等提出了一种支持动态扩充的随机化词典组,增强了目标对象的可区分性.在文献[12]中,王宇新等将图像进行划分,对各子区域进行特征提取和聚类,进而构建整个训练图像集的空间视觉词袋模型.在文献[13]中,刘硕研等提出一种基于上下文语义信息的图像块视觉单词生成方法,在一定程度上提高了视觉单词的区分性.

在获得了有效的特征和视觉词典后,需要将它们组织起来表示图像,这就是特征编码过程.在文献[14]中,Wang 等采用稀疏编码方式对图像进行描述,该方法基于空间金字塔划分后再进行稀疏编码,将得到的特征向量直接相连构成一个向量,再用该向量来表述图像.在文献[15]中,亓晓振等提出用多个稀疏向量来共同表示一幅图像,该方法通过在空间上对图像进行金字塔划分,对每个金字塔层次分别稀疏编码,并将该层次特征转化为一个向量表示.这些稀疏编码方法都获得了对分类区分能力较好的图像表述.

鉴于此,本文提出了一种融合多种特征及空间信息的图像分类方法,首先提取图像的多个特征,将不同的特征结合起来,生成一个联合向量作为新的特征向量.在构造视觉词典时,提出了一种多级空间视觉词典集体构造方法,构造不同层级上的视觉词典,从全局视觉词典一直到划分比较细的子空间视觉词典,从而可以综合不同层级或不同粒度的图像信息进行分类.在构建了从全局视觉词典到不同细分程度的空间子区域视觉词典后,将图像位于不同层次,不同空间位置的特征基于其对应的空间视觉词典进行稀疏编码.得到量化特征后,再根据各层次,各子区域面积的大小赋予对应量化特征不同的权重.最后将所有加权量化特征向量连接形成最终的图像描述.虽然文献[12]也采用图像划分手段得到多个局部空间视觉词典,但文献[12]的工作只利用了单级划分上的局部空间视觉词典进行量化分类,没有考虑到综合利用多个划分级别上的信息.而且文献[12]在后续的量化时,将各子区域简单统计量化后得到直方图特征进行距离度量,然后将全部子区域的距离度量值相加得到两副图像间的距离.本文提出的方法融合了不同特征,不同粒度信息和局部空间信息对图像进行分类,并在融合不同层级的信息时考虑到特征对识别的贡献程度进行了加权处理.第二小节详细论述了本文提出的方法;第三小节通过在两个标准数据集上的实验,验证了本文方法的有效性和鲁棒性;最后阐述了结论.

2　多级空间视觉词典集体用于图像分类

本文提出了一种基于多级空间视觉词典集体的图像分类方法,融合多种特征来分类图像,该方法命名为 FFSVD(Method Based on Feature Fusion and Spatial Visual Dictionary).FFSVD 算法有三个主要创新点,首先,提出了对 SIFT 特征和 KDES-G[16]两种特征进行串行融合,将两组特征融合形成最终的特征向量.其次,提出了多级空间视觉词典集体的构造方法.最后,在图像表示阶段,根据各区域的特征向量贡献大小,赋予它们不同的权重值,连接形成最终的图像描述.

FFSVD 算法的结构框图如图 1 所示.首先,提取所有图像的 SIFT 特征与 KDES-G 特征,对这两组特征进行串行融合,生成一个联合向量作为新的特征向量.然后,在构造图像的空间视觉词典时,先对图像进行空间金字塔划分,再把空间各子区域内的特征分别聚类,构建不同层级,不同子空间区域的空间视觉词典.最后,在图像表示阶段,图像不同层上各子区域内的特征基于其对应的空间视觉词典进行 LLC 稀疏编码,根据各子区域对图像贡献程度的不同赋予不同的权重加权处理,并将所有子区域内特征向量连接形成最终的图像描述.得到的图像描述则直接用于线性 SVM 训练与分类,得到最终的分类结果.

2.1　串行特征融合

将不同特征进行串行融合是一种比较有效的提高

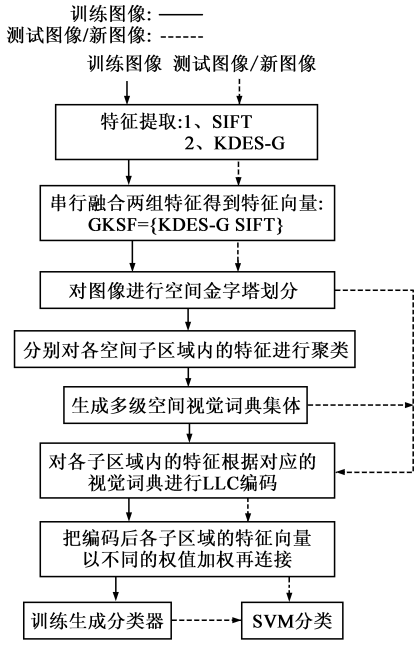

图1 FFSVD算法结构框图

图像分类效果的方法,在文献[17]中,将目标的颜色、纹理、边缘、运动特征统一使用直方图模型进行描述,再将这四种特征通过有效的融合系数进行概率融合,该方法对复杂背景下的跟踪具有较高的鲁棒性.在文献[18]中,将基于方向场特征、基于灰度共生矩阵的纹理特征、基于 LBP 算子的纹理特征的方法和基于细节点特征的方法进行融合,弥补了各个方法的不足,提高了匹配的准确性.在这两个文献中,作者都是通过将多个特征进行融合来提高性能的.

鉴于此,本文也提出将 SIFT 特征与 KDES-G 特征进行串行融合,旨在特征维数增加相对较少的情况下包含更多的图像信息.由于 SIFT 对特征点的尺度、位置、旋转和光照等因素的变化不敏感,具有很强的鲁棒性,已经广泛地应用于计算机视觉的多个方面.而梯度核描述子 KDES-G[16]能将任意类型的基于像素属性(如梯度、颜色和形状)的特征转换为紧密的图像块级的特征,它可以很好的表征局部区域内目标的梯度结构.梯度核描述子 KDES-G[16]首次将核函数的方法用于特征提取.该方法先将图像转化为灰度图像,并逐像素计算其梯度,梯度核描述子的梯度核函数如式(1)所示.

$$K_{\text{grad}}(P,Q)=\sum_{x_P\in P}\sum_{x_Q\in Q}\widetilde{m}(x_P)\widetilde{m}(x_Q)\cdot k_O(\tilde{\theta}(x_P),\tilde{\theta}(x_Q))k_p(x_P,x_Q) \tag{1}$$

其中,P 和 Q 是不同的图像块,x_P 和 x_Q 分别表示图像块 P 和 Q 中的像素.$K_{\text{grad}}(P,Q)$是一个标准的线性核函数,它使用梯度大小 $\widetilde{m}(x)$来衡量每个像素的贡献程度,并结合 $k_p(x_P,x_Q)$和 $k_O(\tilde{\theta}(x_P),\tilde{\theta}(x_Q))$两个核函数来共同提取图像的特征.$k_p(x_P,x_Q)$是高斯位置核函数,用来衡量图像中两个像素间的空间距离.$k_O(\tilde{\theta}(x_P),\tilde{\theta}(x_Q))$是高斯方向核函数,用来计算两个像素梯度方向的相似度.两个像素梯度方向的角度值 $\theta(x)\in[0,2\pi]$在某些情况下会产生错误的相似度,如两个角度值分别为 $2\pi-0.01$ 和 -0.01 时,它们的梯度方向很相似,但空间距离却很大.所以,为了避免像素点 x_P 和 x_Q 梯度方向产生错误的相似度,定义式(2)来标准化核函数 $k_O(\tilde{\theta}(x_P),\tilde{\theta}(x_Q))$中的梯度方向 $\tilde{\theta}(x)$.

$$\tilde{\theta}(x)=[\sin(\theta(x))\ \cos(\theta(x))] \tag{2}$$

梯度大小 $\widetilde{m}(x_P)$和 $\widetilde{m}(x_Q)$的计算如式(3)所示.$m(x_P)$、$m(x_Q)$分别是图像块 P 和 Q 中像素 x_P、x_Q 坐标值的几何均值,ε 是一个很小的正数保证分母大于 0.

$$\widetilde{m}(x_P)=m(x_P)/\sqrt{\sum_{x_p\in P}m(x_P)^2+\varepsilon}$$
$$\widetilde{m}(x_Q)=m(x_Q)/\sqrt{\sum_{x_Q\in Q}m(x_Q)^2+\varepsilon} \tag{3}$$

由于 KDES-G 和 SIFT 特征提取的过程都涉及到对图像块中像素的梯度处理过程,考虑到 SIFT 获得的图像表象特征和 KDES-G 获取的目标梯度结构都有良好的表征能力,所以本文采用这两种特征融合来弥补使用单一特征时造成的信息量不足的缺点.对于 SIFT 特征,使用 4×4 分块,8 个方向的 SIFT 描述子,这样得到的特征具有的维数为 $4\times4\times8=128$ 维.对于 KDES-G 特征,作者利用核主成分分析(KPCA)方法选择的最优的特征维数为 200.最终得到的组合特征向量 **GKSF**(Gradient Kernel Descriptor and SIFT Features)如式(4)所示.

$$\textbf{GKSF}=\{\text{KDES-G weight}\times\text{SIFT}\} \tag{4}$$

其中,weight 是特征融合的权值,由于 SIFT 特征和 KDES-G 的采样方式相同即采样间隔均为 8 像素,图像块为 16×16 像素,而且为了降低计算复杂度,简单把权值 weight 设置为 1,即直接对 SIFT 和 KDES-G 特征串行连接.由于 SIFT 特征维数为 128,KDES-G 特征维数为 200,所以最终通过串行融合这两组特征,得到组合特征向量 **GKSF** 的总维数为 $200+128=328$ 维.

2.2 多级空间视觉词典集体

Lazebnik[3]提出了空间金字塔模型(SPM),该模型的本质是对图像在空间上进行划分,通常在图像的两个坐标方向进行 2 的指数次划分,即 $2^l\times2^l$($l=0,1,\cdots,L$).Lazebnik[3]和 Wang[14]在利用空间金字塔模型时,首先提取图像特征,构造全局视觉词典,再将图像特征空间划分为 $2^l\times2^l$ 的空间子区域,每个空间子区域中的局部特征基于全局视觉词典量化后生成直方图,最后将各区域中的直方图串接形成最后的图像描述.考虑到空间信息的引入虽然在一定程度上等于引入了目标与背景间的位置信息,或者说使得目标与背

景的关系信息更多的表达出来,但是当目标类内出现很大背景差异,或者测试图像中出现了不常见的背景,或者目标尺度及位置极端化的情况时,这种空间信息表达可能会使得分类结果更差.

鉴于此,本文提出了一种多级空间视觉词典集体的方法,在不同划分级别下的空间子区域上分别构建视觉词典,建立从全局视觉词典到不同细分程度的子空间视觉词典.如图 2 所示,本文首先有层次的把图像进行空间金字塔划分,再把不同划分级别中的各空间子区域内的特征向量分别聚类,构建属于各子空间区域的空间视觉词典.图 2 中,l 表示对图像划分的层数序号,空间金字塔在水平方向把图像划分成 2^l 块,在竖直方向将图像划分成 2^l 块,最终将图像划分成 $2^l \times 2^l$ 块.因此,定义 $S_{(l,i)}$ 为图像第 l 层划分下的第 i 个子空间区域,N_l 为第 l 层划分下子空间区域的总数,由此可知,$N_l = 4^l$.所以当空间金字塔划分总层次为 L 时,需要构建的空间视觉词典个数 $\varphi(D)$ 如式(5)所示.

$$\varphi(D) = \sum_{l=0}^{L} 4^l \tag{5}$$

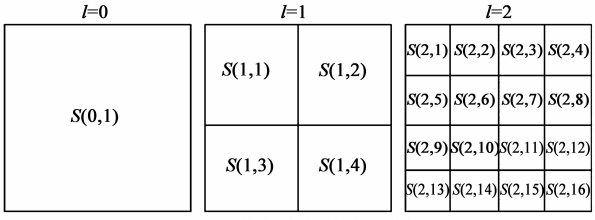

图2　图像的空间金字塔划分

由 2.1 节可知,GKSF 是维度 D 为 328 维的联合特征向量,假设第 j 幅图像的第 $S_{(l,i)}$ 子区域提取得到了 N_u 个 **GKSF** 特征向量,即 $\boldsymbol{F}^j_{S(l,i)} = \{x_1, x_2, \cdots, x_{Nu}\} \in R^{D \times Nu}$.则训练图像集大小为 N 时,在区域 $S_{(l,i)}$ 内提取到的特征向量 **GKSF** 集如式(6)所示,那么区域 $S_{(l,i)}$ 内 N 幅训练图像总特征向量维数为 $N \times N_u$.

$$\begin{aligned}\boldsymbol{X}_{S(l,i)} &= \{\boldsymbol{F}^1_{S(l,i)}, \cdots, \boldsymbol{F}^j_{S(l,i)}, \cdots, \boldsymbol{F}^N_{S(l,i)}\} \\ &= \{x^1_{s(l,i)}, x^2_{s(l,i)}, \cdots, x^r_{s(l,i)}, \cdots, x^{N \times N_u}_{s(l,i)}\} \\ i &= 1,2,\cdots,4^l, l = 0,1,\cdots,L\end{aligned} \tag{6}$$

其中,N 为训练图像的总数目,l 为金字塔划分层数序号,i 为第 l 层划分下的第 i 个子空间区域.分别对 $S_{(l,i)}$ 子区域内的 **GKSF** 特征向量集 $\boldsymbol{X}_{S(l,i)}$ 进行聚类,生成相应子区域的空间视觉词典 $d_{S(l,i)}$,则最终构建的多级空间视觉词典集体 Dic 如式(7)所示.

$$\text{Dic} = \{d_{s(l,i)}\}, i = 1, \cdots, 4^l \tag{7}$$

多级空间视觉词典集体的构建过程示意图如图 3 所示.

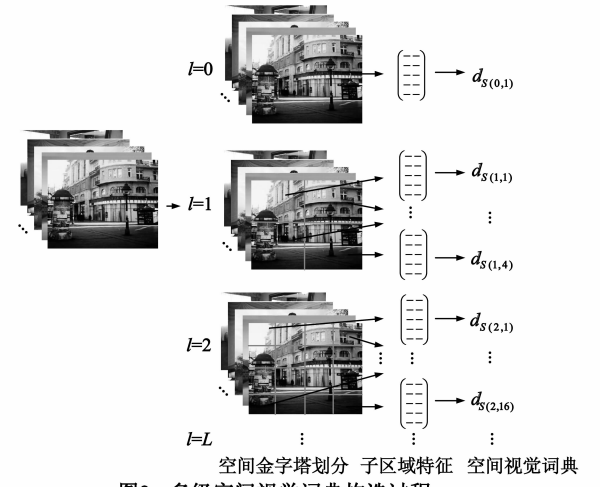

图3　多级空间视觉词典构造过程

2.3　图像描述

由 2.2 小节可知,$\boldsymbol{F}^j_{S(l,i)}$ 是第 j 幅图像的第 $S_{(l,i)}$ 子区域内的 328 维的联合特征向量 **GKSF** 的集合.假设视觉词典的大小为 M,$d_{S(l,i)} = \{b_1, b_2, \cdots, b_M\} \in R^{D \times M}$ 为子区域 $S_{(l,i)}$ 的空间视觉词典,其中 $d_{S(l,i)}$ 内的每个元素 b_m 称为视觉单词.在图像表示阶段,本文将图像 j 各子区域内的特征向量 $\boldsymbol{F}^j_{S(l,i)}$ 基于其对应的空间视觉词 $d_{S(l,i)}$ 分别进行 LLC 编码[14],如式(8)和(9)所示,得到图像表示:

$$h^j_{S(l,i)} = \{y_1, y_2, \cdots, y_{N_u}\} \in R^{M \times N_u},$$

其中 $\boldsymbol{y}_k$ 是一个 M 维的列向量.

$$\arg\min_{h_{S(l,i)}} \sum_{k=1}^{N_u} \| x_k - d_{s(l,i)}\boldsymbol{y}_k \|^2 + \lambda \| D^j_{s(l,i)} \Theta \boldsymbol{y}_k \|^2 \tag{8}$$

$$D^j_{s(l,i)} = \exp\left(\frac{\text{dist}(x_k, d_{s(l,i)})}{\sigma}\right) \tag{9}$$

其中,$\text{dist}(x_k, d_{s(l,i)}) = [\text{dist}(x_k, b_1), \cdots, \text{dist}(x_k, b_M)]$,dist 函数表示图像 j 的子区域 $S_{(l,i)}$ 内特征向量 $\boldsymbol{F}^j_{S(l,i)}$ 中的特征 x_k 和视觉单词 b_m 间的欧氏距离,Θ 表示向量内元素两两相乘,λ 为重构系数.

Lazebnik[3] 在处理多层金字塔之间的核矩阵系数时,三层金字塔结构的核矩阵系数分别设为[0.5,0.25,0.25],而 Wang[14] 直接对稀疏编码后的向量串接组成最终向量,再利用该向量做 SVM 分类.本文提出的 FFSVD 算法则根据各子区域对图像贡献程度的不同,即各子区域面积所占比例的大小,把编码后各子区域的特征向量加权处理并连接.该思想的动机是:一个基于局部信息得到的量化特征所包含的信息量应该比基于全局信息得到的量化特征所包含的信息量要小,例如,通常情况下在只看到四分之一图像时做出的识别决定应该比看到了完整图像做出的识别决定更不可靠,但是比只看到十六分之一图像做出的识别决定要更可靠.因

此,利用区域面积作为加权的一个权重,得到图像 j 的最终描述如式(10)所示.

$$H^j=\bigcup_{l=0}^{L}\left(W_l\cdot\bigcup_{i}^{4^l}h^j_{S(l,i)}\right),\ W_l=1/4^l,\ l=0,\cdots,L \quad (10)$$

式(10)中,$\cup$ 表示特征向量之间的连接,W_l 为不同的金字塔层次的权值,当图像划分的层次 $l=0$ 时,权值 $W_0=1/4^0=1$;层次 $l=1$ 时,权值 $W_1=1/4^1=1/4$,其他层次的权值计算则依此类推.在第三节的实验中,FFSVD 算法的空间金字塔划分层数 L 设为 2,所以,计算得到权值向量

$$\boldsymbol{W}=[W_0,W_1,W_2]=[1,1/4,1/16]$$

2.4 FFSVD 算法

FFSVD 算法训练生成分类器的具体过程描述如算法 1 所示.

算法 1 FFSVD 算法训练生成分类器

输入:训练图像集

输出:分类器 C

Begin

1:特征提取:

1)分别提取图像特征:SIFT 和 KDES-G.

2)根据式(4)将两组特征串行融合得到联合特征向量 **GKSF**.

2:对所有图像进行 L 层次的空间金字塔划分.

3:空间视觉词典的生成:

(1)根据式(6)计算子区域内的 **GKSF**,得到 $\boldsymbol{X}_{S(l,i)}$.

(2)分别对子区域 $S_{(l,i)}$ 内的 $\boldsymbol{X}_{S(l,i)}$ 进行聚类,生成空间视觉词典 $d_{S(l,i)}$.

4:图像表示阶段:

(1)将图像 j 各子区域内的特征 $\boldsymbol{F}^j_{S(l,i)}$ 与其对应的空间视觉词典 $d_{S(l,i)}$ 根据式(8)和式(9)编码得到 $h^j_{S(l,i)}$.

(2)根据式(10)计算得到最终的图像描述 $\boldsymbol{H}=\{H^j,j=1,\cdots,N\}$.

5:将 $\boldsymbol{H}$ 用于线性 SVM,训练并得到分类器 C

End

3 实验与分析

为了验证 FFSVD 算法的有效性和鲁棒性,本文采用了目前图像分类实验上最常用的两个数据集:一个目标分类数据集 Caltech101[19] 和一个场景分类数据集 15 Scenes[3].在这些数据集上分别验证了本文 FFSVD 算法的分类性能,并进一步分析了不同的划分层次、图像表示阶段的权值大小以及空间视觉词典的大小对分类性能的影响.

3.1 实验数据集

图 4 示例了实验中用到的两组数据集 Caltech101 和 15 Scenes 的部分图像.

(1)Caltech101 数据集

Caltech101 数据集共有 9144 幅图像,包括 101 个对象类和一个背景类.每个类的图像数目从 31 到 800 不等,目标一般位于每幅图像的中间并且占据图像的大部分,背景图像与目标类的差异很大.

(2)15 Scenes 数据集

15 Scenes 数据集是一个使用较为广泛的场景识别数据集,它包含 15 个类别的场景,共有 4485 幅图像.每个类的图像数目为 200 到 400 不等,含有建筑、城市、卧室、街道等 15 个场景,每幅图像的平均大小在 200 个像素左右.

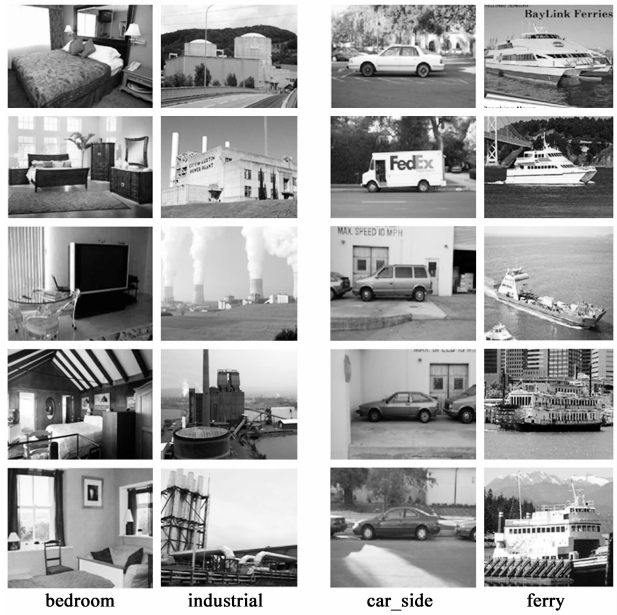

图4 实验数据集中的部分图像

3.2 实验设置

首先将所有图像都处理为灰度图像,然后将图像调整为不超过 300×300 像素大小.在选择训练集和测试集上,分别从每个类中随机选择 n 幅作为训练图像,剩下的作为测试图像.对于 Caltech101 数据集,在实验中 n 分别设为 10、15、20、25、30;15 Scenes 数据集的 n 分别设为 10、20、30、60、100.在同一数据集上的实验,重复 10 次,每次随机选择训练图像和测试图像,论文中报导的结果是 10 次结果的平均值.

对于特征的提取,实验中提取 SIFT 特征与 KDES-G 特征的采样间隔均为 8 像素,图像块为 16×16 像素.产生空间视觉词典时,分别从每幅训练图像的相应区域提取得到的联合特征 **GKSF** 集中随机选择不超过 70 个特征,构成特征集用于生成此区域空间视觉词典.综合考虑到性能与效率因素,FFSVD 算法中空间金字塔层数 L 设置成 2,视觉词典的大小均设为 1000.聚类方法使用的是 k-means 算法,支持向量机采用 LIBSVM 工具包的 libsvm3.12 工具箱[20].

3.3 实验结果

为了验证本文提出的 FFSVD 算法具有良好的分类性能,本文在两个数据集上分别和不同研究者的实验结果进行了对比.

3.3.1 Caltech101 数据集上分类准确率对比

首先,在 Caltech101 数据集上,分析比较了 FFSVD 算法与当前分类准确率较高的一些研究者的方法,如 Lazebnik[3]的空间金字塔匹配模型、Wang[14]的稀疏编码、亓晓振[15]的多核学习方法以及 Liefeng Bo[16]的核描述子的方法.在每类分别使用 10~30 幅训练图像时的实验结果如表 1 所示.从表 1 可以看出,FFSVD 算法在每类训练图像个数不同的情况下,都获得了比其他方法更高的准确率.

表 1　在 Caltech101 数据集上的图像分类准确率(%)对比

每类训练图像个数(n)	Lazebnik[3]	Wang[14]	亓晓振[15]	Liefeng Bo[16]	FFSVD 算法
10	– –	59.77	60.88±0.8	– –	65.78
15	56.40	65.43	65.42±0.4	– –	69.90
20	– –	67.74	69.38±0.5	– –	72.00
25	– –	70.16	70.54±0.6	– –	74.12
30	64.60	73.44	73.58±1.4	76.40±0.7	77.94

Liefeng Bo 在文献[16]中提出了核描述子的方法,它提出将核函数的方法应用于特征提取,并将任意像素属性(如梯度、颜色和形状等)特征转换为紧密的图像块级的特征,当仅使用梯度描述子 KDES-G 时,在每类训练图像的个数为 30 时获得准确率为 75.2%,表 1 中报告的则是它将梯度、颜色和形状三种核描述子集成再经过 EMK[21]编码后得到的准确率 76.4%.而本文提出的 FFSVD 算法将 SIFT 与 KDES-G 特征简单串行融合,在构造了有效的多级空间视觉词典集体的基础上,在每类训练图像的个数为 30 时获得了约 77.94% 的准确率,较 Liefeng Bo[16]的集成结果高出了近 1.5%,比他的单个特征 KDES-G 时的准确率提高了约 2.8%.FFSVD 算法比亓晓振[15]的多核学习方法提高了约 4.4%.由此可知,FFSVD 算法是一种能充分和有效融合多特征和空间信息的方法,具有很好的分类性能.

然后,为了验证 FFSVD 算法构造的空间视觉词典集体和特征向量加权处理的有效性,本文还单独与 Wang[14]的方法进行了对比.在仅利用 SIFT 特征时,Wang[14]的方法基于传统的 SPM,对 LLC 编码后的特征向量只是简单串接,在每类训练图像的个数为 30 时得到的准确率为 73.44%.而 FFSVD 算法基于 SPM 构造多级空间视觉词典集体,根据各子区域对图像贡献程度的不同,将 LLC 编码后各子区域的特征向量加权处理,得到的准确率为 76.67%,比 Wang[14]提高了约 3.3%.这说明,FFSVD 算法中构造的多级空间视觉词典集体和对特征向量的加权处理方法获得了显著的效果.

3.3.2 15 Scenes 数据集上分类准确率对比

本实验在 15 Scenes 数据集上分析比较了 FFSVD 算法与当前分类准确率较高的一些研究者的方法:Lazebnik[3]的空间金字塔匹配模型、Yang[22]的稀疏编码、亓晓振[15]的多核学习方法以及 Liefeng Bo[16]的核描述子方法,实验结果如表 2 所示.从 Liefeng Bo 的文献[16]可知,作者将三种特征集成并用线性 SVM 进行训练获得了约 81.9% 的准确率,而在拉普拉斯核函数的 SVM 上获得了高达 86.7% 的准确率,由于本文使用的是线性 SVM,所以选择了文献[16]中与本文实验环境最相近的一组实验结果即 81.9% 做对比.由表 2 可以看出,FFSVD 算法比 Liefeng Bo[16]线性 SVM 上的集成结果在每类训练图像个数为 100 时高出约 1.0%.

FFSVD 算法的分类准确率在每类训练图像的个数为 100 时比 Lazebnik[3]的基于词典树和空间金字塔划分方法有了大幅度的提高(提高了 11.51%).比 Yang[22]的基于稀疏编码的方法提高了约 3.5%.亓晓振[15]的多核学习方法在每类训图像个数为 100 时准确率为 83.1%,FFSVD 算法得到的准确率为 83.71%,虽然相对于亓晓振[15]的没有明显优势,但是亓晓振[15]的多核学习需要解决凸优化问题来求解各个核矩阵的权重,导致运行速度慢、消耗的时间较长.而 FFSVD 算法只使用了简单的线性 SVM,实现简单,运行速度快.FFSVD 算法在其他训练图像集大小情况下也获得了较亓晓振[15]高的分类准确率.

表 2　在 15 Scenes 数据集上的图像分类准确率(%)对比

每类训练图像个数(n)	Lazebnik[3]	Yang[22]	亓晓振[15]	Liefeng Bo[16]	FFSVD 算法
10	– –	– –	67.10±1.4	– –	68.91
20	– –	– –	72.71±1.4	– –	73.64
30	– –	– –	75.33±0.5	– –	76.98
60	– –	– –	79.81±0.8	– –	81.23
100	72.20	80.28	83.1±0.7	81.9±0.6	83.71

同时,为了比较本文提出的方法与文献[12]中的方法,在 Oliva 和 Torralb 提供的 8 类场景数据库[23]上进行了实验对比.在相同的实验设置下,文献[12]在空间划分层次 4 时,空间相似度被最大程度地挖掘出来,得到了 63.40% 的准确率.而本文提出的算法 FFSVD 在 $L=2$ 时建立了从全局到不同细分程度的局部空间视觉字典共 21 个,获得了 86.388% 的分类准确率,这说明本文提出的方法可以综合利用不同粒度层次的信息,使得识别效果显著提升.

3.3.3 特征融合方式对 FFSVD 算法分类性能的影响

本实验分析了不同的特征融合方式对 FFSVD 算法的分类性能影响.为了验证本文的 SIFT 与 KDES-G 特征串行融合的方法具有性能提升的潜力,本文进一步实验了不同特征融合方式下对 FFSVD 算法的影响,实验的主要内容包括:

(1)方式 1:只用 SIFT 特征生成空间视觉词典,再进行 LLC 编码.

(2)方式 2:只用 KDES-G 特征生成空间视觉词典,再进行 LLC 编码.

(3)方式 3:分别利用两种特征生成空间视觉词典,再根据两种特征的视觉词典分别利用 LLC 编码,然后把编码后的特征进行串接的方法.

(4)方式 4:直接串行融合 SIFT 与 KDES-G 特征(FFSVD 算法).

实验结果如表 3 所示.在表 3 中,Caltech101 数据集(30)表示 Caltech101 数据集中分别从每个类中随机选择 30 幅作为训练图像.15 Scenes(100)表示每类训练图像为 100 的 15 Scenes 数据集.从表 3 的实验结果可以看出,单独使用 SIFT 特征或 KDES-G 特征时获得的分类准确率在两个数据集上都比串行融合两个特征(FFSVD 算法)要低约 1.5%,由此可知,串行融合 SIFT 和 KDES-G 特征比单独使用一种特征时效果更好.方式 3 获得的分类准确率与本文方法结果相近,但采用每种特征分别生成视觉词典的方法,造成视觉词典过多,后续加权连接形成的最终图像特征维度过高.

表 3 不同特征融合方式在各数据集上的分类准确率(%)对比

不同特征融合方式	Caltech101 数据集(30)	15 Scenes(100)
方式 1	76.67	81.68
方式 2	77.05	82.56
方式 3	77.89	83.77
方式 4(FFSVD)	77.94	83.71

3.3.4 空间视觉词典中层次 L 对 FFSVD 算法分类性能的影响

本实验分析了在不同金字塔层次 L 时,构造的空间视觉词典对 FFSVD 算法分类性能的影响.首先,为了验证 FFSVD 算法构造的空间视觉词典中的空间信息的有效性,在 Caltech101 数据集上,将 FFSVD 算法与分别只利用第 0 层,第 1 层和第 2 层信息进行分类的结果进行了分析对比.图 5 显示的是本文 FFSVD 算法在 Caltech101 数据集上每类分别使用 10 ~ 30 幅训练图像时的分类结果.

从图 5 中可以看出,没有加入空间信息即第 0 层($l=0$)时的方法和 Lazebnik[3] 及 Wang[14] 的一样,训练得到了一个全局视觉词典,从图 5 中可以看出该视觉词典对数据集的图像分类能力很弱.只利用第 1 层($l=1$)时训练得到 $N_l=4^1=4$ 个视觉词典,只利用第 2 层($l=2$)时训练得到 $N_l=4^2=16$ 个视觉词典,这两种空间分层训练得到的视觉词典对图像类别的区分能力都较强,在每类训练图像为 30 时分别获得了 75.85% 和 77.01% 的准确率.本文 FFSVD 算法综合第 0 ~ 2 层的信息,比第 0 层提高了约 4.0%,这是因为:

(1)空间金字塔能有效保存图像的空间信息,第 1 和第 2 层都考虑了视觉词汇在空间的分布特点,充分利用了空间金字塔划分后各子区域语义构成的上下文信息.

(2)FFSVD 算法集成多个空间分层的方法尽可能利用了图像不同层次的空间信息,来构造不同的空间视觉词典,能够弥补单个空间分层信息的不足,得到比仅使用单一层次更好的性能.

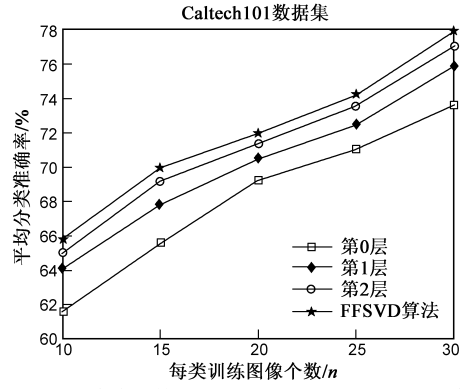

图5 FFSVD算法与单层视觉词典在Caltech101数据集上的分类结果对比

为了分析空间层次 L 更大时对 FFSVD 算法的性能影响,在 Caltech101 数据集上将 FFSVD 算法的 L 设置为 3 进行实验,得到的空间视觉词典集 $\varphi(D)$(式(5))大小为 85,在每类训练图像为 30 时获得的准确率为 78.33%.虽然准确率略高于综合 0 ~ 2 层的准确率 77.94%,但是综合考虑到性能与效率因素,从本小节实验的结果可知 FFSVD 算法中 L 设置成 2 效果比较好.

3.3.5 图像表示阶段的权值的大小对 FFSVD 算法分类性能的影响

本实验分析了 FFSVD 算法图像表示阶段,特征向

量加权连接时权值的大小对分类性能的影响,分析了在三种权值设置情况下 FFSVD 算法的分类性能:相同权值大小,按面积比例设置权值大小,随机设置权值大小.

(1)相同权值

权值大小设置为 $W=[1,1,1]$,即空间金字塔三个层次中的特征向量都赋予相同的权值.

(2)按面积比例设置权值

由 2.2 小节可知,本文空间金字塔划分方式是均匀划分,每个层次中的块都是一样大小,且第 l 层中块的数目为 $N_l=4^l(l=0,1,\cdots,L)$.本实验按照每个层次中块数目的比例来选择权值.第一,选择与各层次中块数目成反比的一组权值,也就是与各区域所占面积比例成正比,即 $W=1/N_l=[1,1/4,1/16]$.第二,选择与各层次中区域所占面积成反比的一组权值即 $W=[1/16,1/4,1]$.

(3)随机设置权值

随机选择一些任意大小的权值来进行实验分析,首先,选择了与文献[3]中核矩阵系数一致的权值即 $W=[1/2,1/4,1/4]$.然后,随机选择一组与金字塔层次中块数目无关的权值即 $W=[1/8,1,1/2]$.

各权值设置情况下,在 Caltech101 数据集上的 FFSVD 算法分类性能比较如图 6 所示.从图 6 中可以看出,当对特征向量直接串接,即 $W=[1,1,1]$时得到的准确率是比较低的.在空间三个层次中,赋予了第 $l=0$ 层权值比其他两层权值更大时,获得的准确率都比较高,如 $W=[1/2,1/4,1/4]$]和本文的 $W=[1,1/4,1/16]$.由此可知,本文根据空间层次中区域面积来设置权值的方法有效地利用了空间信息,获得了相对较高的准确率.

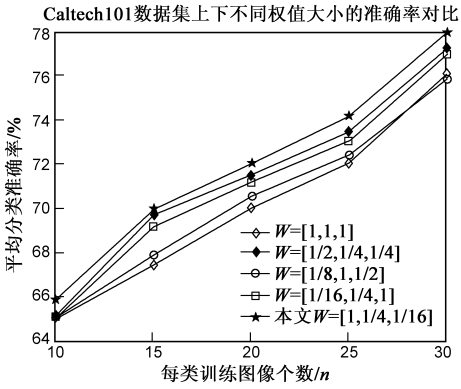

图6　Caltech01数据集上不同权值大小对分类性能的影响

3.3.6　空间视觉词典大小对 FFSVD 算法分类性能的影响

本实验分析了空间视觉词典大小对 FFSVD 算法分类性能的影响.首先,为了分析不同空间金字塔划分层次下,各区域都采用相同大小的空间视觉词典时对 FFSVD 算法分类性能的影响,分别采用了 4 种不同大小的视觉词典,即 500、1000、1500 和 2000,在数据集 Caltech101 和 15 Scenes 上进行实验,它们分类性能的比较情况如图 7 所示.

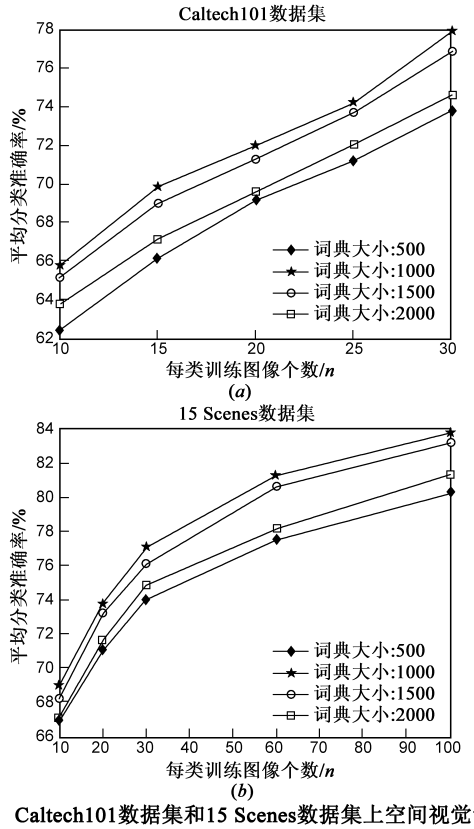

图7　Caltech101数据集和15 Scenes数据集上空间视觉词典大小对FFSVD算法分类性能的影响

由图 7(a)和(b)都可以看出,空间视觉词典大小为 1000 时在两个数据集上都取得了最好的分类效果.并且,观察图 7(a)和(b)可知,词典大小为 1000 与 1500 时的结果很接近,在词典大小为 500 和 2000 时的结果都不理想,在词典大小为 500 时最低.这是因为,视觉词典过小时(如 500),不同语义概念的图像特征可能被标记为相似的视觉单词,导致生成的视觉词典分类性能降低.随着视觉词典大小增加(如 1000、1500),分类性能也在稳步提高,但词汇本大小达到一定程度时(如 2000),反而会使得相同的特征被表示为多个不同的视觉单词,造成分类性能降低.

然后,考虑到不同空间金字塔划分层次下得到的子区域大小不同,实验根据层次来分配空间视觉词典的大小,$l=0,1,2$ 层次下的图像子区域中的词典大小分别设置为 1000、500、250,与各区域都采用相同的空间视觉词典大小 1000 时,在 Caltech101 数据集上分类性能比较如图 8 所示.

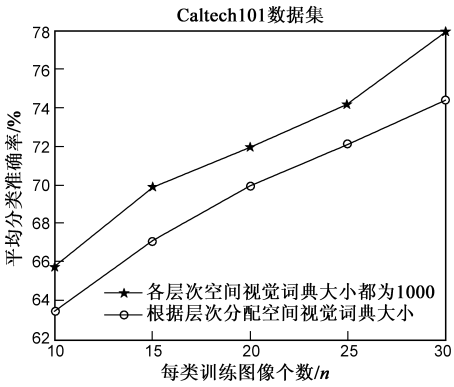

图8 不同分配方式下的空间视觉词典对FFSVD算法分类性能的影响

由图 8 可知,各层次空间视觉词典大小都为 1000 时,比根据层次来分配空间视觉词典的大小时得到的准确率更高.这可能是因为第 1 层和第 2 层的视觉词典过小,使得它们的表征区分能力降低.

综合考虑到分类效率与性能因素,从本小节的实验结果可知,FFSVD 算法各层次下的空间视觉词典大小都设为 1000 至 1500 左右可能效果比较好.

4 结论

本文提出了一种基于特征融合与空间视觉词典的图像分类方法即 FFSVD 算法.FFSVD 算法将 SIFT 特征与 KDES-G 特征进行串行融合,生成两组特征的联合向量 GKSF 作为特征向量.并进一步引入图像空间信息及融合不同粒度信息,提出了一种多级空间视觉词典集体的构造方法,先对图像进行空间金字塔划分,再对空间相对应的子区域内的特征分别聚类,构建不同层级,不同子空间区域的空间视觉词典.在图像表示阶段,分别对图像不同层上各子区域的特征基于其对应的空间词典进行 LLC 编码,将编码后各子区域的特征向量赋予不同的权重加权处理,连接形成最终的图像描述.FFSVD 算法在两个常用数据集上都得到了很好的分类效果,实验结果表明 FFSVD 算法能有效融合多个特征与空间信息,而且算法简单,运行速度快,各子区域上的视觉字典的构造和各子区域的量化都可以并行进行.

参考文献

[1] Gabriella C, Christopher R Dance, et al. Visual categorization with bags of keypoints[A]. Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision[C]. Prague: Xerox, 2004. 1 – 22.

[2] 郭立君,赵杰煜,史忠植.生成模型与判别方法相融合的图像分类方法[J].电子学报,2010,38(5):1141 – 1145.
Guo Li-jun, Zhao Jie-yu, Shi Zhong-zhi. Image categorization of integrated generative models and discriminative methods[J]. Acta Electronica Sinica, 2010, 38(5): 1141 – 1145. (in Chinese)

[3] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [A]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2006. 2169 – 2178.

[4] 汪成亮,周佳,黄晟.基于高斯混合模型与 PCA-HOG 的快速运动人体检测[J].计算机应用研究,2012,29(6):2156 – 2160.
Wang Cheng-liang, Zhou Jia, Huang Sheng. Motion human detection based on mixture of Gaussians and PCA-HOG[J]. Application Research of Computers, 2012, 29(6): 2156 – 2160. (in Chinese)

[5] Dalal N, Bill Triggs. Histograms of oriented gradients for human detection[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2005. 886 – 893.

[6] 张静,董伟,李红娟.基于改进的图像兴趣点特征提取匹配研究[J].计算机仿真,2012,29(2):288 – 291.
Zhang Jing, Dong Wei, Li Hong-juan. Points of interest based on improved image feature extraction algorithm[J]. Computer Simulation, 2012, 29(2): 288 – 291. (in Chinese)

[7] Harzallah H, Jurie F, Schmid C. Combining efficient object localization and image classification[A]. Proceedings of IEEE 12th International Conference on Computer Vision[C]. USA: IEEE, 2009. 237 – 244.

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

[9] 高常鑫,桑农.整合局部特征和滤波器特征的空间金字塔匹配模型[J].电子学报,2011,39(9):2034 – 2038.
Gao Chang-xin, Sang Nong. Unifying local features and filterbank features in the spatial pyramid matching model[J]. Acta Electronica Sinica, 2011, 39(9): 2034 – 2038. (in Chinese)

[10] 程刚,王春恒.基于结构和纹理特征融合的场景图像分类[J].计算机工程,2011,37(5):227 – 229.
Cheng Gang, Wang Chun-heng. Scene image categorization based on structure and texture feature fusion[J]. Computer Engineering, 2011, 37(5): 227 – 229. (in Chinese)

[11] 赵永威,李弼程,彭天强.一种基于随机化视觉词典组和查询扩展的目标检索方法[J].电子与信息学报,2012,34(5):1154 – 1160.
Zhao Yong-wei, Li Bi-cheng, Peng Tian-qiang. An object retrieval method based on randomized visual dictionaries and query expansion[J]. Journal of Electronic & Information Technology, 2012, 34(5): 1154 – 1160. (in Chinese)

[12] 王宇新,郭禾,何昌钦.用于图像场景分类的空间视觉词

袋模型[J].计算机科学,2011,38(8):265 – 268.
Wang Yu-xin, Guo He, He Chang-qin. Bag of spatial visual words model for scene classification[J]. Computer Science, 2011,38(8):265 – 268.(in Chinese)
[13] 刘硕研,须德,冯松鹤.一种基于上下文语义信息的图像块视觉单词生成算法[J].电子学报,2010,38(5):1156 – 1161.
Liu Shuo-yan, Xu De, Feng Song-he. A novel visual words definition algorithm of image patch based on contextual semantic infornlation[J]. Acta Electronica Sinica, 2010,38(5): 1156 – 1161.(in Chinese)
[14] Wang J J, Yang J C, Yu K. Locality-constrained linear coding for image classification[A]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2010. 3360 – 3367.
[15] 亓晓振,王庆.一种基于稀疏编码的多核学习图像分类方法[J].电子学报,2012,40(4):773 – 779.
Qi Xiao-Zhen, Wang Qing. An image classification approach based on sparse coding and multiple kernel learning[J]. Acta Electronica Sinica, 2012,40(4):773 – 779.(in Chinese)
[16] Bo L F, Ren X F, Fox D. Kernel descriptors for visual recognition[A]. Proceedings of Advances in Neural Information Processing Systems[C]. USA: NIPS, 2010. 244 – 252.
[17] 王欢,王江涛,任明武,等.一种鲁棒的多特征融合目标跟踪新算法[J].中国图象图形学报,2009,14(3):489 – 498.
Wang Huan, Wang Jiang-tao, Ren Min-wu, et al. A new robust object tracking algorithm by fusing multi-features[J]. Journal of Image and Graphics, 2009, 14(3): 489 – 498.(in Chinese)
[18] 韩智,刘昌平.基于多种特征融合的指纹识别方法[J].计算机科学,2010,37(7):255 – 259.
Han Zhi, Liu Chang-ping. Fingerprint recognition method based on multi-feature fusion[J]. Computer Science, 2010, 37(7):255-259.(in Chinese)
[19] Li Fei-Fei, et al. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 2004, 106(1):59 – 70.
[20] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):1 – 25.
[21] BO L F, Sminchisescu C. Efficient match kernel between sets of features for visual recognition[A]. Proceedings of Advances in Neural Information Processing Systems[C]. USA: NIPS, 2009. 135 – 143.
[22] Yang J, Yu K, Gong Y H. Linear spatial pyramid matching using sparse coding for image classification[A]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2009. 1794 – 1801.
[23] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope[J]. International Journal of Computer Vision, 2001, 42(3): 145 – 175.

作者简介

罗会兰　女,1974 年 9 月生,江西省上高县人.2008 年在浙江大学获工学博士学位,现为江西理工大学图像处理实验室教授、硕士生导师,主要从事机器学习、模式识别等方面的研究.
E-mail: luohuilan@sina.com

郭敏杰　女,1989 年 7 月生,江西省宜丰县人.2007 年进入江西理工大学,现为江西理工大学硕士研究生,主要从事模式识别及图像分类与识别方面的研究.
E-mail: jaymi890711@sina.cn