一种多层次抽象语义决策图像分类方法

刘 鹏 [1] 叶志鹏 [1] 赵 巍 [1] 唐降龙 [1]

摘 要 视觉词包 (Bag-of-visual-words, BoVW) 模型是一种有效的图像分类方法. 本文提出一种基于语义抽象的多层次决策 (Multiple layer decision, MLD) 方法, 通过在 BoVW 中引入抽象语义进行多层次扩展, 采用语义保留方法生成具有语义的视觉词典, 利用自底向上的方式逐层传递语义, 训练上层语义分类器; 分类时采用自顶向下方式逐层判断待测样本的类别. 用标准数据集验证方法的分类性能. 结果表明, 本文提出的方法与主流分类方法相比具有更好的分类性能.

关键词 图像分类, 图像模糊分类, 视觉词包模型, 决策树, 多层次决策

引用格式 刘鹏, 叶志鹏, 赵巍, 唐降龙. 一种多层次抽象语义决策图像分类方法. 自动化学报, 2015, **41**(5): 960−969

DOI 10.16383/j.aas.2015.c140238

A Multiple Layer Abstract Semantic Decision Method for Image Classification

LIU Peng[1] YE Zhi-Peng[1] ZHAO Wei[1] TANG Xiang-Long[1]

Abstract Bag-of-visual-words (BoVW) is an effective method in image categorizing and retrieving task. A multiple layer decision method (MLD), which introduces abstract semantics of image categories into BoVW to carry out middle-level and upper-level extensions, is proposed in this paper. Semantics is preserved at the stage of generating visual vocabulary, based on which classifiers are trained in a bottom-up way. Abstract semantics is transferred during the training step. After that, the category of a test image is estimated gradually by classifier through each layer in a top-down way. Experiments on standard datasets show that the proposed method achieves better performance compared with mainstream classification methods.

Key words Image classification, image fuzzy classification, bag-of-visual-words (BoVW), decision tree, multiple layer decision (MLD)

Citation Liu Peng, Ye Zhi-Peng, Zhao Wei, Tang Xiang-Long. A multiple layer abstract semantic decision method for image classification. *Acta Automatica Sinica*, 2015, **41**(5): 960−969

近年来, 计算技术及图像传感器的快速发展, 极大地方便了日常生活中图像的获取与分享. 流行的图像分享网站 Flickr 存储的图像数量已经超过 60 亿; 知名图像社交网站 Instagram 的活跃用户数量突破了一亿. 一方面, 丰富的图像数据可为用户提供更优质的信息资源; 另一方面, 海量的图像数据使手工类别标注几乎成为不可能完成的任务. 因此需要对图像类别进行自动标注并分类存储以提高用户检索效率.

视觉词包 (Bag-of-visual-words, BoVW) 模型由 Csurka 等于 2004 年提出[1], 是一种基于频率统计的图像分类方法. 该方法借鉴了文本分类中词包 (Bag-of-words, BoW) 模型的思想, 由 4 个部分组成: 1) 特征提取, 通常使用尺度不变特征变换算法 (Scale invariant feature transform, SIFT) 描述符对特征点进行描述; 2) 构建视觉词典, 利用 K-means 聚类方法将特征点聚成数类, 视觉词典由聚类中心形成的视觉词汇组成; 3) 利用特征局部投影生成码书 (Codebook), 即构造 Bag-of-features (BoF) 特征; 4) 分类器训练. 利用 BoF 特征训练分类器, 获取分类模型, 对待分类图像特征进行预测. BoVW 及其改进模型由于其易实现且对遮挡、光照改变等具有鲁棒性, 因而成为视觉物体的一种主要表示方式[2], 广泛应用于图像中目标与场景分类[3−4]、视频场景事件检测[5] 与行为识别[6] 等领域.

然而, 视觉词包模型也存在局限性: 1) 仅考虑视觉单词的频率分布而忽略了图像中普遍存在的空间相关性[7]; 2) K-means 采用的 Hard-assignment 未必会找到最优的词典, 针对视觉特征的聚类导致产生同义的视觉单词使视觉模式发生过表示, 在视觉词典中引入同义性和不确定性[8]; 3) 采用聚类方法生成单词会导致语义丢失, 而语义关系对于场景理解与目标识别具有重要意义.

针对上述词包模型的不足,研究人员针对词包模

收稿日期 2014-04-23 录用日期 2015-01-06
Manuscript received April 23, 2014; accepted January 6, 2015
国家自然科学基金 (61171184, 61201309, 61440025) 资助
Supported by National Natural Science Foundation of China (61171184, 61201309, 61440025)
本文责任编委 贾云得
Recommended by Associate Editor JIA Yun-De
1. 哈尔滨工业大学计算机科学与技术学院模式识别与智能系统研究中心 哈尔滨 150001
1. Pattern Recognition and Intelligent System Research Center, School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

型进行了优化, 主要包括: 1) 基于分割的优化, 这种优化方法认为图像分类问题中, 前景为待识别目标, 背景为与目标类别无关的噪声. 通过图像分割提供目标的感兴趣区域 (Region of interest, ROI), 从而排除背景噪声, 提高词典生成质量. Du 等提出在词典生成阶段使用有监督 Mean-shift, 仅将目标区域作为 ROI, 去除了无关背景对词典的影响[9]; Chai 等针对弱标记数据集提出三级 (类型级、图像级、数据级) 类型判别一致分割方法用于自动分割出类别区分能力最佳的前景区域进行分类器训练, 提高了细粒度图像分类中前景目标的可分性[10]. 2) 编码优化. Krapac 等提出基于高斯混合模型 (Gaussian mixture model, GMM) 的图像特征空间分布编码方法以提高视觉单词的信息量和表达能力[11]; Bolovinou 等提出一种空间视觉单词包 (Bag of spatio-visual words) 的场景分类方法, 通过在视觉单词中加入上下文信息编码, 利用 Spherical K-means 算法生成空间视觉单词[12]; Wang 等提出一种局部约束线性编码 (Locality-constrained linear coding, LLC) 模式用于取代传统的矢量量化编码, 不仅提高了线性支持向量机 (Support vector machine, SVM) 的分类效果, 同时可以显著降低算法复杂度[13]. 3) 基于二义性的优化. 传统 BoVW 方法只用一个视觉单词描述一个图像特征, 视觉单词的二义性是将连续的图像特征映射到离散的视觉单词上, 即利用两个或多个视觉单词描述同一图像特征. 在视觉单词中引入二义性可以增加模型的表述能力, 从而提升模型的分类效果[14]; Liu 等提出了基于组件 (Components) 的图像表示方法, 每个组件结合了附近视觉单词的空间相关性, 使用视觉单词的频率分布作为其特征表示, 比单个视觉单词具有更好的描述能力[15]; Avrithis 等针对大规模图像分类问题, 结合高斯混合模型构造视觉单词提出一种改进的期望最大化算法, 该方法采用近似最近邻提高搜索效率, 并利用其 EM (Expectation maximization) 算法的迭代性进行增量搜索, 从而提高分类精度[16]; Mikulík 等针对大规模图像分类的 BoVW 类方法提出一种无监督相似度测量方法, 与 L2 软分配 (Soft assignment) 及汉明嵌入 (Hamming embedding) 相比具有更好的区分能力[17]. 4) 语义与词典压缩优化, 低层特征与高层语义间存在语义鸿沟. 语义鸿沟使信息间缺乏一致性, 即从图像中提取的视觉特征与该特征所反映的目标类别不一致[18]. 为了克服语义鸿沟, 研究人员提出了对场景语义建模的图像描述方法; 词典大小是影响 BoVW 模型的重要因素之一, 因此在分类性能损失可以接受的范围内获取更小、更紧凑的词典, 可以提高 BoVW 模型的分类效率. Wu 等引入马氏距离作为语义鸿沟的度量标准并通过距离度量学习缩小语义鸿沟以便尽可能地降低语义损失[19]; Ji 等提出基于条件随机场 (Conditional random field, CRF) 的统一模型, 用于缩小语义上下文与视觉特征间的语义鸿沟[20]; Liu 等提出由中级特征学习语义视觉词典, 每个中级特征由驻点共同信息 (Pointwise mutual information, PMI) 矢量表示, 通过计算传播距离作为不相似度获取特征间的关系, 最终获取压缩词典[21]; Penatti 等提出视觉单词空间分布 (Word spatial arrangement, WSA) 方法用于表示 BoVW 模型中视觉单词的空间分布信息. 该方法将图像以每个特征点为中心划分为 4 个象限, 统计每个象限内的其余特征点个数作为空间分布编码. 与空间金字塔相比, WSA 提供了足够的空间信息, 同时使特征矢量更紧凑[22].

以往的研究多侧重于利用图像中的低层和高层特征, 引入度量或学习手段对视觉词包模型进行改进, 没有解决 BoVW 分类效果受样本数量的影响, 对未经训练类别的样本分类效果欠佳的问题. 因此, 有必要在初级视觉词典的基础上构建更高层级的视觉词典, 以提升分类效果. 目前, 有别于利用传统的低层特征进行分类的方法, 部分学者利用图像间一般到特殊的关系, 针对图像语义层次的构造展开了研究. Li 等从构造具有语义意义的图像层次以减轻组织样本工作量的角度, 提出一种图像层次自动发现模型用于指导样本分类[23]; Bannour 等提出了“语义–视觉概念关系” (Semantico-visual relatedness of concepts, SVRC) 方法, 用于度量图像中语义概念间的相似程度, 并利用构建的语义层次对图像进行分类, 其基本思想是针对图像的视觉、概念及语义相似性度量, 生成层次构造规则, 从而达到自动生成图像层次的目的[24]; Bannour 等[25] 针对层次图像分类问题, 提出了一种图像数据集层次结构学习方法 One-versus-opposite-nodes. 该方法通过将分类问题分解为数个子问题, 从而对大规模数据集具有较好的性能. 本文从语义抽象的角度出发, 针对未经训练样本的分类问题提出一种多层次决策模型 (Multi-level decision model, MLD), 通过多层次语义决策逐步缩小搜索范围, 给出与未经训练类别的样本相近的分类结果, 从而有效提高针对未经训练类别的样本的分类效果. 本文以等概率从每类视觉语义属性集中选取视觉单词, 使每类视觉语义都可以被选择到, 从而避免了出现频率低的语义无法参与上层语义的构造而导致的分类性能下降; 另外, 在上层语义和具体语义之间增加了中间语义层, 进一步缩小了语义鸿沟.

1 多层次决策图像分类方法

图像分类问题可表示为一个最小化问题, 问题

的本质即通过特征提取与量化手段, 在已训练样本特征集合内搜索与待分类样本距离之和最短的类别作为分类结果, 可用式 (1) 描述, 其中 F_t 为待分类图像的特征, F_i 为第 i 个类别图像的特征集合, d 为度量函数. BoVW 即属于上述分类模型的一种. BoVW 图像分类框架仅有一层类别数据集作为语义类别, 其优点是简单快速易实现, 但也具有明显的不足, 即分类能力有限, 对图像的分类结果依赖于具体类别的种类与数量, 使其在小规模数据集上获取的经验未必在大规模数据集成立[26]. 若输入待分类图像的类别并未出现于训练数据集, 则产生的分类结果往往不尽人意. 本文通过引入抽象语义作为分类器的视觉先验知识, 构造多层决策模型来预测样本的类别.

$$C = \arg\min_i d(F_t, F_i) \tag{1}$$

1.1 多层次决策模型

人们利用已有视觉知识学习新的目标类别. 本文通过在传统 BoVW 中结合抽象方法增加公共类别语义层次作为视觉知识, 提高模型的描述能力. 抽象技术最早由面向对象程序设计方法提出, 是一种对现实世界理解和抽象的设计理念, 目的是提高程序的灵活性和可维护性, 目前已在人工智能领域得到广泛应用[27]. 抽象语义类别 (Abstract semantic category, ASC) 定义了某种事物的特征模式, 通常包括事物的属性 (如眼睛、毛发) 及其行为 (如跑、叫等). 与具体类别中的单一图像相比, 抽象模式具有更强的描述能力. 抽象类别与底层具体类别判决分类器距离越远, 其抽象程度越高. 本文将数据集中的图像类别称为具体类别, 利用抽象技术获取的上层类别称为抽象语义类别, 其语义由来自不同下级类别的视觉语义属性组成. 将具有相同基本属性的具体类别用一个 ASC 描述, 如上层 ASC "动物" 是 "猫"、"狗"、"人" 等中层 ASC 的集合, 其相似之处是都具有头、眼睛、腿等特征.

本文提出的多层次决策方法 MLD 的抽象体系及其层级如图 1 所示. 同一个抽象类中的不同具体类别之间既有区别又相互联系, 不同抽象类间的具体类别的属性则差异较大. 通过在具体类别中提取共有的上层抽象模式有助于产生类间距离大而类内距离小的码书, 从而提升分类效果. BoVW 算法及其改进方法仅包含具体层一个层次, 本文加入了上层与中层 ASC 组成的两层抽象层级, 分别记为 UASC 和 MASC. 若不使用抽象层, 直接使用具体层 BoF 描述图像数据集, 则 MLD 退化为单层次 BoVW 算法, 故 MLD 是 BoVW 算法的超集.

本文中的度量模型的形式为

$$\begin{aligned} u &= \arg\min_i d(F_t, F_i) \rightarrow \\ m &= \arg\min_j d(F_t, F_j^u) \rightarrow \\ c &= \arg\min_k d(F_t, F_k^m) \end{aligned} \tag{2}$$

其中, F_i、F_j^u 与 F_k^m 分别为第 i、j 与 k 个 UASC、MASC 属性和下层视觉特征集合. 式 (1) 和 (2) 将图像分类问题划分为 3 个决策步骤, 将待分类图像 I 输入 U-SVM (上层分类器) 得到该图像所属的上层抽象类别, 接下来将 I 输入 u 所包含的 M-SVM (中层分类器), 获得该图像所属的中层抽象类别, 再次缩小具体类别的搜索范围, 最后利用 BoVW 方法根据分类器输出将 I 标注为具体类别 c, 完成整个分类过程. 由于每进行一步搜索均需根据上一步分类器输出决定下一步搜索路径, 因此本文模型需对抽象层中每个类别训练一个分类器, 但这个代价仅在训练时存在.

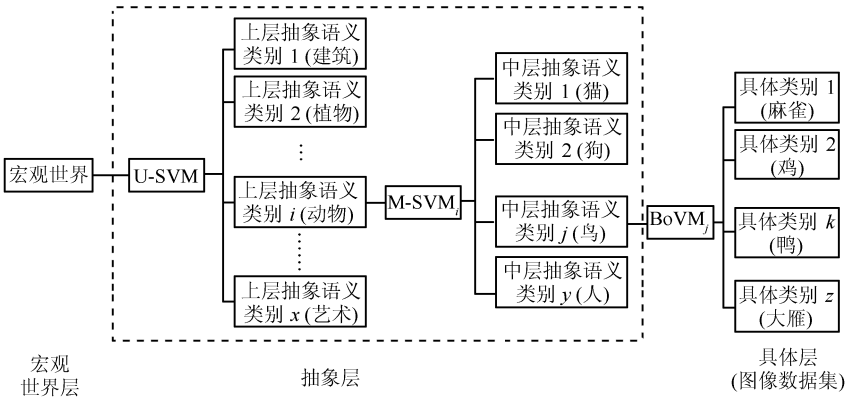


图 1 MLD 模型层次结构

Fig. 1 Hierarchical structure of MLD model

1.2 视觉语义学习与预测

视觉语义属性可用于目标识别及图像分类. 本文使用视觉语义训练抽象层分类器, 使用视觉单词训练具体层分类器, 即抽象层的输入为保留了视觉语义的视觉单词, 称为视觉属性, 具体层的输入为图像样本. 本文采用语义保留视觉词包 (Semantics-preserving bag-of-words, SPBoW)[27] 在生成视觉词典的同时保留其中的语义信息, 其本质是将语义识别特征映射到视觉单词中, 通过语义识别特征间的距离作为语义鸿沟的度量手段, 在最小化语义鸿沟的同时学习最优码书, 以最大程度地保留视觉单词的语义信息. 视觉属性的学习是视觉语义分类器不断积累与更新视觉知识的过程. 分类器的训练采用自底向上的方式, 算法 1 描述了上述策略, 其中变量 i, j, k, x, y, z 与图 1 对应, UASC_i 表示第 i 个上层抽象语义类, MASC_j 表示第 j 个中层抽象语义类.

算法 1 (MLD 学习算法).

样本准备阶段:

步骤 1. 利用 SPBoW 产生具体类 k 的视觉单词 v_q 及其语义 s_q 的集合 $Inh_k^j = \{(v_q, s_q)\}_{q=1}^{c}$, c 为码书大小; 令 MASC_j 的视觉单词集合 $M_j = \bigcup_{k=1}^{z} Inh_k^j$. 对 UASC_i 包含的所有 MASC 迭代该过程, $M\_A_i = \bigcup_{j=1}^{y} M_j$;

步骤 2. 从 UASC_i 包含的每个 Inh_k^j 以等概率随机选择语义视觉单词构造集合 $U\_ABS_i = \bigcup_{j=1}^{y}\bigcup_{k=1}^{z} Inh_k^j$;

步骤 3. 对每个 UASC 重复上述步骤, 直至所有 UASC 均被访问, $U\_A = \bigcup_{i=1}^{x} U\_ABS_i$;

步骤 4. 量化上述 U_A 及每个 M_A_i 与 Inh_k^j, 得到相应视觉词包特征 BoF_{up}、BoF_{mid} 与 BoF_{inh}.

训练阶段:

步骤 5. 分别利用 BoF_{mid} 训练 MSVM_i, BoF_{inh} 训练 BoVW_j. 重复该过程直至所有中层与底层分类器训练完毕;

步骤 6. 利用 BoF_{up} 训练 U_SVM.

随着抽象层次的上升, 每个具体类别的视觉属性在抽象属性集合中所占的比例随着抽象程度的升高逐渐降低. 例如, MASC "鸟" 的视觉属性由于其抽象程度的提升, 加入了其他具体类别的视觉属性, 在可以表述更大范围的目标类别的同时, 来自于具体类别 "鸡" 的视觉属性的比例必然下降. 本文为了避免某类样本出现频率较高, 导致分类器只对该部分样本有效, 在学习视觉属性分类器时, 使各具体类别的视觉属性比例均衡, 即使用等概率不放回方式选取语义视觉单词, 这也是本文与其他语义层次结构方法的不同之处.

测试时采用自顶向下方式. 通过方法自身与样本选择两个角度降低后续分类对上层决策的敏感性: 1) 在判断一个样本的类别时, 提取其 Bag-of-features 并记录其通过 U-SVM 及每个 M-SVM 的输出 u', m', 选择其和最大的中层抽象类别作为下一步分类的起始类别; 最后将图像对应的 Bag-of-features 输入每一个具体层分类器, 计算分在该类的可能性 $p_A^1, p_A^2, \cdots, p_A^n$, n 为具体类别数. 将最大输出值所对应的类别作为测试图像的类别, 即:

$$C = \arg\max_{t=1}^{n} p_A^t \tag{3}$$

2) 利用 SPBoW 方法生成最优的语义视觉单词, 提高分类性能.

1.3 图像分类

本文考虑精确图像分类和模糊图像分类两种情形:

1) 精确图像分类. 对实际类别为 C_i 的待分类图像, 训练阶段使用的图像具体类别集合 T, 若 $C_i \in T$, 这种情形属于精确图像分类, 即图像中目标类别对分类器是已知的. 这种情形的分类结果正确与否较容易判断, 仅需按式 (4) 比较模型对图像类别的预测输出与实际图像类别是否一致即可, 即

$$\text{correct\_rate} = \frac{\sum_j \delta(C_j - C_i)}{N} \times 100\,\% \tag{4}$$

其中, $\delta(C_j - C_i) = \begin{cases} 1, & C_j = C_i \\ 0, & \text{其他} \end{cases}$; N 为 C_i 的测试样本数.

2) 模糊图像分类. 很多情形下, 图像中目标所属的具体类别对于分类器是未知的, 用户希望通过输入图像寻找与图像中目标相似的类别. 本文假设用户进行模糊分类时, 期望反馈包含与被分类目标具有相似类别的图像, 即对每个待测样本 T, 分类算法的正确率如式 (5) 所示:

$$\text{correct\_rate} = \frac{\sum_j \delta(U_j - U_i)}{N} \times 100\,\% \tag{5}$$

其中, $\delta(U_j - U_i) = \begin{cases} 1, & U_j = U_i \\ 0, & \text{其他} \end{cases}$; U_i 为测试图像的实际上层抽象类别, U_j 为模型预测的上层抽象类别. MLD 与 BoVW 的模糊分类的过程如图 2 所示, 图 2 中未列出所有的图像类别与视觉单词. 可见模糊分类是一种 "尽力而为" 的分类过程, 其并不要求计算出完全精确的图像类别, 仅需计算出被分类图像的上层抽象类别即可.

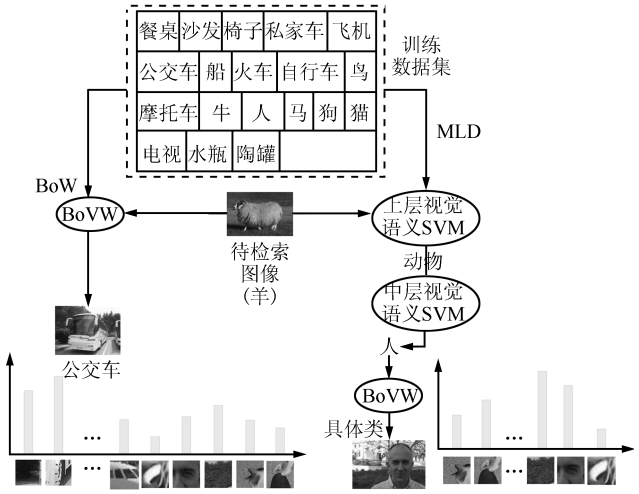


图 2 PASCAL VOC 2007 数据集 BoVW 和 MLD 图像模糊分类比较

Fig. 2 Classification results of BoVW and MLD on PASCAL VOC 2007 dataset

本文提出的多层次 MLD 与单层次 BoVW 相比具有以下优势:

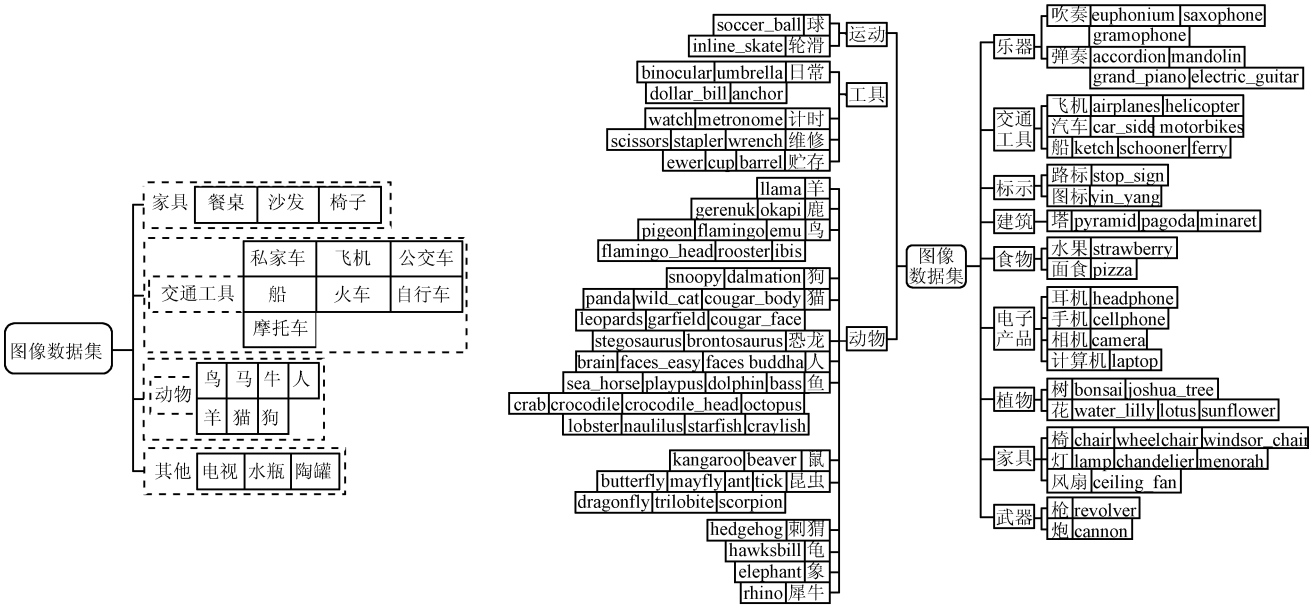
1) BoVW 尝试直接利用数据集大小和种类有限的具体类别描述类别数量近乎无穷的宏观世界层目标, 其精度必然会受到类别数据集大小的约束, 这是其方法的不足之处. 本文通过引入抽象技术, 在抽象层中屏蔽了不同具体类别数据的细节而保留了共同属性, 并且抽象层次可根据需要任意扩展与压缩, 拥有与宏观世界类别近乎等价的描述能力.

2) BoVW 模型通常对于模糊分类效果不佳, 而这种分类方式在客观世界中存在且经常发生. 这是由于分类模型不可能穷尽所有类别的输入样本进行训练. MLD 模型则可以通过上层与中层抽象类别判断给出抽象模式相同且具体类别相近的结果, 对于图像模糊分类中经常遇到的未知类别样本的预测更为有效, 分类结果更加符合抽象层语义的一致性, 分类结果更为合理.

2 实验结果

本文采用了目前广泛应用于图像分类领域的主流数据集, 包括 PASCAL VOC 2007[28], Caltech-101[29]. 前者具有完整详细的标注信息, 包含 20 个类别共 9 963 幅图像. 后者包含 101 个类别共 9 146 幅图像, 目标大多居中并经过处理旋转到一个适于分类的位置; 实验所采用的数据集层次结构如图 3 (a) 和 (b) 所示. 采用线性核 LIBLINEAR SVM[30] 进行分类, 利用 One-vs-all 策略训练分类器, 码书大小为 1 000.

与 Caltech-101 数据集不同, PASCAL VOC 2007 并未提供图像的整体类别, 仅给出每幅图像中物体类别的标注信息, 给抽象层次构建造成了一定的困难. 因此, 根据该数据集自身的特点, 依据文献 [25] 构建的语义层次模型, 本文给出了包含上层与中层抽象类别的改进结构, 每个中层抽象类下所有样本不再细分, 共同构成一个具体类别. 根据数据集的层次结构, 按照层次遍历顺序将各数据集参数设



(a) PASCAL VOC 2007 层次结构
(a) Semantical structure of PASCAL VOC 2007

(b) Caltech-101 层次结构
(b) Semantical structure of Caltech-101

图 3 各数据集抽象层次结构

Fig. 3 The structure of datasets

置如下: PASCAL VOC 2007 数据集: $x' = 4$, $y' = \{3,7,7,3\}$, 每类随机选择训练样本数 $z' = 30$, 数据集中其余图像作为测试样本; 所有方法均采用 SIFT 描述符; Caltech-101 数据集: $x' = 12$, $y' = \{2,4,14,2,3,2,1,2,4,2,3,2\}$, $z' = \{1,1,4,2,3,3,1,2,6,2,6,2,6,2,4,12,2,7,1,1,1,1,3,4,2,2,3,1,1,3,1,1,1,1,1,1,2,3,3,3,1,1,1\}$. 本文率先针对 Caltech-101 的层次结构进行了研究, 该层次结构的组织同时考虑样本真实类别间的概念相似性及视觉相似性. 每个类别随机选择 30 幅图像作为训练样本. 选择 PASCAL VOC 2007 而非最新的 PASCAL VOC 2012 是基于以下考虑: 1) 可以进行公正的对比. PASCAL VOC 2012 中, 由于提供了更为丰富的场景信息, 很多方法均结合各类技术, 如 ROI 及低层特征优化、描述符降维、换用其他分类器等, 得到的结果往往是多种因素融合的结果, 不能公正地反映在模型本身的对比结果[31]; 2) PASCAL VOC 2007 每个类别均有详细的类别标注信息并可以较为容易地抽象为多层次数据类别. 首先考察不同中间层数对 MLD 分类性能的影响. 在一台配置 2.13 GHz 四核 CPU 及 12 GB 内存的工作站上进行实验. 3 层 MLD 包含 UASC、MASC 及具体层各 1 层、4 层与 5 层 MLD 分别增加 1、2 层 MASC. 分别从 PASCAL VOC 2007 和 Caltech-101 的每个类别随机选取一半图像作为两数据集的子集进行测试, 所有图像的平均分类性能与所消耗时间分别记为 P_1、T_1 和 P_2、T_2. 其中, $P_1 = 0.58$, $T_1 = 0.283$ s; $P_2 = 0.741$, $T_2 = 0.336$ s. 实验结果列于表 1. 该实验结果表明, 单层及二层 MLD 方法虽然可以减少算法运行时间, 但代价是性能显著降低. 例如, 与三层 MLD 相比, 对两个数据集, 单层 MLD 消耗时间分别减少了 62.8 % 及 57.9 %, 算法性能分别下降了 70.6 % 及 67.4 %; 二层 MLD 消耗时间分别减少了 19 % 及 32.7 %, 算法性能分别下降了 38.7 % 及 37.2 %. 另外, 层次的数量对分类性能的提升是微弱的, 但增加层次数量会使时间消耗显著增加. 例如, 对于 PASCAL VOC 2007 数据集, 使用 5 个层次, 分类性能提高仅 1.3 %, 而时间消耗却增加了 190 %, 因此从分类性能和消耗时间两方面综合考虑, 本文采用三层结构.

接下来与其他图像分类方法进行对比, 包括传统 BoVW[1]、SPBoW[19]、文献 [32] (Gaussian expectation-maximization purity, GEMP)、文献 [33] (Semantic attribute) 及文献 [13] (LLC) 的方法. 其中, BoVW、SPBoW、GEMP 及 LLC 为单层语义模型, Semantic attribute 与 MLD 为公共上层语义模型, 通过提取视觉语义属性作为上层特征用于分类. 分别针对精确与模糊分类两部分进行测试, 均利用 SIFT 特征产生视觉单词. 根据文献 [13] 所述, 将 Caltech-101 数据集中每个样本类别分别划分为 $5, 10, \cdots, 30$, 其对应的测试样本数不超过 50, 采用均值平均精度 (Mean average precision, MAP) 作为评价标准[32]. PASCAL VOC 2007 数据集精确分类结果如图 4 (a) 与表 3 所示, Caltech-101 数据集的实验结果如图 4 (b) 与表 4 所示. 从中可以看出, MLD 与 Semantic attribute 同为多层次语义结构方法, 在大多数测试中分类效果优于 BoVW、SPBoW 等单层次模型, 说明引入多语义层次能够提高分类精度, 这与语义抽象构造多层次分类模型的目的及已有研究的结论是一致的. 其中, MLD 具有最高的均值平均精度. 这是由于本文引入了 UASC 与 MASC 层次结构, 缩小了视觉单词与图像间的语义鸿沟; 通过自动学习抽象层语义分类器, 避免了在预测未知样本类别时, 较不合理的语义重要性排列策略. 在更具有挑战性的 PASCAL VOC 2007 数据集中, 目标可能以任意尺度出现于图像中任何角落. MLD 与单层模型相比提升幅度较大. 本文方法与部分主流方法的对比结果如表 2 所示. 实验结果表明, 本文方法在两个数据集上的平均识别率达到了可比的结果并优于部分主流方法.

表 1　MLD 层数与分类性能及复杂度测试结果

Table 1　Experimental results on performance and complexity tests between MLD layers

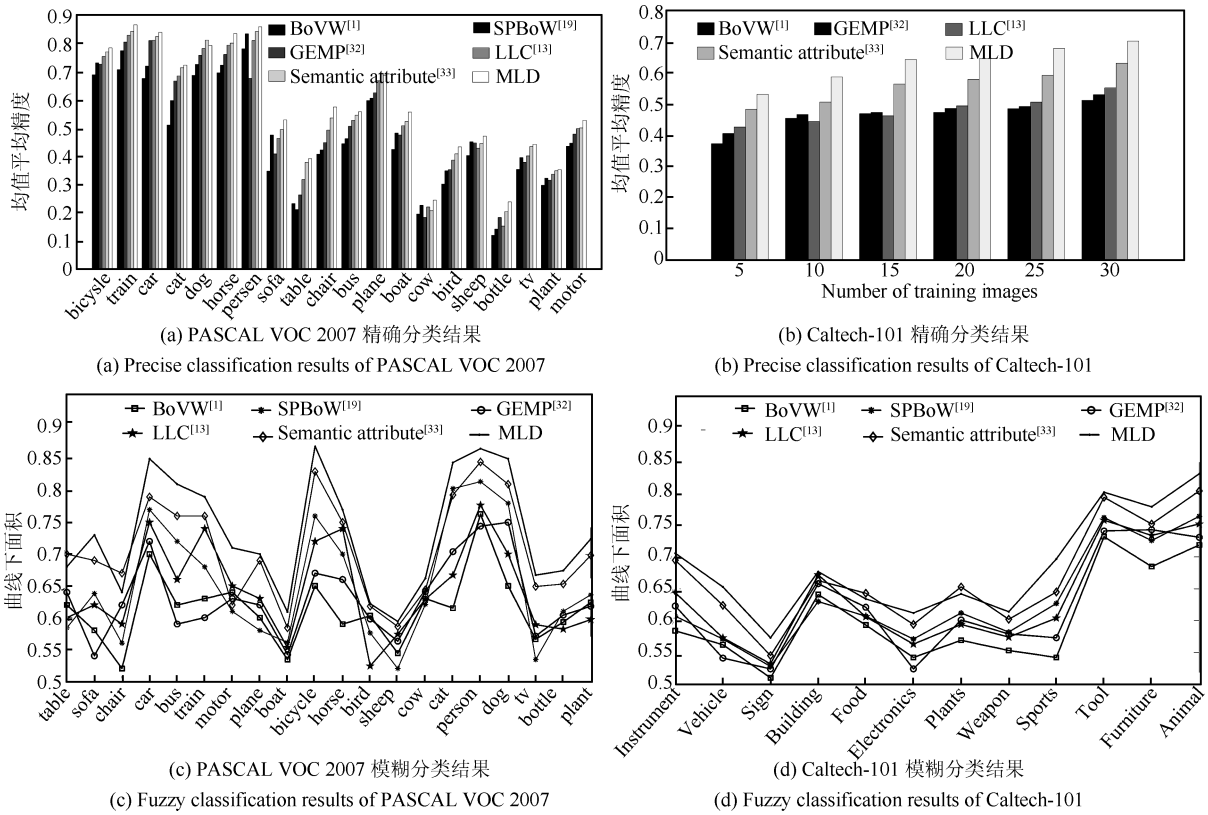
MLD	PASCAL VOC 2007		Caltech-101	
层数	分类性能	消耗时间	分类性能	消耗时间
单层	$0.294P_1$	$0.372T_1$	$0.326P_2$	$0.421T_2$
2 层	$0.613P_1$	$0.81T_1$	$0.628P_2$	$0.673T_2$
3 层	P_1	T_1	P_2	T_2
4 层	$1.015P_1$ (+1.5 %)	$1.6T_1$ (+60 %)	$1.003P_2$ (+0.3 %)	$2.35T_2$ (+135 %)
5 层	$1.013P_1$ (+1.3 %)	$2.9T_1$ (+190 %)	$1.007P_2$ (+0.7 %)	$3.6T_2$ (+260 %)

表 2　与部分主流方法的平均准确率对比结果

Table 2　Comparing results with some state-of-the-arts

	PASCAL VOC 2007	Caltech-101
Lazebnik[7]	–	0.646 ± 0.008
Perronnin[34]	0.583	–
Zhong[35]	–	0.69
Maji[36]	–	0.566 ± 0.008
Bilen[37]	0.571	0.753 ± 0.68
Proposed	0.586	0.709 ± 0.12

进行模糊分类的实验时, 上层抽象类 A 所包含



(a) PASCAL VOC 2007 精确分类结果

(a) Precise classification results of PASCAL VOC 2007

(b) Caltech-101 精确分类结果

(b) Precise classification results of Caltech-101

(c) PASCAL VOC 2007 模糊分类结果

(c) Fuzzy classification results of PASCAL VOC 2007

(d) Caltech-101 模糊分类结果

(d) Fuzzy classification results of Caltech-101

图 4 各数据集分类结果

Fig. 4 Classification results on each dataset

表 3 PASCAL VOC 2007 各类别实验结果

Table 3 Experimental results on PASCAL VOC 2007

类别	Bicycle	Train	Car	Cat	Dog	Horse	Person	Sofa	Table	Chair	Bus
BoVW[1]	0.689	0.707	0.676	0.511	0.687	0.696	0.781	0.348	0.233	0.408	0.445
SPBoW[19]	0.731	0.774	0.72	0.598	0.726	0.723	0.834	0.476	0.212	0.423	0.463
GEMP[32]	0.727	0.805	0.81	0.667	0.758	0.761	0.676	0.409	0.264	0.449	0.507
LLC[13]	0.754	0.829	0.811	0.685	0.782	0.794	0.81	0.464	0.318	0.493	0.528
Semantic attribute[33]	0.77	0.842	0.824	0.714	0.812	0.801	0.842	0.495	0.379	0.532	0.546
MLD	0.784	0.866	0.839	0.723	0.793	0.835	0.859	0.529	0.392	0.575	0.558
类别	Plane	Boat	Cow	Bird	Sheep	Bottle	TV	Plant	Motor	平均	
BoVW[1]	0.598	0.425	0.196	0.302	0.403	0.121	0.354	0.298	0.436	0.466	
SPBoW[19]	0.606	0.483	0.227	0.349	0.452	0.143	0.396	0.323	0.447	0.505	
GEMP[32]	0.625	0.476	0.184	0.353	0.448	0.184	0.379	0.316	0.479	0.514	
LLC[13]	0.668	0.509	0.221	0.387	0.429	0.153	0.402	0.337	0.498	0.544	
Semantic attribute[33]	0.694	0.524	0.208	0.409	0.446	0.205	0.435	0.349	0.501	0.566	
MLD	0.686	0.557	0.245	0.434	0.472	0.239	0.443	0.353	0.528	0.586	

的中层抽象类别记为 $X_i, i = 1, 2, \cdots, 48$, 具体抽象类记为 $y_j, y_j \in X_i$. 对 $\forall y_j$ 进行模糊分类时, 训练集由 $X_i \backslash y_j$ 及其他所有抽象类别随机选择的样本集 OC 组成, 测试集由 y_j 及其他所有抽象类别随机选择的样本组成 (与 OC 类别相同但具体测试样本不重复), 采用曲线下面积 (Area under curve, AUC) 对各方法性能进行评价[19]. 该值越大, 其对应方法的性能越好. 每个抽象类训练样本数为 30;

设图像数据集类别集合为 U, 当前测试上层抽象类 A, 分别从 A 与 $\forall X_i \in U\backslash A$ 中随机选择一定数量的图像, 分别记为 V、W, 共同组成测试样本集, $i = 1, 2, \cdots, Q$, Q 为数据集上层抽象类别数. 对于 Caltech-101 数据集, $V_c = \{10, 20, 30, 40, 50, 60\}$, $W_c = \{5, 10, 15, 20\}$; 对于 PASCAL VOC 2007 数据集, $V_p = \{5, 10, 15, 20, 25, 30\}$, $W_p = \{5, 10, 15\}$. 各参数的取值为 $\{(x, y)|x = 50, y = 5, x \in V_c, y \in W_c\}$ 与 $\{(x, y)|x = 25, y = 5, x \in V_p, y \in W_p\}$ 时取得最佳分类性能, 实验结果如图 4 (c) 和 (d) 所示. 受篇幅所限, 本文不能列出全部实验结果, 仅在表 5 中给出 Caltech-101 数据集 "动物" 抽象类的模糊分类结果. 从实验结果可知, 本文提出的 MLD 方法在大部分测试中优于已有方法. 这是由于 MLD 方法通过在公共上层语义搜索确定目标的抽象类别作为视觉先验知识为分类提供指导, 克服了 BoVW 方法

表 4　图 4 对应的 Caltech-101 实验结果

Table 4　Experimental results for Fig. 4

训练样本数	5	10	15	20	25	30
BoVW[1]	0.372	0.454	0.469	0.473	0.485	0.512
SPBoW[19]	0.396	0.472	0.481	0.495	0.514	0.537
GEMP[32]	0.405	0.466	0.473	0.486	0.492	0.53
LLC[13]	0.426	0.443	0.462	0.494	0.506	0.552
Semantic attribute[33]	0.483	0.506	0.564	0.579	0.592	0.631
MLD	0.531	0.587	0.633	0.646	0.677	0.709

表 5　Caltech-101 数据集 "动物" 抽象类模糊分类结果

Table 5　Fuzzy classification results of category "Animal" from Caltech-101 dataset

X_i	$A\backslash X_i$	AUC
羊	{鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.852
鹿	{羊, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.843
鸟	{羊, 鹿, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.845
狗	{羊, 鹿, 鸟, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.831
猫	{羊, 鹿, 鸟, 狗, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.825
恐龙	{羊, 鹿, 鸟, 狗, 猫, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.843
人	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 鱼, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.784
鱼	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鼠, 昆虫, 刺猬, 龟, 象, 犀牛}	0.864
鼠	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 昆虫, 刺猬, 龟, 象, 犀牛}	0.753
昆虫	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 刺猬, 龟, 象, 犀牛}	0.861
刺猬	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 龟, 象, 犀牛}	0.833
龟	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 象, 犀牛}	0.821
象	{羊, 鹿, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 犀牛}	0.887
犀牛	{羊, 鸟, 狗, 猫, 恐龙, 人, 鱼, 鼠, 昆虫, 刺猬, 龟, 象}	0.792
	平均	0.831

受限于具体且有限的训练数据集进行训练的缺陷和文献 [33] 缺乏中层语义的不足. 因此, 本文提出的 MLD 方法对于模糊分类取得了更好的分类性能.

3　结论

本文提出了多层次抽象类别图像分类方法. 通过两个抽象层次, 使图像分类性能得到显著提高. 该方法不但对精确图像分类具有较高的准确性, 且对于模糊图像分类也是有效的. 在多层次决策过程中, 上层抽象语义和中层抽象语义对未经训练的具体类别样本的分类发挥重要作用, 使分类的可靠性和准确性明显提高. 实验结果表明, 本文提出的方法对图像分类的性能优于主流分类方法.

References

1 Csurka G, Dance C R, Fan L X, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Proceedings of the 2004 Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision. Berlin, Germany: Springer Berlin Heidelberg, 2004. 1−2

2 Zhang Su-Lan, Guo Ping, Zhang Ji-Fu, Hu Li-Hua. Automatic semantic image annotation with granular analysis method. *Acta Automatica Sinica*, 2012, **38**(5): 688−697
(张素兰, 郭平, 张继福, 胡立华. 图像语义自动标注及其粒度分析方法. 自动化学报, 2012, **38**(5): 688−697)

3 Qin J Z, Yung N H C. Scene categorization via contextual visual words. *Pattern Recognition*, 2010, **43**(5): 1874−1888

4 Elfiky N M, Khan F S, van De Weijer J, Gonzàlez J. Discriminative compact pyramids for object and scene recognition. *Pattern Recognition*, 2012, **45**(4): 1627−1636

5 Wang F, Jiang Y G, Ngo C W. Video event detection using motion relativity and visual relatedness. In: Proceedings of the 16th ACM International Conference on Multimedia. NY, USA: ACM, 2008. 239−248

6 Liu J E, Yang Y, Saleemi I, Shah M. Learning semantic features for action recognition via diffusion maps. *Computer Vision and Image Understanding*, 2012, **116**(3): 361−377

7 Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006. 2169−2178

8 Yuan J S, Wu Y, Yang M. Discovery of collocation patterns: from visual words to visual phrases. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN: IEEE, 2007. 1−8

9 Du R, Wu Q, He X J, Yang J. Object categorization based on a supervised mean shift algorithm. In: Proceedings of the Computer Vision — EECV 2012 Workshops and Demonstrations. Berlin, Germany: Springer Berlin Heidelberg, 2012. 611−614

10 Chai Y N, Rahtu E, Lempitsky V, van Gool L, Zisserman A. TriCoS: a tri-level class-discriminative co-segmentation

method for image classification. In: Proceedings of the 2012 European Conference on Computer Vision. Berlin, Germany: Springer Berlin Heidelberg, 2012. 794−807

11 Krapac J, Verbeek J, Jurie F. Modeling spatial layout with fisher vectors for image categorization. In: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 1487−1494

12 Bolovinou A, Pratikakis I, Perantonis S. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition*, 2013, **46**(3): 1039−1053

13 Wang J J, Yang J C, Yu K, Lv F J, Huang T, Gong Y H. Locality-constrained linear coding for image classification. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA: IEEE, 2010. 3360−3367

14 van Gemert J C, Veenman C J, Smeulders A W M, Geusebroek J M. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(7): 1271−1283

15 Liu J, Zhang C J, Tian Q, Xu C S, Lu H Q, Ma S D. One step beyond bags of features: visual categorization using components. In: Proceedings of the 18th IEEE International Conference on Image Processing (ICIP). Brussels, Belgium: IEEE, 2011. 2417−2420

16 Avrithis Y, Kalantidis Y. Approximate Gaussian mixtures for large scale vocabularies. In: Proceedings of the 12th European Conference on Computer Vision. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2012. 15−28

17 Mikulík A, Perdoch M, Chum O, Matas J. Learning a fine vocabulary. In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2010. 1−14

18 Tang J H, Zha Z J, Tao D C, Chua T S. Semantic-gap-oriented active learning for multilabel image annotation. *IEEE Transactions on Image Processing*, 2012, **21**(4): 2354−2360

19 Wu L, Hoi S C H, Yu N H. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, 2010, **19**(7): 1908−1920

20 Ji C J, Zhou X D, Lin L, Yang W D. Labeling images by integrating sparse multiple distance learning and semantic context modeling. In: Proceedings of the 12th European Conference on Computer Vision. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2012. 688−701

21 Liu J E, Yang Y, Shah M. Learning semantic visual vocabularies using diffusion distance. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE, 2009. 461−468

22 Penatti O A B, Silva F B, Valle E, Gouet-Brunet V, Torres R S. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 2014, **47**(2): 705−720

23 Li L J, Wang C, Lim Y, Blei D M, Li F F. Building and using a semantivisual image hierarchy. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA: IEEE, 2010. 3336−3343

24 Bannour H, Hudelot C. Building semantic hierarchies faithful to image semantics. In: Proceedings of the 18th International Conference on Advances in Multimedia Modeling. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2012. 4−15

25 Bannour H, Hudelot C. Hierarchical image annotation using semantic hierarchies. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2012. 2431−2434

26 Deng J, Berg A C, Li K, Li F F. What does classifying more than 10 000 image categories tell us? In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2010. 71−84

27 Lorenza S, Jean-Daniel Z. *Abstraction in Artificial Intelligence and Complex Systems*. New York: Springer-Verlag New York Inc., 2013. 273−325

28 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

29 Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 2007, **106**(1): 59−70

30 Fan R E, Chang K W, Hsieh C J, Wang X R, Lin C J. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008, **9**: 1871−1874

31 Zhang Lin-Bo, Wang Chun-Heng, Xiao Bo-Hua, Shao Yun-Xue. Image representation using bag-of-phrases. *Acta Automatica Sinica*, 2012, **38**(1): 46−54
(张琳波, 王春恒, 肖柏华, 邵允学. 基于 Bag-of-phrases 的图像表示方法. 自动化学报, 2012, **38**(1): 46−54)

32 Fernando B, Fromont E, Muselet D, Sebban M. Supervised learning of Gaussian mixture models for visual vocabulary generation. *Pattern Recognition*, 2012, **45**(2): 897−907

33 Su Y, Jurie F. Improving image classification using semantic attributes. *International Journal of Computer Vision*, 2012, **100**(1): 59−77

34 Perronnin F, Sánchez J, Mensink T. Improving the Fisher kernel for large-scale image classification. In: Proceedings of the 11th European Conference on Computer Vision. Berlin, Germany: Springer-Verlag Berlin, Heidelberg, 2010. 143−156

35 Zhong J, Wang J, Su Y T, Song Z J, Xing S K. Balance between object and background: object-enhanced features for scene image classification. *Neurocomputing*, 2013, **120**: 15−23

36 Maji S, Berg A C, Malik J. Efficient classification for additive kernel SVMs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 66−77

37 Bilen H, Namboodiri V P, Van Gool L J. Object and action classification with latent window parameters. *International Journal of Computer Vision*, 2014, **106**(3): 237−251

刘 鹏 哈尔滨工业大学计算机科学与技术学院副教授. 2007 年获得哈尔滨工业大学微电子与固体电子学博士学位. 主要研究方向为图像处理, 视频处理, 模式识别, 超大规模集成电路设计.
E-mail: pengliu@hit.edu.cn
(**LIU Peng** Associate professor at the School of Computer Science and Technology, Harbin Institute of Technology. He received his Ph. D. degree of microelectronics and solid state electronics from Harbin Institute of Technology in 2007. His research interest covers image processing, video processing, pattern recognition, and design of very large scale integrated (VLSI) circuit.)

叶志鹏 哈尔滨工业大学计算机科学与技术学院博士研究生. 2013 年获得哈尔滨工业大学计算机应用技术硕士学位. 主要研究方向为模式识别, 机器学习.
E-mail: yezhipeng@hit.edu.cn
(**YE Zhi-Peng** Ph. D. candidate at the School of Computer Science and Technology, Harbin Institute of Technology. He received his master degree of computer application technology from Harbin Institute of Technology in 2013. His research interest covers image processing and machine learning.)

赵 巍 哈尔滨工业大学计算机科学与技术学院副教授. 2006 年获哈尔滨工业大学计算机应用技术博士学位. 主要研究方向为模式识别, 图像处理.
E-mail: zhaowei@hit.edu.cn
(**ZHAO Wei** Associate professor at the Pattern-Recognition Research Center, School of Computer Science and Technology, Harbin Institute of Technology. She received her Ph. D. degree of computer application technology from Harbin Institute of Technology in 2006. Her research interest covers image pattern recognition and image processing.)

唐降龙 哈尔滨工业大学计算机科学与技术学院教授. 1995 年获得哈尔滨工业大学计算机应用技术博士学位. 主要研究方向为模式识别, 图像处理, 机器学习. 本文通信作者.
E-mail: tangxl@hit.edu.cn
(**TANG Xiang-Long** Professor at the School of Computer Science and Technology, Harbin Institute of Technology. He received his Ph. D. degree of computer application technology from Harbin Institute of Technology in 1995. His research interest covers pattern recognition, image processing, and machine learning. Corresponding author of this paper.)