第 38 卷 第 1 期 自 动 化 学 报 Vol. 38, No. 1
2012 年 1 月 ACTA AUTOMATICA SINICA January, 2012

基于 Bag-of-phrases 的图像表示方法

张琳波 [1,2] 王春恒 [1] 肖柏华 [1] 邵允学 [1]

摘 要 在过去的几年, 将图像内容表示为特定 "视觉词" 出现次数直方图的 Bag-of-words 模型, 展示了其在图像内容分类方面的强大优势. 然而, 在这种统计特定 "视觉词" 出现次数直方图的模型中, "视觉词" 之间的相互位置关系几乎被完全丢弃了. 本文从分析 Bag-of-words 模型在文本分类和图像内容分类领域的对应关系的角度出发, 提出一种加入 "视觉词" 之间的相互位置关系的图像表示方法 — Bag-of-phrases 模型. 在标准数据集上验证了该图像表示方法对图像内容分类性能的影响. 实验结果显示, 本文提出的方法相对于传统的 Bag-of-words 模型可以达到更好的分类性能.

关键词 图像表示, 空间排列, Bag-of-words, Bag-of-phrases, SIFT 描述子

DOI 10.3724/SP.J.1004.2012.00046

Image Representation Using Bag-of-phrases

ZHANG Lin-Bo[1,2] WANG Chun-Heng[1] XIAO Bai-Hua[1] SHAO Yun-Xue[1]

Abstract Bag-of-words representation, with which an image is represented as a histogram of the numbers of occurrences of particular visual words, has demonstrated impressive levels of performance in the past few years. However, the relative position information between the visual words are almost entirely ignored. In this paper, the potential strength of this relative position information is investigated and a new kind of representation named Bag-of-phrases is proposed. The effectiveness of this strategy is validated on two benchmark databases. The classification results demonstrate that our Bag-of-phrases strategy can achieve better results compared to Bag-of-words method.

Key words Image representation, spatial layout, Bag-of-words, Bag-of-phrases, scale-invariant feature transform (SIFT) descriptor

随着数码相机等数码产品的广泛应用以及互联网的飞速发展, 可以利用的数字图像的数目正在以指数级的速度快速增长. 为了充分利用这些数字资源, 开发一套数字图像自动标注系统变得非常重要. 而图像内容分类的水平, 直接决定着自动图像标注系统的性能. 要开发一个良好的图像内容分类系统, 需要处理许多因素对分类性能的影响: 视角变化、亮度变化、背景干扰、图像内目标被遮挡、类内差异以及尺度变化等.

在现有机器学习领域分类技术性能确定的情况下, 图像内容表示方法的优劣, 直接决定着最终图像分类系统性能的高低. 因为如果不同类别的图像在被映射到的模式空间中不具备可分性 (如, 重叠), 再好的分类器也无法得到令人满意的图像分类性能. 鉴于图像内容表示的重要地位, 寻找更好的图像内容表示方法一直是图像内容分类的热点问题. 按照底层特征的不同, 图像内容分类方法大致可以分为: 基于全局特征的表示方法和基于局部特征的表示方法. 全局特征提取的对象通常是整幅图像, 将整幅图像看作一个整体并抽取图像中包括感兴趣区域和背景区域在内的全部信息. 常见的全局特征有: LBP (Local binary pattern)[1]、HOG (Histogram of oriented gradients)[2] 以及颜色直方图[3] 等. 在图像分类领域, 全局特征以其抽取速度快、可以有效利用图像上下文信息等优点, 在场景分类等图像分类系统中得到了广泛的应用. 然而, 全局特征描述的是整幅图像的信息, 而不能描述图像中的单独目标. 在这种情况下, 基于全局特征的图像表示方法对图像亮度变化、目标遮挡、目标位移甚至同一目标的变形都比较敏感, 严重限制了其在图像内容分类中的应用.

与基于全局特征的表示方法不同, 基于局部特征的图像内容表示方法由于可以在同一类别、千变万化的图像中寻找不变的部分, 受到了广大研究者的青睐. 基于局部特征的图像内容表示方法可以从 Neisser 的工作[4] 中找到认知生理学的支持: 人类的早期视觉分为两个阶段 — 前意识阶段 (Pre-attentive stage) 和意识阶段 (Attentive stage). 在前意识阶段, 只有那些 "显著" (pop-out) 局部区域

收稿日期 2011-07-28 录用日期 2011-10-09
Manuscript received July 28, 2011; accepted October 9, 2011
国家自然科学基金 (60802055, 60835001) 资助
Supported by National Natural Science Foundation of China (60802055, 60835001)
本文责任编委 刘一军
Recommended by Associate Editor LIU Yi-Jun
1. 中国科学院自动化研究所复杂系统智能控制与管理国家重点实验室 (筹) 北京 100190 2. 交通运输部科学研究院 北京 100029
1. Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. China Academy of Transportation Sciences, Beijing 100029

才会被人类视觉系统注意到; 而到了意识阶段, 人类视觉系统开始试着根据这些局部区域的特点建立它们之间的关系, 进而完成对所见内容的识别或者分类.

对于前意识阶段, 研究者提出了大量的局部特征. 例如: 局部 "显著" 区域可以利用角点检测子、水滴状检测子或者 "特殊" 区域检测子进行定位[5]. 然后利用基于滤波器的描述子、基于分布统计的描述子或者边缘相关的描述子将这些局部显著区域描述为局部特征向量[6−7].

对于意识阶段, 为了建模 "显著" 区域 (以下称局部特征) 之间的相互关系, 研究者提出了很多模型, 主要包括以下几类: 基于码本的模型[8−9]、部分 – 结构 (Part-structure) 模型[10−11]、轮廓 – 边缘 – 切割相关模型[12]、生物和认知相关模型[13] 以及场景相关模型[14] 等.

在上述建立局部特征关系的方法中, 基于码本的 Bag-of-words 模型[8] 由于其策略简单、对图像内目标位置和形变具有鲁棒性等优点, 在过去的几年得到了广泛的关注和显著发展. 图像分类中的 Bag-of-words 模型, 是受到文本分类中 Bag-of-words 模型的启发而发展起来的. 这种方法最初由 Zhu 等[15] 引入计算机视觉并应用于图像检索算法中. Csurka 等[8] 进一步给出了 "视觉词 (Visual words)" 的概念; 这里的 "视觉词" 相对于 Zhu 等[15] 的 "图像块 (Keyblock)" 具备了一定的不变性.

图像分类中 Bag-of-words 表示模型的基本思路如下: 首先, 在每幅图像中按照某种规则抽取一组局部特征; 然后, 选择部分局部特征, 用某种方式 (K-均值, Mean-shift[16] 等) 得到具有一定不变性的 "视觉词" 作为 "码字", 这些码字作为一个整体称为字典; 将每幅图像中被选择的所有局部特征按照某种规则 (最近邻、软分配 (Soft assignment)[17] 等) 映射到字典中的某个或某几个码字上; 最后, 在每幅图像中统计字典中各个码字在该图像中出现的次数, 得到直方图向量. 得到的直方图向量称为 Bag-of-words 向量, 表示对应的图像, 如图 1 所示.

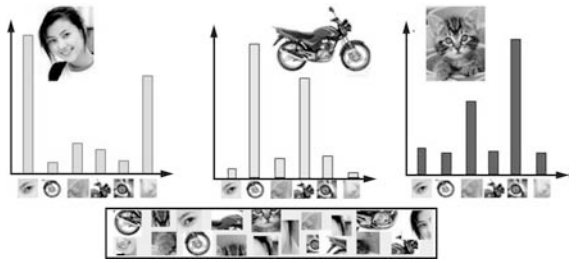


图 1　Bag-of-words 模型示意图

Fig. 1　Illustration of Bag-of-words model

Bag-of-words 方法丢弃了局部区域之间的位置排列关系. 然而, 局部特征之间位置关系的丢失是一把双刃剑. 一方面使得该图像表示方法具备一定的稳定性; 另一方面却限制了该图像表示方法的表示性能. 为了在 Bag-of-words 图像表示方法中加入位置信息, 研究者做了大量工作, 其中最知名的是 Lazebnik 等[18] 提出的空间金字塔 (Spatial pyramid) Bag-of-words 模型.

本文的总体架构安排如下: 第 1 节介绍空间金字塔 Bag-of-words 方法及其存在的问题; 第 2 节从分析图像分类领域 Bag-of-words 方法与文本分类领域 Bag-of-words 之间的关系出发, 寻找在基本 Bag-of-words 方法中加入局部特征之间位置排列关系的新思路; 第 3 节详细介绍本文的 Bag-of-phrases 模型; 第 4 节给出本文方法在国际公用数据集上的实验结果; 最后是总结和未来的工作.

1　问题描述

基于局部特征的 Bag-of-words 方法的一个重要不足就是丢弃了局部特征之间的空间排列信息, 而研究发现, 这种局部特征之间空间位置关系对提高图像表示性能是非常重要的. Grauman 等提出了快速计算两个局部特征集合之间核函数的空间金字塔匹配 (Spatial pyramid matching, SPM) 模型[19]. 该方法在匹配时首先对每个局部特征集合按照局部特征之间距离关系的不同度量分别建立直方图金字塔 (Histogram pyramids), 进而通过匹配不同金字塔层的直方图来判断两组局部特征是否相似. 在空间金字塔匹配模型的启发下, Lazebnik 等[18] 提出了空间金字塔 Bag-of-words 模型.

空间金字塔 Bag-of-words 模型首先逐层将图像细化分成一系列子区域, 然后分别计算各个子区域的 Bag-of-words 直方图表示, 从而为整幅图像建立起一种简单有效的 "空间金字塔" 直方图表示, 如图 2 所示. 最后采用金字塔匹配核实现图像分类. 这种建立空间金字塔 Bag-of-words 的方法在实际应用中取得了非常好的效果, 并得到了广泛应用.

然而, 随着分级层数的增加, 这种加入位置信息的策略对目标在图像中的位置、姿态越来越敏感. 当分层到一定深度, 该策略相对于基本 Bag-of-words 方法的优势就消失了. 因此, 在计算空间金字塔模型的过程中, 最多将图像细分为 4×4 个子图, 如图 2 所示.

即使这样, 这种加入位置信息的策略还是存在不足. 考虑 $L = 1$, 即将整幅图像划分为 2×2 个子图的情形.

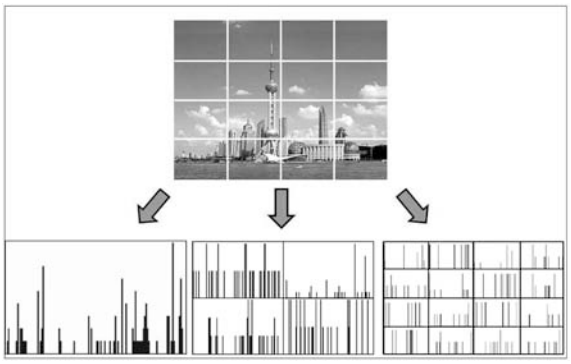


图 2 空间金字塔 Bag-of-words 模型示意图

Fig. 2 Illustration of spatial pyramid Bag-of-words models

如果目标在图像中间位置, 该策略可以很好地建立局部特征之间的几何对应关系. 然而, 如果目标出现在偏离中心区域的位置, 则会产生另外的效果. 考虑一种极端情况, 由于拍摄过程中镜头的小偏移, 使得目标分别出现在图像偏左和偏右的位置, 分别如图 3 (a) 和 3 (b) 所示. 从人类视觉的角度考虑, 图 3 (a) 和 3 (b) 两幅图像是同一只狗在相同背景条件下的图片, 而图 3 (c) 是一只猫的图片. 但是对于空间金字塔 Bag-of-words 图像表示方法来说, 图 3 (c) 却有着与图 3 (b) 更相似的 2 × 2 Bag-of-words 直方图. Bag-of-words 模型的最大优点就是对目标在图像中的姿态、位置具有很好的稳定性. 但是如果采用了这种空间金字塔 Bag-of-words 策略, 这种稳定性就会受到削弱, 当子图划分到某种层次的时候, 这种稳定性就会消失.

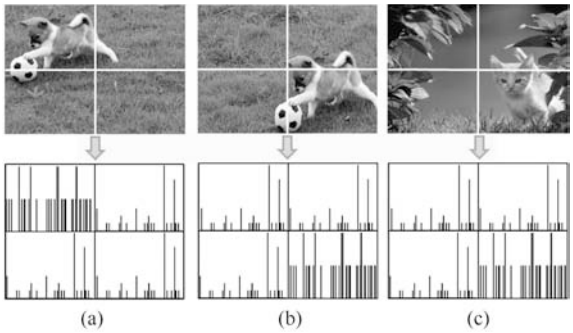


图 3 不同内容图像的空间金字塔 Bag-of-words 表示向量

Fig. 3 Spatial pyramid Bag-of-words vectors of different image contents

虽然空间金字塔 Bag-of-words 模型在建立局部特征之间的几何对应关系方面存在一定的局限性, 但以其优于基本 Bag-of-words 方法的性能给了我们两点启示: 1) 局部特征点之间的空间排列信息是非常有用的; 2) 应该存在另外一种方法, 既能加入了空间信息, 又可以充分发挥 Bag-of-words 模型的优势.

来自仿生学的研究结果表明[13], 在人类等灵长类视觉系统的高层 (High-level), 空间位置信息变得不再重要并被丢弃. 这个结论很好地解释了基本 Bag-of-words 表示方法的成功; 然而却与空间金字塔 Bag-of-words 模型拥有更好的性能相矛盾. 因为如果空间位置信息真的没用, 为什么空间金字塔 Bag-of-words 图像表示方法会取得比基本 Bag-of-words 好得多的性能? 前面已经提到过基于部分的模型与人类视觉系统在大方向上是基本吻合的[4], 难道 Bag-of-words 图像表示方法与人类视觉是相矛盾的吗? 我们认为出现上述问题的原因是并不确定图像中的局部特征的组织阶段是否达到了文献 [13] 中所谓高层. 也就是说, 局部特征的组织阶段可能尚未达到对应人类视觉系统所谓的高层的层次. 而在人类视觉系统的初始阶段, 位置信息的重要性是显而易见的. 因此, 本文的研究目标就是提高局部特征的层次, 使之达到或者接近位置信息可以被抛弃的层次, 进而充分发挥 Bag-of-words 图像表示方法稳定性的优势, 从而避免由于位置信息的引入而出现对目标位置、姿态敏感的情况.

由于计算机视觉界的 Bag-of-words 图像内容表示方法是从文档分类领域引进的, 因此本文从分析图像分类领域 Bag-of-words 图像表示方法与文档分类领域 Bag-of-words 文档表示方法之间关系的角度出发, 试图寻找一种既能在基本的 Bag-of-words 模型中引入局部特征之间的空间排列信息, 又不失该模型稳定性的方法.

2 图像内容分类与文档分类中概念之间的关系

在大部分 Bag-of-words 图像表示模型中, 图像和局部特征直接被作为文本分类中文档 (Document) 和词 (Word) 的等价物. 一般来说, 让图像和文档之间建立联系是合理的, 因为, 不论一幅图像还是一篇文档, 都在给人们传递一个事件. 但是, 直接让图像中的局部特征对应于文档中的词就带有随意性, 因为目前还没有现成的理论可以为这种对应提供依据.

严格来讲, 计算机视觉领域的局部特征可以对应文本分类领域中 “文本” 这个概念以下的所有概念, 而不仅仅是 “词”. 大多数情况下, 局部特征中包含的信息, 远远少于如 “头”、“耳朵” 或 “车轮” 这样的词在文本中所包含的信息, 如图 4 所示. 当然, 局部特征也不应该对应着文本分类领域的 “笔画” (英语中的 “字母”); 否则, 计算机视觉中的 Bag-

of-words 模型就会对应着文本分类中统计笔画 (或者字母) 的出现次数直方图, 而这种无意义的模型不会得到 Bag-of-words 图像表示所呈现出的良好性能.

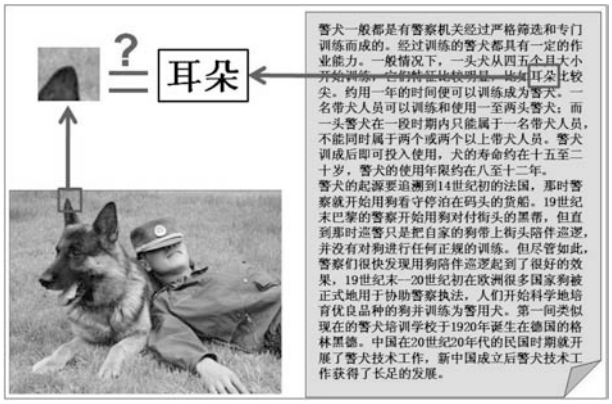


图 4　图像中的局部特征与文档中词对应关系

Fig. 4　The relationship between local features in image and words in document

基于这些考虑, 本文假设计算机视觉中的局部特征对应着文本分类领域的介于笔画 (或者字母) 与词语之间的某种概念 (如, 部首、偏旁). 如果该假设成立, 通过为每种局部特征增加它与其邻域内其他局部特征之间的位置关系, 进而提高局部特征对应的语义层次, 对提高图像表示性能将很有帮助. 因为词语是由笔画或者偏旁、部首与其他笔画或偏旁之间的位置关系构成的. 即使上面的假设不成立 (如, 局部特征对应文档中的词语), 那增加与其他词语之间的位置关系会将词语提升到更高的层次 (如, 短语或句子), 这对表示性能也是有改进的, 因为统计短语或者句子的出现次数直方图也是有意义的.

本文接下来的部分将详细介绍如何为指定局部特征增加其邻域内其他局部特征的位置信息, 使之提高到一个更高层次 (指的是从 "视觉词" 提高到 "视觉短语"), 进而将局部特征之间的空间位置关系融入到图像的表示中.

3　Bag-of-phrases 模型

由于在计算机视觉界 "视觉词 (Visual words)" 已经被广泛地用来指图像中的局部特征, 本文把融入了与相邻局部特征位置排列关系的局部特征称为 "视觉短语 (Visual phrase)". 然后, 根据 Bag-of-words 模型的方法可以直接得到 Bag-of-phrases 模型.

3.1　主要框架

本文的 Bag-of-phrases 模型的主要框架为:

1) 显著区域检测. 应用某种或者几种局部特征检测子, 在每幅图像上得到一组显著特征点. 假设在图像 I_i 中检测到 N_i 个显著区域, 记为 $\boldsymbol{P}_i = \{p_{i1}, p_{i2}, \cdots, p_{iN_i}\}$;

2) 局部区域纹理描述. 利用某种局部区域描述算子来描述所有局部显著特征点 p_{ij} 周围的区域, 可以得到点 p_{ij} 周围的纹理描述, 记为局部特征 $\boldsymbol{f}_{ij}$. 这个过程也可以概括为: 利用视觉字 (Visual word) $\boldsymbol{f}_{ij}$ 来描述点 p_{ij} 的语义.

3) 邻域内视觉词空间排列建模. 对每个视觉词 (如 $\boldsymbol{f}_{ij}$), 将其邻域内其他视觉词的空间排列信息编码在向量 $\boldsymbol{s}_{ij}$ 中;

4) 视觉短语计算. 将点 p_{ij} 对应的 $\boldsymbol{f}_{ij}$ 和 $\boldsymbol{s}_{ij}$ 组合在一起, 形成 "视觉短语", 记为 $\boldsymbol{v}_{ij}$;

5) 码本构建. 在大量视觉短语基础上, 采用现有 Bag-of-words 模型中常用的某种码本构建技术得到固定数量的具有代表性的 "视觉短语" 作为 "码本短语 (Code phrase)", 这些码本短语作为一个整体称为码本;

6) 利用 Bag-of-phrases 模型进行图像表示. 将每一幅图像里面所有的 Visual phrase 都映射到码本中的 Code phrases 上. 然后, 统计每个 Code phrase 出现的次数得到直方图向量, 称为 Bag-of-phrases 向量.

此时, 得到的 Bag-of-phrases 向量可以直接被用来作为图像的向量表示用于分类. 需要注意的是, 除了邻域内视觉词空间排列建模部分, 每一步在 Bag-of-words 里都有现成的技术可以利用. 下面给出邻域内视觉词空间排列建模的详细介绍.

3.2　邻域内视觉词空间排列建模

在本文方法中, 视觉词之间的空间排列建模是受 Shape-context 方法[20] 的启发而得到的. 即利用直方图向量来描述视觉词的局部空间排列. 以 $\forall I_i$ 图像为例 (如图 5 (a) 所示), 详细描述视觉短语的形成过程.

1) 对图像 I_i 中的每一个特征点 p_{ij} 周围邻域, 沿着半径可以分成 R 个箱, 按方向可以分为 O 个箱 (如图 5 (b) 所示). 需要注意的是, 最小的圆里面并没有被按照方向被分为 O 个箱. 因为在这个圆中, 某个视觉词的轻微移动, 就会使得特征点邻域的点落入不同的方向上; 如果对这个小圆也进行划分, 将会导致结果不稳定. 最小的半径不能太小, 否则里面点的分布不稳定; 最大的半径不能太大, 否则得到的局部短语不符合局部的优势而导致性能的下降.

2) 基于这一划分, 统计每个箱中视觉词出现的

次数, 可以得到直方图向量 $\boldsymbol{s}_{ij}$.

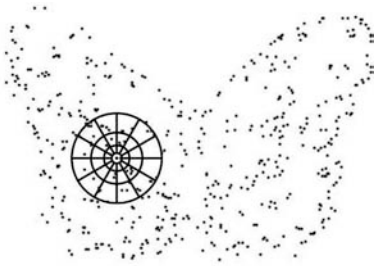
3) 为了使得这种描述空间排列的方式对旋转鲁棒, 在统计直方图向量 $\boldsymbol{s}_{ij}$ 的时候借助选择主方向的方法. 具体方式如下: 将同一个方向上所有箱中的视觉词个数相加, 然后选择具有最多视觉词的方向作为特征点 p_{ij} 的主方向. 统计每个箱内视觉词出现次数直方图时, 不同圆环按照从内到外的顺序. 同心环不同方向的箱, 从主方向开始按照顺时针方向依次统计.

4) 一般来说, 每个视觉词对距离比较近的视觉词比距离比较远的视觉词要更加敏感一些. 因此, 对直方图 $\boldsymbol{s}_{ij}$ 中各分量的数值按照它对应的箱距离中心点的远近进行加权. 例如: 将直方图中从第 t_{th} 个圆环上统计得到的分量都乘以 $1/\delta_t$.

最后, $\boldsymbol{s}_{ij}$ 即为关键点 p_{ij} 邻域内其他视觉词的空间排列信息的向量化描述.

(a) 原始图像中局部特征位置

(a) Local features on original image



(b) 同心圆圆心位置局部区域内其他视觉词建模 (4 个半径格, 12 个方向格)

(b) Spatial layout of neighbor visual words around the center of concentric circles in 4 radius bins and 12 orientation bins

图 5 视觉词空间排列建模示意图

Fig. 5 Modeling spatial layout of visual words

3.3 其他细节

理论上说, 在关键点检测和视觉词生成阶段, 任意现有的局部检测子和描述子都可以直接使用. 然而, 在本文方法的当前阶段, 视觉词用来表示关键点区域的语义信息, 而与其他视觉词的关系是通过它们的位置得到的. 因此, 理想的检测子应该是能够捕获图像中的结构信息, 而描述子最好选择适合表示纹理的描述子.

在视觉短语产生阶段, 对每一个关键点 p_{ij}, 它的 "语义" f_{ij} 和它邻域内其他视觉词的空间排列信息 $\boldsymbol{s}_{ij}$ 组合在一起得到视觉短语 v_{ij}. 完成这个操作的一种方式是直接把两个向量链接成一个长向量. 然而, 由于两个向量是采用不同的机制计算得到的, 因此在进行链接操作之前要进行向量归一化或者对两个向量进行加权.

由于目标在图像中变化的连续性以及目标遮挡变形等因素的影响, 相同目标局部区域在不同图像中得到视觉短语的变化也是连续的. 为了将这些连续变化用固定长度的向量表示, 本文采用 Bag-of-words 模型的组织策略. 即固定数量的具有代表性的视觉短语被选择或者计算出来作为词典短语. 这个阶段与基本 Bag-of-words 模型中通过视觉词计算码本词 (Code word) 的原理和方式没有明显差异. 因此, 在 Bag-of-words 模型中现有的技术 (如, K-均值、Mean-shift 或高斯混合模型等) 都可以直接用于完成对 Bag-of-phrases 模型中码本的构建.

同样的, 在计算 Bag-of-phrases 向量时, 所有计算 Bag-of-words 向量的方法都可以直接应用, 如: 最近邻, 软分派[17] 等.

4 实验结果与分析

为了验证本文的 Bag-of-phrases 图像表示方法的有效性, 将提供本文方法在两个标准国际数据集, 即 Caltech-101 和 PASCAL Visual object challenge (VOC) 2006 上的测试结果. 实验具体设置如下:

1) 在显著区域检测阶段, 选择基于小波的检测子[21]. 其中, 每幅图像需要得到的关键点的个数设置为 N=1 000;

2) 每个关键点 p_{ij} 周围区域的纹理特性 $\boldsymbol{f}_{ij}$ 采用 SIFT 描述子[22] 进行描述;

3) 当按第 3.2 节所述计算关键点 p_{ij} 邻域内其他视觉词的局部空间排列信息时, M 和 O 分别设置为 5 和 12, 其中 5 个圆的半径分别为: 4, 8, 14, 22, 32;

4) 在 $\boldsymbol{s}_{ij}$ 除以 δ_t 之前, 它的每个分量首先被缩放到 $[0, 255]$, 以协调其每个分量和 SIFT (Scale-invariant feature transform) 描述子分量部分. δ_t 的设置遵循 $\delta_1 < \delta_2 < \delta_3 < \delta_4 < \delta_5$. 初始阶段设置 $\delta_t = t$, 然后对它们进行微调, 来决定空间位置信息的影响;

5) 在计算码本的过程中, 选择一部分视觉短语进行 K-均值聚类, 这个过程和 Csurka 等[8] 的

Bag-of-words 方法中计算码本的过程相同;

6) Bag-of-phrases 向量计算完成后, 采用 LibSVM 分类器进行训练和测试. SVM 核函数矩阵的计算采用 $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\{-d_{\chi^2}(\boldsymbol{x}_1, \boldsymbol{x}_2)/\sigma\}$, 其中, σ 设置为 100. $d_{\chi^2}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ 计算如下:

$$d_{\chi^2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{2}\sum_{q=1}^{Q}\frac{[\boldsymbol{x}_1(q) - \boldsymbol{x}_2(q)]^2}{\boldsymbol{x}_1(q) + \boldsymbol{x}_2(q)}$$

其中, $\boldsymbol{x}_1(q)$, $\boldsymbol{x}_2(q)$ 分别指 Q-维向量 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的第 q 个分量.

4.1 Caltech-101 数据集实验结果

Caltech-101 包含 101 个目标类, 共 9 197 幅图像[23]. 为了利用现有的数据集训练分类器并测试算法的性能, 在每个类别中随机选择 30 幅图像组成训练图像集合, 其余图像构成测试图像集合. 也就是说, $101 \times 30 = 3\,030$ 幅图像被用来训练分类器, 所有剩余图像用来评价训练出来的分类器性能, 也间接地测试本文提出的 Bag-of-phrases 模型的图像表示能力.

每一个类别都训练 1 个一对多的支持向量机 (Support vector machine, SVM) 分类器. 训练完成后, 可以得到 101 个分类器, 记为: $\zeta_1, \zeta_2, \cdots, \zeta_{101}$. 每幅测试图像 $I\tau$ 依次通过这些分类器, 得到 101 个概率值, 记为 $p_\tau^1, p_\tau^2, \cdots, p_\tau^{101}$. 最后, 将给出最大概率值的分类器所对应的类别作为测试图像 I_τ 的预测标签 $l_\tau^{'}$, 即:

$$l_\tau^{'} = \arg_t \max_{t=1}^{101} p_\tau^t$$

如果预测标签 $l_\tau^{'}$ 与 I_τ 的真实类别标签 l_τ 相同, 则分类系统对测试图像 I_τ 分类正确; 否则, 分类错误. 表 1 给出了本文方法在 Caltech-101 数据集上的分类结果, 对比实验分别是基本的 Bag-of-words 图像表示方法[8] 和空间金字塔 Bag-of-words 方法[18].

从表 1 可以看出, 本文的 Bag-of-phrases 方法可以得到比基本 Bag-of-words 方法更高的准确率. 然而, 它比空间金字塔 Bag-of-words 表示方法仅仅改进了一点. 原因可能是 Caltech-101 数据集的每幅图像只包含一个对象, 且所含对象全部位于图像的中间位置并占据图像大部分区域, 而这恰好有利于空间金字塔 Bag-of-words 表示方法发挥优势. 基于这方面的考虑, 本文设计的第二个实验选择更具有挑战性的 PASCAL VOC 2006 数据集. 在这个数据集上, 目标对象可能以任意尺度出现在图像的任意位置.

表 1　数据集 Caltech-101 上的精确度对比 (%)
Table 1　Comparison of precisions on Caltech-101 dataset (%)

方法	精确度 (%)
Bag-of-words, $K = 1\,000$	60.97 ± 0.015
空间金字塔Bag-of-words, $K = 1\,000$	63.01 ± 0.054
Bag-of-phrases, $K = 1\,000$	$\mathbf{64.02 \pm 0.048}$
Bag-of-words, $K = 2\,000$	59.81 ± 0.022
空间金字塔Bag-of-words, $K = 2\,000$	62.83 ± 0.066
Bag-of-phrases, $K = 2\,000$	$\mathbf{64.85 \pm 0.032}$

4.2 PASCAL VOC 2006 数据集上的实验结果

选择 PASCAL VOC (Visual object challenge 2006) 数据集而不是最新的 PASCAL VOC 2010 数据集是基于以下考虑: 1) PASCAL VOC 2010 测试数据集的标准标注文件没有公布. 2) 在 PASCAL VOC 2006 竞赛中, 有参赛队伍选择单纯采用简单的 Bag-of-words 方法或者空间金字塔 Bag-of-words 图像表示方法, 有利于将本文的方法和其他方法进行一个公正的对比. 而在 PASCAL VOC 2010 竞赛中, 很多参赛队伍为了达到好的结果, 将 Bag-of-words 图像表示策略和其他各种各样的技术融合在一块; 这样得到的结果往往是所有这些技术共同作用的结果, 因而不适合一个简单算法与之进行性能比较. 3) PASCAL VOC 2006 数据集包含 10 个类别的 5 304 幅图像, 而 PASCAL VOC 2010 数据集中包含了 20 个类别的 21 738 幅图像, 选择 PASCAL VOC 2006 数据集可以减少实验的时间. 此外, 如果采用 PASCAL VOC 2010 数据集, 在训练一对多的分类器时, 问题的焦点很容易被转移到如何处理正样本数目远远少于负样本数目的情况. 因此, 考虑本文方法的特点, 我们选择 PASCAL VOC 2006 数据集作为本文第二个实验的数据集.

PASCAL VOC 2006 数据集中共包含 10 个类别的 5 304 幅图像, 10 个类别为: 自行车 (Bicycle)、公交车 (Bus)、轿车 (Car)、猫 (Cat)、母牛 (Cow)、狗 (Dog)、马 (Horse)、摩托车 (Motorbike)、人 (Person) 和绵羊 (Sheep). 根据 PASCAL VOC 2006 组织者的划分, 这些图像中的 2 618 幅图像被用来作为训练图像, 剩下的 2 686 幅图像被作为测试图像来测试待评价方法的性能.

在所有参赛队伍中, INRIA Larlus 组[24] 采用的是基本 Bag-of-words 法, 因此选择他们的结果与本文的方法在本数据集上的结果进行对比; 为了达到最好的性能, INRIA Larlus 组在每幅图像中检

测了多达 10 000 个显著区域并将码本大小设置为 $K = 5\,000$. 此外, 本文也给出了空间金字塔 Bag-of-words 在 PASCAL VOC 2006 数据集上的性能. 在空间金字塔 Bag-of-words 实验中, 每幅图像被划分为 1×1 和 2×2 的子图像.

为了与 PASCAL VOC 2006 竞赛中的评价策略一致, 本实验的性能采用 ROC (Receiver operating characteristic) 曲线下的面积进行计算. 也就是说, ROC 曲线下的面积越大, 它对应的算法的性能越好. 实验的对比结果见图 6, 其中空间金字塔 Bag-of-words 在图 5 中标记为 Spatial bag-of-words.

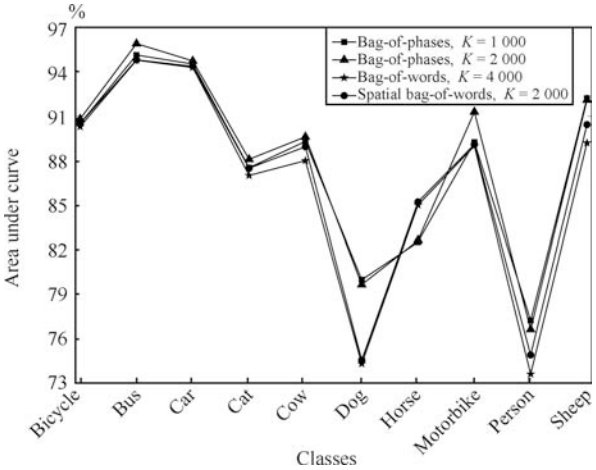


图 6 图像中的局部特征相相对于文档中的词之间的关系

Fig. 6 The relationship between words in document and local features in image

从图 6 可以看到, 虽然本文的 Bag-of-phrases 方法在每幅图像中仅仅使用了 1 000 个显著区域, 并使用比 INRIA Larlus 组[24] 小得多的码本, 却可以达到更好的性能. 此外, 本文的方法对狗、摩托车、人以及绵羊类别图像分类性能比 Bag-of-words 方法、空间金字塔 Bag-of-words 方法优势明显, 在马类别图像上性能最差; 对剩余的类别, 本文的方法比基本 Bag-of-words 方法和空间金字塔 Bag-of-words 方法好一点, 优势不明显. 同时, 利用大小 $K = 2\,000$ 的码本比 $K = 1\,000$ 的码本, 可以达到略好的性能.

4.3 讨论

从上面提供的实验结果来看, 本文的 Bag-of-phrases 方法性能最好, 这也恰好与本文的观点是一致的. 然而, 也可以很明显地看到, 本文的 Bag-of-phrases 方法的性能并不如预期的好, 它仅仅比空间金字塔 Bag-of-words 好一点, 而总体性能比目前国际上计算机视觉界在这两个数据集上的最好结果差很多. 出现这种情况的原因可能是, 本文在计算关键点 (如 p_{ij}) 处视觉短语时, 仅仅考虑了关键点 p_{ij} 所在的局部区域的 "语义信息"(也就是纹理特性) 和在其邻域内其他视觉词的位置信息, 却没有考虑点 p_{ij} 周围其他视觉词的语义信息. 换句话说, 目前本文的方法仅仅考虑了点 p_{ij} 周围其他词是怎么分布的, 却没有关注那些词是什么.

基于这方面的考虑, 研究一种方法在计算视觉短语的时候不但考虑其他视觉词的位置排列情况, 并且对那些视觉词的 "语义信息" 也进行建模, 是改进这个方法的关键所在. 也许, 结构模式识别的某些知识可以对这个问题有所帮助.

此外, 本文的 Bag-of-phrases 方法是建立在这样的假设之上的: 计算机视觉中的局部特征对应着文本分类中介于笔画 (或字母) 和词之间的某个概念. 实际上, 大部分情况也确实如此, 例如图 4 中, 局部特征仅占据了耳朵的一小部分. 然而, 当目标在图像中比例非常大或者非常小的时候, 情况就比较复杂了. 例如: 当整幅图像只包含狗的一个耳朵时, 局部特征可能仅占据整个耳朵的很小一部分, 这种情况下应用本文的策略还是问题不大; 但是, 当图像中包含的是巨大建筑物下面站着一只小狗的时候, 局部特征可能包含了整个狗的全部信息, 如果此时再采用本文的策略, 就不是那么合理了.

从上面的分析可以得出如下结论: 在多个尺度上对图像进行表示和分析应该是更合理的方式. 例如, 给定一幅图像, 可以在多个尺度上计算不同的直方图向量. 在局部特征的基础上, 计算 Bag-of-words 直方图向量; 建模局部区域内各局部特征之间的排列和语义等信息, 可以得到视觉短语, 然后在这些视觉短语的基础上计算 Bag-of-phrases 直方图向量; 建模短语之间的关系得到视觉句子 (Visual sentences), 然后在这些句子的基础上计算 Bag-of-sentences 直方图向量等. 最后, 利用所有的这些直方图向量共同表示一幅图像, 并输入分类器, 让分类器实时决定在遇到特定向量的时候, 哪些信息重点采用, 哪些信息仅供参考, 哪些信息放弃.

此外, 文献 [25−26] 也用到了 Visual phrase 的概念. 与本文利用局部特征之间的位置关系提高局部特征层次不同的是, 文献 [25] 利用局部特征 (Visual words) 之间的共现关系确定 Visual phrase, 然后利用得到的 Visual phrase 重新处理局部特征, 达到去除局部特征中常见的同义词和多义词现象的目的. 与本文和文献 [25] 都不同的是, 文献 [26] 将 Visual word 定义为图像中的单个目标, 而将目标之间的关系定义为 Visual phrase, 然后利用单目标检测和目标之间关系检测结合的方式提高目标检测性能.

5 结论

本文分析了计算机视觉领域图像内容分类中 Bag-of-words 方法的优越性能以及它与文本分类领域中 Bag-of-words 的对应关系. 然后, 提出了视觉短语的概念. 在视觉短语中, 考虑了局部特征和周围其他局部特征之间的空间排列关系; 在视觉短语的基础上, Bag-of-words 表示模型被直接提升到了 Bag-of-phrases 表示模型. 实验结果表明, 本文提出的 Bag-of-phrases 图像表示方法是行之有效的.

最后, 提出了两种改进本文方法的思路: 第一种针对计算某个位置的视觉短语阶段, 在计算该位置视觉词 (也称语义信息) 和周围其他视觉词位置分布的同时, 也要考虑周围其他视觉词的语义信息; 另一种就是, 图像表示可以考虑采用多尺度表示, 然后由分类器最终决定采用哪种信息.

References

1 Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(7): 971−987

2 Dalal N, Triggs B, Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893

3 Swain M J, Ballard D H. Color indexing. *International Journal of Computer Vision*, 1991, **7**(1): 11−32

4 Neisser U. Visual search. *Scientic American,* 1964, **210**(6): 94−102

5 Tuytelaars T, Mikolajczyk K. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision,* 2008, **3**(3): 177−280

6 Mikolajczyk K, Schmid C. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2005, **27**(10): 1615−1630

7 Li J, Allinson N. A comprehensive review of current local features for computer vision. *Neurocomputing,* 2008, **71**(10−12): 1771−1787

8 Csurka G, Dance C R, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision. Prague Czech Republic: ECCV, 2004. 1−22

9 Yang J C, Yu K, Gong Y H, Huang T. Linear spatial pyramid matching using sparse coding for image classication. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Mimi, USA: IEEE, 2009. 1794−1801

10 Fergus R, Perona D, Zisserman A. Object class recognition by unsupervised scale-invariant learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA: IEEE, 2003. 264−271

11 Felzenszwalb P F, Huttenlocher D P. Pictorial structures for object recognition. *International Journal of Computer Vision,* 2005, **61**(1): 55−79

12 Shotton J, Blake A, Cipolla R. Multiscale categorical object recognition using contour fragments. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2002, **30**(7): 1270−1281

13 Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2007, **29**(3): 411−426

14 Torralba A, Murphy K P, Freeman W T. Contextual models for object detection using boosted random fields. In: Proceedings of the Neural Information Processing Systems. Vancouver, Canada: NIPS, 2004. 1401−1408

15 Zhu L, Rao A B, Zhang A D. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems,* 2002, **20**(2): 224−257

16 Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2002, **24**(5): 603−619

17 Gemert J C, Geusebroek J M, Veenman C J, Smeulders A W M. Kernel codebooks for scene categorization. In: Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2008. 696−709

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

19 Grauman K, Darrell T. The pyramid match kernel: discriminative classication with sets of image features. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Beijing, China: IEEE, 2005. 1458−1465

20 Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2002, **24**(4): 509−522

21 Loupias E, Sebe N, Bres S, Jolion J M. Wavelet-based salient points for image retrieval. In: Proceedings of the International Conference on Image Processing. Vancouver, Canada: IEEE, 2000. 518−521

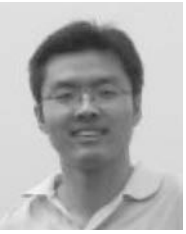
22 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision,* 2004, **60**(2): 91−110

23 Li F F, Fergus F, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop. Washington D. C., USA: IEEE, 2004, 178−187

24 Everingham M, Zisserman A, Williams C, Gool L V. The PASCAL visual object classes challenge 2006 (VOC2006) results [online], available: http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, July 20, 2011

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

26 Sadeghi M A, Farhadi A. Recognition using visual phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2011. 1745−1752



张琳波 中国科学院自动化研究所博士研究生. 主要研究方向为模式识别与机器学习, 图像内容分类. 本文通信作者.
E-mail: linbo.zhang@ia.ac.cn
(**ZHANG Lin-Bo** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, machine learning, and content based image classification. Corresponding author of this paper.)

王春恒 中国科学院自动化研究所研究员. 主要研究方向为模式识别与智能系统, 图像处理, 字符识别, 人工智能.
E-mail: chunheng.wang@ia.ac.cn
(**WANG Chun-Heng** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, intelligent systems, image processing, and character recognition, and artificial intelligence.)

肖柏华 中国科学院自动化研究所研究员. 主要研究方向为模式识别与智能系统, 多媒体信息处理与检索.
E-mail: baihua.xiao@ia.ac.cn
(**XIAO Bai-Hua** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, intelligent systems, and multimedia information processing and retrieval.)

邵允学 中国科学院自动化研究所博士研究生. 主要研究方向为模式识别与机器学习, 手写汉字识别.
E-mail: yunxue.shao@ia.ac.cn
(**SHAO Yun-Xue** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, machine learning, and handwritten Chinese character recognition.)