# 基于BoF模型的图像表示方法研究

梁 晔[1,2] 于 剑[2] 刘宏哲[3]

(北京联合大学信息学院 北京 100101)[1] (北京交通大学计算机与信息技术学院 北京 100044)[2]

(北京联合大学北京市信息服务工程重点实验室 北京 100101)[3]

**摘 要** 设计合适的图像表示是计算机视觉中最重要的问题之一。BoF特征表示方法非常流行,已经广泛应用于图像分类、对象识别、图像检索、机器人定位和纹理识别。BoF特征是将图像表示为无序的特征集合。这种方法虽然缺乏结构信息和空间信息,但概念简洁、计算简单,在某些应用上取得的效果甚至可以与当前最好的方法媲美。仔细研究了BoF模型,着重对BoF模型中的3个阶段:局部特征提取、特征量化和编码、特征汇集所涉及到的典型技术进行了讨论。最后在分析各类研究方法的基础上,总结了目前研究存在的问题及可能的发展方向。

**关键词** 特征包,局部特征,特征量化,特征汇集,计算机视觉

**中图法分类号** TP317.4 **文献标识码** A

## Study of BoF Model Based Image Representation

LIANG Ye[1,2] YU Jian[2] LIU Hong-zhe[3]

(College of Information, Beijing Union University, Beijing 100101, China)[1]

(Institute of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)[2]

(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)[3]

**Abstract** Designing a suitable image representation is one of the most fundamental issues of computer vision. BoF model is very popular and used extensively in image classification, video search, robot localization and texture recognition. BoF feature is an orderless collection of quantized local image descriptors. While this feature representation discards structural and spatial information, BoF model is conceptually and computationally simple, even as good as stateof- the-art methods. Three steps in the popular BoF were studied in detail, including feature extraction, feature coding and feature pooling. In the end, the main problems and challenges were highlighted based on analysis of current research technique.

**Keywords** BoF, Local features, Feature quantization, Feature pooling, Computer vision

在过去的10多年里面,BoF模型在计算机视觉中非常重要,是近年来在计算机视觉领域应用最广泛的一类特征,已经应用于图像分类、对象识别、图像检索、机器人定位和纹理识别。大量研究结果表明BoF特征在计算机视觉中具有很好的性能。BoF特征[1],也称为Bag-of-Features或者Bag-of-Visterms,其思想来源于文本信息检索和分类任务中的文档表示技术,亦即将图像表示为无序的特征集合。BoF特征是通过统计局部不变特征的全局出现情况来实现的,其特征既保留了局部特征的不变性又增加了全局特征的鲁棒性,同时与数量庞大的局部不变特征相比还能起到简化特征的作用,是对图像的压缩表示,但是丢失了特征的空间、相对位置、尺度和方向信息。构建BoF图像表示的过程如图1所示,包括特征提取、字典的生成、特征量化和编码以及特征的汇集。

要将文档表示的词袋描述方法应用到图像表示领域,最关键的步骤就是构建出与文本中的词相对应的图像的视觉单词(Visual Words)。这一过程首先是通过聚类训练图像库中的局部不变特征建立视觉词典(Visual Vocabulary),聚类方法可以是K均值聚类[2]、层次K均值聚类[3]、高斯混合模型[4]、Mean-Shift[5]等,然后再建立图像中出现的局部不变特征与视觉词典中视觉单词的映射来实现。在获取了视觉词典和图像中出现的视觉单词后,就能通过统计图像中每个视觉单词出现的次数来生成最终的BoF特征。

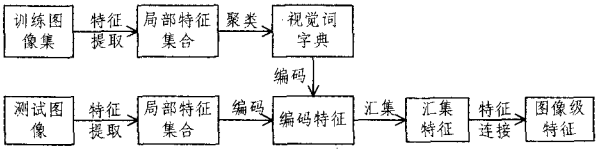


图1 基于BoF模型的图像表示流程

BoF表示的每个步骤在设计上都有很多选择。本文将仔

到稿日期:2013-05-20 返修日期:2013-08-01 本文受国家自然科学基金项目(60905028,61033013),北京市自然科学基金(4112046),北京联合大学"新起点"项目(ZK201211)资助。

梁 晔(1978—),女,博士生,讲师,主要研究领域为图像处理,E-mail:liangye@buu.edu.cn;于 剑(1969—),男,教授,博士生导师,主要研究领域为机器学习、图像处理等;刘宏哲(1971—),女,博士,副教授,主要研究领域为语义计算和图像处理,E-mail:xxtliuhongzhe@buu.ed.cn(通信作者)。

细讨论 BoF 模型，着重对局部特征提取、特征的量化和编码、特征的汇集这 3 个阶段所需要的典型技术进行综述，使读者充分了解 BOF 模型，并探讨未来发展遇到的问题和挑战。

本文第 1 节论述局部特征提取技术，包括特征点的检测和描述；第 2 节讨论局部特征的量化和编码技术；第 3 节讨论特征的汇集技术；最后对存在的问题进行展望。

## 1 局部特征提取

图像特征是对图像特性或属性的描述。图像特征的提取和表示是图像处理的基础，在对图像进行 BoF 表示时，特征提取非常重要。当前，在 BoF 处理流程中提取的特征为图像的局部特征。局部特征的提取包括特征点检测和特征点描述。目前存在大量的局部特征提取方法，但是判断什么是好特征的标准依赖于应用，没有统一的标准来衡量。

### 1.1 特征点检测

特征点检测的主要目的是确定特征点的位置和尺度等参数。特征检测的输出是关键点的集合，这些关键点指定了在相应尺度和方向下抽取的像素的位置。关键点和特征描述子是不一样的，特征描述子包含了关键点周围邻域像素的信息。因此，特征检测对于特征表示来说是个独立的过程。目前有大量的文献至少从计算机视觉的两个子领域对特征检测的位置和范围进行研究。其中一种研究方法是从图像配准发展起来的，目的是在图像有轻微的仿射变化和光照变化时仍然能够找到稳定的关键点，这些特征检测方法被称为兴趣点检测子。另一种特征检测方法是基于人类视觉注意计算模型发展起来的，这种方法关注于视觉显著的位置。此外，也有研究者建议可以通过金字塔结构、网格或随机抽样的方式获得关键点。

#### 1.1.1 基于图像配准的兴趣点检测子

虽然有许多变种方法，但是典型的兴趣点检测子都是基于图像的尺度空间表示的。尺度空间表示就是在多个分辨率上表示图像，通过多值的高斯核函数与图像卷积得到。读者想要详细地了解尺度空间表示，可以参看文献[12]。兴趣点检测子能够检测局部的具有判别力的特征点，例如角点、斑点区域或曲线。检测这些特征的过滤器响应位于 3 维的坐标系空间 $(x,y,s)$，$(x,y)$表示响应像素位置，$s$ 表示尺度。在$(x,y,s)$局部邻域内响应的极值被称为兴趣点。基于图像配准的兴趣点检测子大体都是围绕旋转不变性、尺度不变性和仿射不变性进行提取的。

兴趣点检测子最早能够追溯到 Moravec 角点检测子[9]。Moravec 角点检测子通过寻找最小强度变化的局部最大值来检测角点。Moravec 角点检测子的缺点是响应是各向异性、充满噪音的，并且对边比较敏感[10]。为了克服这些缺点，Harris[7] 角点检测子出现了。这个检测子是应用比较广泛的角点检测方法，具有很好的旋转不变性。此外，SUSAN 检测子[8]也是一种较流行的具有旋转不变性的检测子。然而，Harris 角点检测子并不能够对图像中存在的尺度变换进行处理，因此构建能够处理尺度问题的检测器是非常重要的。Lowe[23] 率先提取了尺度不变的局部特征 SIFT。SIFT 特征提取包含一个检测子和一个描述子。SIFT 检测子通过查找图像中高斯差分(DoG)的局部最大值来定位兴趣点。Mikolajczyk 和 Schmid[15] 开发了 Harris-Laplace 检测子，这个检测子结合了 Harris 角点检测子和能够进行尺度不变特征选择的 Laplace 函数。为了处理视角的变化，仿射不变的检测子被提出来了。Mikolajczyk 和 Schmid[15] 提出了 Harris (Hessian)仿射检测子，这个检测子合并了 Harris 角点检测子、尺度选择和基于二阶矩阵的椭圆估计。Tuytelaars 和 Van Gool 开发了一个基于边的区域检测子[13]，这种检测子采用曲线边和直线边来构建与 Harris 角点相关联的平行四边形。他们还提出了基于灰度的检测子[13]，这种检测子从灰度的局部极值开始，构建类似椭圆的区域，并且从这些极值发射大量的射线。基于边和灰度的方法都保留了仿射不变性。Matas[18] 提出了最大极值稳定区域(MSER)检测子，这种检测子通过分水岭过程来发现椭圆区域，具有仿射不变性。此外，显著区域检测子[19]也是一种仿射不变的检测子，其采用熵函数来定位区域，且提取的区域具有不规则的形状。Mikolajczyk[20] 对于各种仿射特征点提取算法进行了比较，读者可以参考。

#### 1.1.2 基于视觉注意计算模型的兴趣点检测子

许多仿生学思想也被应用到计算机视觉的兴趣点检测方法中，这些方法是基于视觉注意计算模型的。Frintrop[16] 详细讨论了视觉注意方法。Itti 和 Koch 在 2000 年提出了一个非常流行的模型[17]，这个模型是建立在 Koch 和 Ullman[21] 于 1985 年的研究基础上的。Itti 和 Koch 的显著性模型在一个高的层次上通过多个特征通道，例如颜色、方向和灰度等，寻找中央-周边模式的极值。中央-周边极值能够通过拉普拉斯-高斯(LoG)或者差分高斯(DoG)滤波得到，与 Lindeberg[22]、Lowe[14] 的模型是相似的。Bruce 和 Tsotsos 在 2009 年提出了用于显著性的信息论理论方法[80]，这种方法和早期的 Kadir 和 Brady 的尺度显著性兴趣点检测子是类似的。局部区域被看作是兴趣点并不只是因为它是基于一定模式或响应的，而是因为它和周边邻域的像素是显著不同的。在交叉学科领域仍然可以找到其他的视觉显著性计算模型[81-83]，这些交叉领域关注于哪些计算模型能够最好地预测人类的注意机制，从而能够定位兴趣点。研究显示仿生学领域的相关模型研究是对计算机视觉的有益补充。

#### 1.1.3 抽样问题

上面讨论了应该使用哪个兴趣点检测子的问题，但是还有一个更加基础的问题：是否需要使用兴趣点检测子的问题。研究人员把局部特征的抽取问题看成一个抽样问题。虽然兴趣点运算子对于图像配准是有用的，但其对于图像匹配和分类是否也是个理想的方式仍是一个需要讨论的问题。

Mar'ee 等人描述了一个以随机多尺度子窗口为抽取工具并且集成了随机决策树的图像分类算法[84]，虽然这个算法并不是一个严格的 BoF 方法，但是却说明了随机抽样的有效性。Nowak、Jurie 和 Triggs 在 2006 年探索了应用于 BoF 图像分类的抽样策略[85]，结果显示当使用足够多的抽样时，随机抽样的性能超过了兴趣点检测子。他们证明最重要的因素是图像中抽取的斑点的数量，因此认为随机的密抽样是最好的策略。空间金字塔匹配[66]使用的 SIFT 描述子是 8 像素间隔的密集网格抽样。在金字塔表示中，图像被逐层细化为多个子区域，每个子区域都形成自己的直方图，基层相当于是整个图像的标准的 BoF 表示。空间金字塔除了标准的无序的

BoF 表示外，还在一定程度上获得了位置空间信息。

### 1.2 特征点描述

为了表示检测到的点和区域，需要对特征点进行描述。特征描述就是对特征信息进行表述，要求描述后的特征能够对不同区域具有较好的区分性，并且对于光照和几何变换具有很好的不变性。目前，已经提出了大量的局部特征描述方法，大致可以分为基于分布的特征描述方法、基于滤波的特征描述方法、基于变换的描述方法和生物学启发的特征描述方法。此外的研究还包括如何在特征描述子中嵌入更多的空间信息，以及如何通过机器学习的方法使特征描述子的性能更好。

近年来，基于分布的特征描述方法受到广泛的关注和应用，理论和实验都证明这是一种非常有效的方法。SIFT 描述符子[23]是最有代表性的一种。SIFT 描述子将特征点周围的 16 × 16 窗口分割成 16 个子窗口，统计每个子窗口的 8 个方向梯度直方图，最终形成 128 维的特征向量。SIFT 描述子对于光照变化、背景遮挡、旋转和尺度变换具有很好的不变性，被认为是性能最好的描述符之一[11]。针对 SIFT 特征维数过高的问题，Ke 和 Sukthankar[24] 使用 PCA 归一化梯度窗口，简化了 SIFT 描述子，加快了匹配的速度，这种描述子称为 PCA-SIFT[24]。此外，HOG[25]、GLOH[11] 也被认为是 SIFT 描述子的扩展和改进，区分能力和鲁棒性都优于 SIFT 描述子。和 PCA-SIFT 类似，GLOH 同样使用 PCA 来降低描述子的维数。Lazebnik 等[86] 提出了旋转不变特征描述子(RIFT)，将归一化的圆形区域划分为同心环，每个环都和梯度方向直方图相关联。形状上下文描述子[26]统计边缘点位置和方向，形成直方图，广泛地应用于形状匹配和目标识别领域。

Freeman[29]等提出了方向可调滤波器，它线性结合了若干基滤波器，能够进行方向和尺度的选择。为了进行有效的多视角匹配，Baumberg[30] 和 Schaffalitzky 等[31] 使用复杂滤波器产生核函数来提取特征。方向可调滤波器与复杂滤波器的区别在于滤波器的不同形式。

LBP 指局部二值模式[32]，是一种基于变换的特征描述方法，具有光照不变性，广泛用于纹理描述，在人脸图像分析[33]等领域也取得了很好的效果。DCT 描述子是对特征点区域进行 DCT 变换，然后取前 $n$ 个 DCT 系数作为特征描述子。傅立叶方法对特征区域进行傅立叶变换，然后取低频部分作为特征描述符，该方法存在的一个问题是没有反映特征的空间信息。Marcelja[51] 和 Daugman[87,88] 通过一系列的 Gabor 函数来对哺乳动物的视觉响应进行建模，这些函数适合于皮层简单细胞感受野表示，因此 Gabor 变换适合于局部特征的描述。Wavelet 变换对于多分辨率分析也是很有效的，也能够表示局部特征。Gabor 和 Wavelet 变换能够反映空间信息，在一定程度上克服了缺少空间信息的缺点[89]。

Mikolajczyk 和 Schmid 的综述[11] 比较了多个特征描述子。然而，在评价的描述子中缺少颜色信息，这和仿生视觉团体形成鲜明的对比。相对于特征表示，仿生视觉团体通常会加入颜色信息。Jiang 等人在 2007 年证明在特征提取和描述中加入颜色信息能够改善 BoF 图像检索的性能[27]。van de Sande 等人详细比较了颜色特征描述子的性能，结果表明结合了颜色的描述子在图像分类性能上已经超过了 SIFT。在颜色描述子中，OpponentSIFT 在通常情况下都是非常有用的[28]。

Dongjin Song 在 2010 年提出了如何利用生物学角度的特征进行图像分类。文中作者采用了多种受生物学启发得到的特征[34]，并对这些特征进行降维，降维后的新特征既保留了空间特性又增加了区分能力，此外还构建了新的图像分类框架，实验证明了方法的有效性。

Tatsuya Harada 在文献[35]中提出了这样一个问题："给定的局部特征描述子，如何加入局部和全局的空间信息，并且获得表达图像更简洁、更具判别能力的特征?"。作者在文中提出了一个通用的框架，它能够在局部特征中嵌入全局的和局部的空间信息；另外还提出了简单有效的不同类型特征融合的方法。实验证明了这种方法的有效性。此外，最近的研究也已经针对给定的任务学习一个具有判别能力的特征而不是使用一个先验选取的特征。这方面的努力包括 Karlinsky 等人提出的无监督学习判别特征子集和检测参数的方法[36]，还包括 Winder 等人提出的特征描述子的模块分解及其如何优化性能的方法[37,38]。Winder 证明许多通用的特征描述子，例如 SIFT，都能够通过这种方法分解，并且可以通过学习使其性能更好。

## 2 特征的量化和编码

### 2.1 特征量化和编码的相关说明

图像特征的提取和表示是图像处理的基础，但是从图像中提取的局部特征数量巨大，不适合直接用于后续的处理。最近的工作[39,40]表明对于识别任务，字典的设计相对于后面的阶段(特征编码和特征汇集)来说重要性小一些，可以通过字典对局部特征进行表示的编码阶段来获得令人满意的性能。目前，存在大量的特征量化和编码策略。

在讨论各种特征编码技术之前，先给出符号的表示及含义。$X_{d\times N}=(x_1,x_2,\cdots,x_i,\cdots,x_N)$代表局部特征描述子的集合，其中 $x_i\in R^d$，$d$ 表示特征的维数，$N$ 表示局部特征的个数；$B_{d\times M}=(b_1,b_2,\cdots,b_j,\cdots,b_M)$代表视觉词字典，其中 $b_j\in R^d$，$d$ 表示特征的维数，$M$ 表示视觉词的个数；$U=(u_1,u_2,\cdots,u_N)$代表局部特征描述子编码的集合，$u_i$ 表示特征描述子 $x_i$ 的编码向量，$u_{ij}$ 表示视觉词 $b_j$ 对 $x_i$ 的编码。特征编码和特征汇集的过程可以用图 2 来表示。

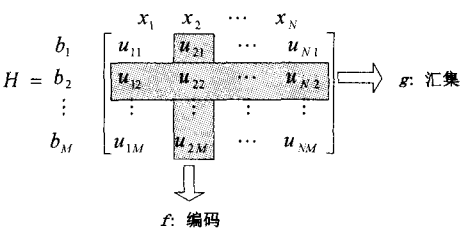


图 2 特征编码和汇集原理图

简单地说就是需要找到两个函数 $g$、$f$，函数 $f$ 负责对特征表述子进行编码，函数 $g$ 负责对视觉词进行空间汇集操作，来获得图像更紧凑、更鲁棒的表示。此部分将详细地讨论局部特征的量化和编码技术。论文第 3 部分将详细地讨论特征汇集技术。

## 2.2 典型的特征编码方法

### 2.2.1 经典的硬指派编码和软指派编码方法

原始的 BoF 方法[41]采用硬指派的方法对局部描述子进行编码。硬指派的过程是在生成视觉词典后进行局部特征的视觉单词映射时，根据局部特征与视觉单词所对应的特征向量的近邻程度来判定局部特征所对应的视觉单词，特征描述子被分配给最近的一个视觉词，被分配的视觉词对应的编码为 1，其余的视觉词的编码为 0，公式表示如下：

$$u_{ij}=\begin{cases}1, & \text{如果} \arg\min\limits_{j=1\cdots M}\| x_i-b\| \\ 0, & \text{否则}\end{cases} \tag{1}$$

硬指派编码方法存在一些局限性，一则是对字典的失真错误非常敏感[42,43]；二则是视觉单词不确定性和似然性导致的视觉词如何选择的问题。视觉单词的不确定性是指从两个或者多个相关的候选值中选择正确词汇的问题。硬指派方法仅仅选择了最好的表示单词，但却忽略了其它相关的候选值。视觉单词的似然性主要是指在视觉词典中没有合适的候选值时也选择了一个视觉单词，导致编码的不准确性。

软指派编码[44,45]是针对硬指派编码的缺点提出的一种改进方法。在软指派编码中一个特征描述子用多个视觉词来描述。软指派编码的优点是概念简单、计算有效，整个计算过程不需要优化，只需要计算局部特征和每个视觉词之间的距离。软量化的编码公式如下：

$$u_{ij}=\frac{\exp(-\beta\| x_i-b_j\| )_2^2}{\sum\limits_{k=1}^{M}\exp(-\beta\| x_i-b_k\| )_2^2} \tag{2}$$

### 2.2.2 改进的软指派编码方法

(1)稀疏编码

稀疏编码[46,47]作为一种硬指派量化的改进方法，已经显著地提高了硬指派编码问题的鲁棒性。稀疏编码可以看作基向量的稀疏子集的线性组合，并通过 l1 范式进行正则化的近似。

$$u_i=\arg\min_{u\in R^M}\| x_i-Bu\| _2^2+\lambda\| u\| _1,\lambda\in R \tag{3}$$

然而，这种方法在优化时计算代价太大，而且还会产生相似的描述子的编码并不一致的问题[48,49]。产生这个现象的原因是由于字典是过完备的，因此相似的描述子选择的视觉词可能是不同的，这就会导致在表示相似的描述子时产生很大的偏差。

(2)局部性约束编码

针对稀疏编码存在的问题，K. Yu[49,50]提出了更有效、更一致的编码方法，这种方法建立在局部性约束的基础上。作者提出的目标函数如下：

$$\begin{aligned}&u_i=\arg\min_{u\in R^M}\| x_i-Bu\| _2^2+\lambda\| d_i\odot u\| _2^2,\lambda\in R \\ &\text{s. t. } 1^{\mathrm{T}}u_i=1\end{aligned} \tag{4}$$

式中，$d_i=\exp(\frac{dist(x_i,B)}{\sigma})$，$dist(x_i,B)=[dist(x_i,b_1)\cdots dist(x_i,b_M)]^{\mathrm{T}}$，代表特征描述子 $x_i$ 和视觉词典中的视觉词的欧式距离，$\sigma$ 是控制局部性衰减速度的权重参数。此外，作者认为描述子都位于邻近描述子的低维的流形空间内，使用欧几里得距离时将描述子分配给邻近空间内的视觉词才是有意义的，因此在编码时应该选择局部的基才是合理的。局部性编码(LLC)[49]的编码过程可以通过下面的公式来近似：

$$\begin{aligned}&\min_U\sum_{i=1}^{N}\| x_i-B_iu_i\| ^2 \\ &\text{s. t. } 1^{\mathrm{T}}u_i=1\end{aligned} \tag{5}$$

能够加速的原因是根据局部性约束原理在对局部特征 $x_i$ 编码时不是采用所有的基 $B$，而是采用 $k$ 最近邻构成的局部基 $B_i$。

Xinggang Wang[78]提出了径向基编码方法，这种方法也是利用局部的近邻约束来找到要编码的视觉词，每个视觉词对应的编码如下：

$$\begin{aligned}&u_{ij}=\begin{cases}\frac{1}{\sqrt{2\pi}\sigma_j}\exp(-\| x_i-b_j\| ^2), & \text{如果 } b_j\in N_k(x_i) \\ 0, & \text{否则}\end{cases} \\ &\text{s. t. } \sum_{j=1}^{M}u_{ij}=1\end{aligned} \tag{6}$$

Lingqiao Liu[52]在传统的软分配量化编码方法的基础上加入了局部性约束，编码的公式如下：

$$u_{ij}=\frac{\exp(-\beta\hat{d}(x_i,b_j))}{\sum\limits_{k=1}^{M}\exp(-\beta\hat{d}(x_i,b_k)} \tag{7}$$

式中，$\hat{d}(x_i,b_j)=\begin{cases}d(x_i,b_j), & \text{如果 } b_j\in N_k(x_i) \\ \infty, & \text{否则}\end{cases}$，

$\hat{d}(x_i,b_j)$代表特征描述子 $x_i$ 和 $k$ 最近邻的视觉词 $b_j$ 之间的距离，并且将不是 $k$ 近邻的其他视觉词的距离设置为∞，这正是对传统的软量化的改进。

显著性编码[53]是另外一种可选的办法，在保留简单有效的基础上显示出非常好的性能。这种方法利用了局部性约束来代替传统的硬编码，描述子 $x_i$ 和最近的视觉词 $b_j$ 之间的显著性定义为：

$$u_{ij}=\phi(\frac{\| x_i-\hat{b}_j\| _2^2}{\frac{1}{k-1}\sum\limits_{m\neq j}^{k}\| x_i-\hat{b}_m\| _2^2}) \tag{8}$$

$\phi(\cdot)$是一个单调递减函数，$\{\hat{b}_m\}$，$m=1,\cdots,k$，是局部特征 $x_i$ 的 $k$ 个最近邻视觉词。相对于描述子的维数，当显著性编码应用于数量小的局部基的时候，显著性编码提升了局部约束编码的一致性。

(3)拉普拉斯稀疏编码

S. Gao 于 2010 年提出了另外一种改进稀疏编码一致性的方法[48]，这种方法在目标函数(3)里面增加了拉普拉斯矩阵来进行字典和编码的学习，目标函数表示为：

$$\begin{aligned}&\arg\min_{B,U}\| x_i-BU\| _2^2+\lambda\| u_i\| _1+\beta tr(ULU^{\mathrm{T}}),\lambda\in R \\ &\text{s. t. } \| b_j\| \leqslant 1,\forall j\in M\end{aligned} \tag{9}$$

式中，$L=A-W$ 是拉普拉斯矩阵；$W$ 是相似度矩阵，表示局部特征之间的关系；$A$ 为对角阵，$A_{mm}=\sum\limits_n W_{mn}$。

在数据集中存在太多的局部特征，同时构建拉普拉斯矩阵和学习稀疏编码是不可行的。在学习局部特征的时候有必要对局部特征进行限制，将其称为模板特征。但是仍然存在模板特征的 $k$ 近邻搜索中计算量还是太大的问题。

上面提到的编码方法除了拉普拉斯稀疏编码外，其余的都是对独立的局部特征进行编码。在拉普拉斯稀疏编码中计算局部特征的全局相似性来约束稀疏性。

#### 2.2.3 图像空间一致性约束编码

上面提到的编码方法都是利用在特征空间的局部性约束，但并没有考虑到图像物理空间的一致性约束。然而，在图像的空间域中，稠密的局部特征会共享一些上下文信息。这可以简单地看作计算两两局部特征之间的相似性，能够反映图像中一些区域的局部关联。在编码阶段失去这种上下文信息将会引起编码在空间上下文方面的编码不一致，并且当进行后续的空间特征汇集操作时得到的最终的图像特征是不可靠的。Shabou, A[54]除了吸收上面编码方法的优点之外，还加入了图像空间域的上下文信息，取得了更好的效果。文献[54]考虑到了局部空间信息能够改善编码的一致性，在文中提出了能量函数来选择更可靠的基，然后再对局部特征进行编码，下面是论文中提到的能量函数：

$$E(Y)=\underbrace{\sum_{p\in P} f_{data}(x_p,\hat{B}_p)}_{E_{data}}+\beta\underbrace{\sum_{p\sim q} f_{prior}(\hat{B}_p,\hat{B}_q)}_{E_{prior}} \tag{10}$$

能量函数包含两部分，Edata 部分代表特征表述子 xp 和选择的基之间的距离，Eprior 部分代表特征描述子 xp 的基和它的近邻所分配的基的距离和。这个模型的 Eprior 部分相当于实现了图像空间上的约束。实验结果证明了方法的有效性。

## 3 特征的汇集

现代的计算机视觉体系结构中通常都加入了特征的空间汇集阶段，这个阶段是对感兴趣的区域的特征描述子的联合分布进行统计。通常来说，汇集的目的就是将特征表示为一种新的更有用的特征，这种特征保留了重要信息，丢弃了不相关的细节；此外，还获得了对于位置变化和光照条件变化的不变性和遮挡的鲁棒性，得到特征的压缩表示。特征汇集的思想最初来源于 1962 年 Hubel 和 Wiesel 的关于视觉皮层细胞的开创性工作[55]，在 1999 年此项工作和 Koenderink 的局部无序图像概念结合起来[56]。局部邻域的特征汇集能够对输入的微小变换产生很好的不变性，已经用于大量的视觉识别的模型中。

### 3.1 汇集操作

#### 3.1.1 经典的汇集操作

Jarrett[60]已经证明了在小训练集的分类中，特征汇集方式比无监督的特征的预训练更重要，如果使用了合适的汇集方式，即使是随机的特征也能够得到好的分类结果。典型的汇集操作有求和汇集(sum pooling)、求平均汇集(average pooling)和求最大汇集(max pooling)。平均汇集操作可以用下面的公式表示：

$$f_a(u)=\frac{1}{N}\|u\|_1 \tag{11}$$

式中，$N$ 表示局部特征的个数，$u$ 表示某个视觉词在所有局部特征上的编码。如果在式(11)中只是求和，没有求平均，则为求和汇集。

最大值的汇集方法可以用下面的公式表示：

$$f_m(u)=\|u\|_\infty \tag{12}$$

Yang[65]在多个基准数据库上的实验结果已经显示，使用特征的最大汇集操作其结果要比平均汇集操作的更好。

最近的识别系统中使用汇集技术来计算局部的或全局的特征包的应用很多。采用特征汇集的仿生图像识别模型包括新认知机[57]、采用平均特征汇集的卷积网络[58]和最大特征汇集的卷积网络[59,60]、采用最大汇集的 HMAX[61]、采用平均汇集的视觉皮层区域 V1 模型[62]。许多流行的特征提取模型也使用了汇集技术，例如 SIFT、HOG 及这些方法的变种。在这些方法中，主梯度方向是在若干个区域中进行计算，在邻近区域进行汇集，形成方向的局部直方图。这种计算是通过字典量化特征描述子来计算局部或全局区域内视觉词出现的次数[63-66]，相当于平均汇集方法。

#### 3.1.2 改进的汇集操作

文献[67]已经说明对硬指派量化的特征采用最大汇集操作和线性分类的效果与 Lazebnik[66]采用直方图交核函数分类的效果是等同的，然而并不清楚为什么最大汇集操作能够取得那么好的结果。Y-Lan Boureau[68]填补了这个空白，从理论上分析了汇集操作；在分类环境下比较了不同的汇集操作，并且检测了相应的统计行为是如何转化到后续分类中的。这篇论文还通过详尽的实验研究了不同汇集操作的区分能力以及影响汇集操作性能的不同因素，包括特征的平滑性和稀疏性，还将几种流行的汇集操作统一为一种连续形式。

最近的理论分析显示平均和最大汇集操作都不是最优的，这两种汇集操作的严格限制来自于对特征的分布进行了过于简单的假设，导致在统计过程中空间信息的不可逆转的丢失。Jiashi Feng 于 2011 年提出了空间 p 范式的汇集方法(GLP)[69]。此汇集函数直接从最大化类间区分能力中学习得到，具有很好的区分能力；汇集函数完全和类特定的每个视觉词的空间模式相对应，因此利用视觉词的空间分布达到了令人满意的程度。这种方法对局部特征之间的关联进行建模，并且对特征分布进行了更合理的假设；此外，还将最大汇集操作和平均汇集操作以一种更合理的框架进行了统一，大大提高了结果特征的判别能力和分类性能。

Lingqiao Liu 指出最大汇集操作的问题在于只估计了视觉词在图像中出现的概率，但是忽略了视觉词出现的频率。针对这个问题，Lingqiao Liu 于 2011 年提出了一种新的 mix-order 最大汇集操作方法[52]。在这种方法中，考虑了每个视觉词在图像中出现的次数，一个视觉词出现次数大于 $k$ 的概率可以通过这个视觉词在图像中局部特征描述子出现概率的最小值的 $k$ 次方来近似。最大汇集操作是 mix-order 最大汇集操作方法中 $k=1$ 的特殊情况。

2011 年 Jimei Yang[70]指出现实生活中绝大多数图像中包含杂乱的背景，这会导致在对图像进行汇集操作时得到的最强响应可能来自于背景，从而导致图像的最终表示是不可靠的。文中针对这个问题，提出了一种新的汇集函数，这个函数在传统的最大汇集函数计算中加入了基于显著性的权重，从而使最终图像级的特征更可靠。

Avila, S. 也不是采用经典的求和汇集和最大汇集策略，而是通过估计特征分布的概率密度分布进行汇集。Avila, S.[71]在 2011 年提出新的汇集函数，此汇集函数是建立在对每个视觉词对应的局部特征的统计分析的基础上的，能够更好地表示编码和描述子之间的联系，这样得到的图像特征称为 BOSSA (Bag Of Statistical Sampling Analysis)，特征的优点是维数合理，且较 BoF 保留了更多的信息。

现有的汇集方法都是将一个 $N\times M$ 的编码矩阵汇集后得到 $M$ 维的向量，这里 $N$ 指特征的个数，$M$ 是字典中视觉词的个数。这样变换后的后果是信息剧烈的下降。Xinnan Yu 针对这种现象，在 2011 年提出了一种新的特征汇集方法[72]，其是基于 2 维直方图表示的，这样对于编码的图像可以保存更多的信息，并且可以很容易地融合到现有的计算机系统框架中。

### 3.2 汇集区域的选择

特征汇集操作的选择是指在不同的图像空间内采用某种汇集操作来对特征的编码进行区域上的汇集。空间金字塔(SPM)[66]就是最经典的一种空间汇集方法。空间金字塔不断地将图像区域进行细分，在金字塔的子区域单元上进行特征的汇集操作而不是在整幅图像上进行特征的汇集，因此融入了更多的空间信息，从而大大提高了性能。空间金字塔的成功也说明了对邻域进行空间汇集操作的重要性。

Harada，T.[73]指出 SPM 在图像的每层金字塔划分时都是基于手工的且固定的划分，划分策略缺少理论依据。Yang Cao[74]从这种角度出发研究了一种新类型的 BoF 模型，以便包含图像中更多的空间信息。和 SPM 划分方法不同的是，图像上的特征首先将特征投影到某些直线或圆上来产生一系列的有序的 BoF 特征族，使用这两种投影策略的原因是线和圆可以看成物体的基本组成元素。然后作者采用类似 boosting 的方法进行特征选择，得到最具有代表性的特征向量。有序的 BoF 特征具有和传统的 BoF 相同的形式，可以看作是 SPM 空间思想的泛化，不但包含了更多的空间信息，而且具有更好的平移、旋转和尺度不变性。Mateusz Malinowski[79]进一步指出了 SPM 空间划分的缺点：区域的划分独立于具体的任务，没有考虑数据的特点；另外，划分策略属于对区域的硬划分，空间上近邻的同质的区域可能会分到不同的子区域。作者针对这种情况提出了联合优化分类器参数和汇集区域的方法，并且为了避免高维特征空间的优化问题和过拟合现象，提出了近似算法，一个近似是通过预汇集操作降低编码的维数，另一个近似是将编码划分为子集以加快优化的速度。Yangqing Jia[75]也研究了图像汇集区域的选择对图像分类的影响，结果显示对汇集区域进行自适应的学习能够大幅地提高系统的性能，即使在编码阶段使用小数量的字典。为了学习优化的学习参数，文中采用了过完备的思想，开始的时候采取了大量的汇集区域作为候选，然后训练结构稀疏的分类器，只使用全部特征的稀疏子集。此外，还提出了一个有效的基于递增特征选择和再训练的算法进行快速的学习，此方法虽然使用更低维的特征空间却取得了更好的性能。

**结束语** BoF 特征作为计算机视觉领域应用最广泛的一类特征，其应用越来越广泛。大量研究结果也表明 BoF 特征在计算机视觉中具有很好的性能。BoF 模型通常包括局部特征提取、字典的生成、特征的量化和编码、特征的汇集。最近的工作[39,40]显示对于识别任务，视觉字典的设计相对于后面的阶段(特征编码和特征汇集)来说重要性小一些，所以本文着重对局部特征提取、特征的量化和编码、特征的汇集技术进行了讨论。尽管已有的研究成果已经证明 BoF 特征在图像分类和基于内容的图像检索方面的性能非常好，但是在某些应用上仍然存在局限性和挑战。存在的问题和挑战如下。

(1)空间信息。空间信息的丢失是 BoF 特征的一个重要弱点。在处理空间任务时，比如在杂乱的背景中定位小的对象或识别对象之间的关系时，由于空间信息的丢失导致 BoF 的效果不佳。如何在保持 BoF 模型的优点的同时嵌入更多的空间信息，仍然是研究的重点。

(2)语义信息。尽管一些视觉词表示对象的不同部分，但是在实际中，大部分视觉词仍然没有明确的语义。在 BoF 表示中缺少内在的语义信息，使得在处理某些任务如在需要关键字或自然语言描述的图像检索任务时存在挑战。2009 年的 ImageNet 数据库[76]是一个缩小视觉词和自然语言词之间鸿沟的工具，但是如何研制减小语义鸿沟的其他技术仍然是值得研究的问题。

(3)特征不变性。文献[77]已经对特征不变性这个重要属性进行了评价。然而随着特征的不变性的增加，特征的区分能力在下降，并且特征不变性还会增加计算的复杂性和特征的再现性。因此一个基本的准则是如果应用中确实不太需要不变性，那么尽量不用不变性。在通常情况下，应该更多地考虑特征检测和描述的鲁棒性，而不是不变性。好的特征的判断标准依赖于应用，没有一个通用的标准。所以在设计好的局部特征时，要从实际应用出发。

(4)特征量化和编码。特征的量化和编码经历了硬量化、软量化、稀疏编码、显著性约束、局部性约束、相似性约束等。在编码时这些约束都体现在目标函数的正则化上。新的编码目标函数的优化目标仍然是相似的特征应该具有相似的编码形式，且得到的编码应该具有更多的空间一致性。此外，好的编码算法应该具有小的时间复杂度和重构错误率。

(5)特征汇集。在特征汇集过程中使用的经典方法是最大值和平均值等统计方法。然而研究表明这些都不是最优的。在统计的过程中应该更多地考虑到数据的空间分布和数据之间的联系，充分考虑到不同类的视觉词的空间分布不同以及特征之间的分布不独立等事实。此外，不同类型的特征也会影响特征汇集的函数及参数的设置。

(6)算法的比较平台。许多研究人员在标准数据集上已经取得了较好的分类效果，但是由于特征计算上的差异、学习算法上的差别以及在分类流程上的各个阶段的调节不同，这些方法在描述上缺少细节，很难直接地对这些方法进行重现和比较。所以有必要设计一个统一的平台来对算法条件进行统一的设置，从而保证比较是公平的。

(7)图像的 BoF 特征是针对局部不变特征的有效利用而提出的。但是，BoF 本质上仍是一种图像的全局特征，这类特征在图像内容比较简单的图像分类问题上往往能获得较好的效果，但对于内容更丰富、背景更复杂的图像分类问题则难以获得很好的效果，所以能否结合除局部特征外的其他特征或其他的统计方法来对图像进行综合描述以取得更好的效果是个值得研究的问题。

## 参 考 文 献

[1] Csurka G, Dance C R, Fan Li-xin, et al. Visual categorization with bags of keypoints[C]// Proceedings of European Conference Computer Vision 2004, workshop on Statistical Learning in Computer Vision, 2004. Prague, Czech Republic: Springer-

Verlag LNCS, 2004: 59-74

[2] MacQueen J B. Some Methods for classification and Analysis of Multivariate Observations[C]// Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967. Berkeley, University of California Press, 1967, 1: 281-297

[3] Arai K, Barakbah A R. Hierarchical K-means: an algorithm for centroids initialization for K-means[J]. Reports of the Faculty of Science and Engineering, 2007, 36(1): 25-31

[4] McLachlan G J, Basford K E. Mixture Models: Inference and Applications to Clustering [M]. New York: Marcel Dekker, 1988

[5] Comaniciu D, Meer P. Mean Shift: A Robust Approach toward Feature Space Analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(5): 603-619

[6] Beaudet P R. Rotationally invariant image operators[C]// Proceedings of the 4th International Joint Conference on Pattern Recognition, 1978. Kyoto, Japan: Institute of Electrical and Electronics Engineers Inc., 1978: 579-583

[7] Hams C, Stephens M. A combined corner and edge detector[C]// Proceedings of Alvey Vision Conference, 1988. University of Manchester, 1988: 147-151

[8] Smith S M, Brady J M. SUSA N: A new approach to low level image processing[J]. International Journal of Computer Vision, 1997, 23(1): 45-78

[9] Moravec H. Towards automatic visual obstacle avoidance[C]// Proceedings of the International Joint Conference on Artificial Intelligence, 1977. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology, 1977: 584

[10] Johnson A, Hebert M. Object recognition by matching oriented points[C]// Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1997. San Juan, Puerto Rico: IEEE Computer Society, 1997: 684-689

[11] Mikolajczyk K, Schmid C. A Performance Evaluation of Local Descriptors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615-1630

[12] Lindeberg T. Feature Detection with Automatic Scale Selection [J]. International Journal of Computer Vision, 1998, 30(2): 79-116

[13] Tuytelaars T, Van Gool L. Matching widely separated views based on affine invariant regions[J]. International Journal of Computer Vision, 2004, 59(1): 61-85

[14] Lowe D G. Object recognition from local scale invariant features [C]// Proceedings of the 7th International Conference on Computer Vision, 1999. Kerkyra, Greece: IEEE Computer Society, 1999: 1150-1157

[15] Mikolajczyk K, Schmid C. Scale&-Affine Invariant Interest Point Detectors[J]. International Journal of Computer Vision, 2004, 60(1): 63-86

[16] Frintrop S, Rome E, Christensen H I. Computational visual attention systems and their cognitive foundations: A survey[J]. ACM Transactions on Applied Perception (TAP), 2010, 7(1): 1-39

[17] Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention[J]. Vision Research, 2000, 40 (10-12): 1489-1506

[18] Matas J, Chum O, Urban M, et al. Robust Wide Baseline Stereo From Maximally Stable Extremal Regions[C]// Proceedings of British Machine Vision Conference, 2002. British: the British Machine Vision Association, 2002: 384-393

[19] Kadir T, Zisserman A, Brady M. An Affine Invariant Salient Region Detector[C]// Proceedings of European Conference on Computer Vision, 2004. LNCS, 2004, 302 1: 228-241

[20] Mikolajczyk K, Tuytelaars T, Schmid C, et al. A Comparison of Affine Region Detectors[J]. International Journal of Computer Vision, 2005, 65(1/2): 43-72

[21] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry[J]. Human Neurobiology, 1985, 4(4): 219-27

[22] Lindeberg T. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention[J]. International Journal of Computer Vision, 1993, 11(3): 283-318

[23] Lowe D. Distinctive Image Features from Scale-invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60 (2): 91-110

[24] Ke Y, Sukthankar R. Pca-sift: A More Distinctive Representation for Local Image Descriptors[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2004. Washington, DC: IEEE Computer Society, 2004: 506-513

[25] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005. San Diego, CA, USA: IEEE Computer Society, 2005: 886-893

[26] Belongie S, Malik J, Puzicha J. shape matchjng and object recognition using shape contexts[J]. IEEE Tran8actions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509-522

[27] Jiang Y G, Ngo C W, Yang J. Towards optimal bag-of-features for object categorization and semantic video retrieval[C]// Proceedings of ACM Conference on Image and Video Retrieval, 2007. New York, NY, USA: ACM, 2007: 494-501

[28] van de Sande K E A, Gevers T, Snoek C G M. Evaluating color descriptors for object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32 (9): 1582-1596

[29] Freeman W T, Adelson E H. The Design and Use of Steerable Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(9): 891-906

[30] Baumberg A. Reliable feature matching across widely separated views[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2000. Hilton Head, SC, USA: IEEE Computer Society, 2000: 774-781

[31] Schaffalitzky F, Zisserman A. Multi-view Matching for Unordered Image Sets[C]// Proceedings of 4th European Conference on Computer Vision, 2002. Copenhagen, Denmark: Springer, 2002: 414-431

[32] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987

[33] Hadid A. Face Description with Local Binary Patterns: Applica-

tion to Face Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12): 2037-2041

[34] Song Dong-jin, Tao Da-cheng. Biologically Inspired Feature Manifold for Scene Classification[J]. IEEE Transactions on Image Processing, 2010, 19 (1): 174-184

[35] Harada T, Nakayama H, Kuniyoshi Y. Improving Local Descriptors by Embendding Global and Local Spatial Information[C]// Proceedings of European Conference on Computer Vision, 2010. Heraklion, Crete, Greece, 2010: 736-749

[36] Karlinsky L, Dinerstein M, Ullman S. Unsupervised Feature Optimization (UFO): Simultaneous Selection of Multiple Features with Their Detection Parameters[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 1263-1270

[37] Winder S, Brown M. Learning Local Image Descriptors[C]// Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2007. Minneapolis, Minnesota, USA: IEEE Computer Society, 2007: 1-8

[38] Winder S, Hua G, Brown M. Picking the best DAISY[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 178-185

[39] Coates A, Ng A Y. The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization[C]// Proceedings of the 28th International Conference on Machine Learning 2011. Bellevue, WA, USA, 2011

[40] Rigamonti R, Brown M A, Lepetit V. Are Sparse Rrepresentations Really Relevant for Image Classification? [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 1545-1552

[41] Sivic J, Zisserman A. Video google: A Text Retrieval Approach to Object Matching in Videos[C]// Proceedings of IEEE International Conference on Computer Vision, 2003. Nice, France: IEEE Computer Society, 2003: 1470-1477

[42] Lazebnik S, Raginsky M. Supervised Learning of Quantizer Codebooks by Information Loss Minimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31 (7): 1294-1309

[43] Mairal J, Bach F, Ponce J, et al. Discriminative learned dictionaries for local image analysis[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008. Anchorage, Alaska, USA: IEEE Computer Society, 2008: 1-8

[44] Gemert J C V, Geusebroek J M, Veenman C J, et al. Kernel codebooks for scene categorization[C]// Proceedings of European Conference on Computer Vision, 2008. Marseille, France: Springer, 2008: 696-709

[45] van Gemert J C, Veenman C J, Smeulders A W M, et al. Visual Word Ambiguity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32 (7): 1271-1283

[46] Lee H, Battle A, Raina R, et al. Efficient Ssparse Coding Algorithms[C] // Proceedings of Advances in Neural Information Processing System, 2007. Vancouver, B. C. , Canada: Springer, 2007

[47] Yang J, Yu K, Gong Y, et al. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 1794-1801

[48] Gao S, Tsang I, Chia L, et al. Local Features Are Not Lonely-Laplacian Sparse Coding for Image Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. San Francisco, CA, USA: IEEE Computer Society, 2010: 3555-3561

[49] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. San Francisco, CA, USA: IEEE Computer Society, 2010: 3360-3367

[50] Yu K, Zhang T, Gong Y. Nonlinear Learning Using Local Coordinate Coding[C]// Proceedings of Advances in Neural Information Processing System, 2009. Vancouver, British Columbia, Canada: Springer, 2009

[51] Marcelja S. Mathematical description of the responses of simple cortical cells[J]. Journal of the Optical Society of America, 1980, 70(11): 1297-1300

[52] Liu Ling-qiao, Wang Lei, Liu Xin-wang. In Defense of Soft-assignment Coding[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 2486-2493

[53] Huang Y, Huang K, Yu Y, et al. Salient Coding for Image Classification[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 1753-1760

[54] Shabou A, LeBorgne H. Locality-constrained and Spatially Regularized Coding for Scene Categorization[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. Providence, RI, USA: IEEE Computer Society, 2012: 3618-3625

[55] Hubel D H, Wiesel T N. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Vsual Cortex[J]. The Journal of Physiology, 1962, 160: 106-54

[56] Koenderink J J, Van Doorn A J. The structure of locally orderless images[J]. International Journal of Computer Vision, 1999, 31(2/3): 159-168

[57] Fukushima K, Miyake S. Neocognitron: A New Algorithm for Pattern Recognition Tolerant of Deformations and Shifts in Position[J]. Pattern Recognition, 1982, 15(6): 455-469

[58] LeCun Y, Boser B, Denker J S, et al. Handwritten digit recognition with a back-propagation network[C]// Proceedings of Conference on Neural Information Processing, 1989. Morgan Kaufmann, 1990: 396-404

[59] Ranzato M, Boureau Y, LeCun Y. Sparse feature learning for deep belief networks[C]// Proceedings of Conference on Neural Information Processing, 2007. Vancouver, B. C. , Canada: Springer, 2007

[60] Jarrett K, Kavukcuoglu K, Ranzato M, et al. What is the Best Multi-stage Architecture for Object Rcognition? [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 2146-2153

[61] Serre T, Wolf L, Poggio T. Object recognition with features inspired by visual cortex[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005. San Diego, CA, USA: IEEE Computer Society, 2005: 994-1000

[62] Pinto N, Cox D, DiCarlo J. Why is real-world visual object recognition hard[J]. PLoS Computational Biology, 2008, 4(1): 151-156,

[63] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[C]// Proceedings of IEEE International Conferenceon Computer Vision, 2003. IEEE Computer Society, 2003: 1470-1477

[64] Zhang J, Marszalek M, Lazebnik S, et al. Local features and kernels for classifcation of texture and object categories: An in-depth study[J]. International Journal of Computer Vision, 2007, 73(2): 213-238

[65] Yang J, Yu K, Gong Y, et al. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 1794-1801

[66] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2006. New York, NY, USA: IEEE Computer Society, 2006: 2169-2178

[67] Boureau Y, Bach F, LeCun Y, et al. Learning mid-level features for recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. San Francisco, CA, USA: IEEE Computer Society, 2010: 2559-2566

[68] Boureau Y, Ponce J, LeCun Y. A theoretical analysis of feature pooling in vision algorithms[C]// Proceedings of International Conference on Machine Learning, 2010. Haifa, Israel: Omnipress, 2010

[69] Feng Jia-shi, Ni Bing-bing, Tian Qi, et al. Geometric p-norm Feature Pooling for Image Classification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 2609-2704

[70] Yang Ji-mei, Yang M-H. Learning Hierarchical Image Representation with Sparsity, Saliency and Locality[C]// British Machine Vision Conference, 2011. British: BMVA Press, 2011: 19. 1-19. 11

[71] Avila S, Thome N, Cord M, et al. Bossa: Extended Bow Formalism for Image Classification[C]// Proceedings of International Conference on Image Processing, 2011. Brussels, Belgium: IEEE Computer Society, 2011: 2909-2912

[72] Yu Xin-nan, Zhang Yu-jin. A 2-D Histogram Representation of Images for Pooling[C]// SPIE. 2011

[73] Harada T, Ushiku Y, Yamashita Y, et al. Discriminative Spatial Pyramid[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 1617-1624

[74] Cao Yang, Wang Chang-hu, Li Zhi-wei, et al. Spatial-Bag-of-Features[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010. San Francisco, CA, USA: IEEE Computer Society, 2010: 3352-3359

[75] Jia Yang-qing, Huang Chang, Darrell T. Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2012. Providence, RI, USA: IEEE Computer Society, 2012: 3370-3377

[76] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. Miami, Florida, USA: IEEE Computer Society, 2009: 248-255

[77] Schmid C, Mohr R, Bauckhage C. Evaluation of interest point detectors[J]. International Journal of Computer Vision, 2000, 37(2): 151-172

[78] Wang Xing-gang, Bai Xiang, Liu Wen-yu, et al. Feature context for image classification and object detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011. Colorado Springs, CO, USA: IEEE Computer Society, 2011: 961-968

[79] Malinowski M, Fritz M. Learnable Pooling Regions for Image Classification[OL]. http://arxiv.org/abs/1301.3516

[80] Bruce N D B, Tsotsos J K. Saliency, attention, and visual search: An information theoretic approach[J]. Journal of Vision, 2009, 9(3): 1-24

[81] Elazary L, Itti L. Interesting objects are visually salient [J]. Journal of Vision, 2008, 8(3): 1-15

[82] Kienzle W, Franz M O, Schölkopf B, et al. Center-surround patterns emerge as optimal predictors for human saccade targets [J]. Journal of Vision, 2009, 9(5): 1-15

[83] Tatler B W, Baddeley R J, Gilchrist I D. Visual correlates of fixation selection: Effects of scale and time[J]. Vision Research, 2005, 45(5): 643-659

[84] Maree R, Geurts P, Piater J, et al. Raet alndom subwindows for robust image classification[C]// Proceedings of IEEE International Conferenceon Computer Vision, 2005. San Diego, CA, USA: IEEE Computer Society, 2005: 34-40

[85] Nowak E, Jurie F, Triggs B. Sampling strategies for bag-of-features image classification[C]// Proceedings of European Conferenceon Computer Vision, 2006. Graz, Austria: Springer, 2006, 3954: 490-503

[86] Lazebnik S, Schmid C, Ponce J. A sparse texture representation using local affine regions[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(8): 1265-1278

[87] Daugman J G. Two-dimensional spectral analysis of cortical receptive field profile[J]. Vision Research, 1980, 20(10): 847-856

[88] Daugman J G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters[J]. Journal of the Optical Society of America A, 1985, 2(7): 1160-1169

[89] Hui Bin, Tang Xu-sheng, Luo Hai-bo, et al. SDF Matched Filter Based on Gabor Wavelet Transform for Face Recognition[J]. Information and Control, 2008, 37(5): 633-636