基于多尺度稀疏表示的场景分类*

段 菲，章毓晋
（清华大学 电子工程系，北京 100084）

摘 要：提出一种基于多尺度稀疏表示的场景分类框架。首先从图像中提取多个尺度的局部特征，然后利用稀疏编码为每个尺度的特征单独学习相应的过完备字典；在图像表示阶段，为图像各尺度上的局部特征依据与其对应尺度的字典进行编码，并按照空间金字塔表示方法和特征各维最大汇总（max pooling）对各尺度上的特征编码分别汇总；最后将不同尺度上汇总的特征串接，形成对图像最终描述的全局向量。在三个常用标准场景库上的分类结果表明，提出的算法由于利用了不同尺度特征间的互补关系，与采用单尺度特征的方法相比，性能有了显著提升。

关键词：稀疏表示；多尺度；场景分类；空间金字塔表示
中图分类号：TP391.4 **文献标志码**：A **文章编号**：1001-3695(2012)10-3938-04
doi:10.3969/j.issn.1001-3695.2012.10.090

Multi-scale sparse representation based scene categorization

DUAN Fei, ZHANG Yu-jin
(*Dept. of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: This paper proposed a novel multi-scale sparse representation based framework of scene categorization. Firstly, it extracted local features of multiple scales, then learned dictionaries of each scale by using sparse coding algorithm. In the stage of image representation, this paper proposed to encode each local feature according to dictionary on the same scale. By using spatial pyramid representation and feature max-pooling, it summarized the features on each scale. In the final step, the proposed algorithm concatenated the pooled features of different scales to form the holistic feature. The experimental results on three popular benchmark datasets show that since the proposed method utilizes the complementary relationship of features on different scales, it obtains better performance compared with methods using features on single scale.

Key words: sparse representation; multi-scale; scene categorization; spatial pyramid representation

0 引言

近年来，词袋模型[1]（bag-of-word, BoW）在场景分类中获得了广泛应用。该模型由局部特征提取、字典学习、特征编码（或称向量量化）、特征汇总四个环节构成。首先从图像中抽取若干局部特征，然后将这些局部特征进行聚类（通常为 K-means）而形成视觉字典；在表示图像时，则依据最近邻匹配准则建立图像局部特征到视觉词汇表中词条的映射，也称做向量量化（vector quantization, VQ）；最后统计出整幅图像中所有视觉词汇的出现频率直方图（通常称之为特征平均汇总（average pooling）），作为最终的图像表示。

可以看出，传统的词袋模型并未对图像中局部特征之间的位置关系予以考虑，这显然会制约其描述能力。针对该问题，Lazebnik 等人[2] 提出空间金字塔表示模型（spatial pyramid matching, SPM），在特征汇总阶段对传统词袋模型进行了扩展，极大地提升了词袋模型的性能。

除缺乏对空间信息的刻画外，在传统词袋模型中，其最为关键的两个环节——字典学习和特征编码也存在一些不足。在使用聚类算法学习字典时，首先面临的问题是确定一个合适的字典长度，过大或过小均会对字典的描述能力带来负面影响。而在对局部特征编码时，传统的词袋模型使用了基于最近邻匹配的向量量化，当数据中带有随机噪声时，不可避免地会带来较大的量化误差。为此，Gemert 等人[3] 提出了软量化（soft assignment）的概念，利用从字典中找到的 K-近邻词汇的自适应加权线性组合来表示待量化的特征。与向量量化相比，该方法能显著减少量化误差，但 K 的选择对量化结果具有较大影响。为使 K 的选取具有自适应性及提高字典学习的质量，Yang 等人[4] 利用稀疏编码[5] 模型中涉及的两个子问题取代了传统词袋模型中的特征编码和字典学习模块，既解决了特征编码的自适应问题，又减少了字典长度对分类性能的敏感度。

虽然稀疏编码和 SPM 的结合将词袋模型的性能提升至一个全新的水平，但该框架仍然有提升的空间。空间金字塔表示方法[2] 按照多尺度特征汇总的方式融合了局部特征的空间布局信息，但这只是在词袋模型的后期阶段实施的。图像的视觉信息往往分布在多个尺度中，而且场景中有的对象结构尺度较大，如果在最初的特征提取阶段从每个采样点的多尺度邻域中提取特征，则与传统的只从单一尺度（通常为 16×16 像素）图像块中提取特征的方法相比，有可能增加局部特征所携带的上下文信息，从而进一步提升词袋模型的描述能力。

从上述考虑出发，本文对基于稀疏表示的图像分类[4] 框

收稿日期：2012-03-11；**修回日期**：2012-04-17 **基金项目**：国家自然科学基金资助项目（61171118）

作者简介：段菲（1979-），男，山西人，博士研究生，主要研究方向为图像处理、机器学习（duanf@163.com）；章毓晋（1954-），男，教授，博士，主要研究方向为其积极倡导的图象工程（图像处理、图像分析、图像理解及其技术应用）和相关学科.

架进行了扩展,提出了一个基于多尺度稀疏表示的场景图像表示方法。在最初的特征提取阶段,本算法保持图像尺寸不变,按照稠密采样的方式从图像中提取多尺度特征。然后利用稀疏编码模型从训练集中为每个尺度的局部特征求取一组过完备基,即为每个尺度的局部特征单独建立一个字典。在图像表示阶段,各自独立地为图像中各尺度上的局部特征依据其对应尺度的字典进行编码,并按照空间金字塔表示方法和文献[4]提出的特征各维最大汇总(max pooling)对各尺度上的特征分别汇总,最后将不同尺度上汇总的特征串接,作为对图像最终描述的全局向量。为了对本文算法的有效性进行验证,在三个常用标准场景库上进行了分类实验。结果表明,由于利用了不同尺度的特征间的互补性,本文所提出的多尺度稀疏表示方法能够显著提升单尺度稀疏表示方法的性能。

1　相关工作

1.1　稀疏表示

稀疏表示[5]的基本假设是,自然信号可用一组过完备基中的少量基向量的线性组合来良好地逼近。可形式化地用下列模型来追求该目标:

$$\min_{B,S} \| X - BS \|_F^2 + \lambda \| S \|_1 \quad \text{s.t.} \quad \| B_{\cdot j} \|_2 \leq 1, \forall j \tag{1}$$

其中:$X = \{x_1, x_2, \cdots, x_N\} \in \mathbb{R}^{d \times N}$为训练样本;$B \in \mathbb{R}^{d \times K}$为所要学习的过完备基;$S = [s_1, s_2, \cdots, s_N] \in \mathbb{R}^{K \times N}$为依据 B 所得的稀疏编码;$B_{\cdot j}$表示 B 的第 j 列。虽然式(1)对于 B 和 S 来说是非凸的,但当固定其中任意一个时,该模型便成为一个凸优化问题。因此,可通过交替优化的方式来求解该模型,即将式(1)分解为如下的两个子问题:

a)字典学习。当固定 S 时,式(1)便转换为如下的待约束的最小二乘问题:

$$\min_{B} \| X - BS \|_F^2 + \lambda \| S \|_1 \quad \text{s.t.} \quad \| B_{\cdot j} \|_2 \leq 1, \forall j \tag{2}$$

该问题可通过梯度投影或将其变换至对偶空间来求解。

b)特征编码。当固定 B 时,式(1)成为如下的无约束最小二乘问题:

$$\min_{S} \| X - BS \|_F^2 + \lambda \| S \|_1 \tag{3}$$

该问题可通过正交匹配寻踪(orthogonal matching pursuit, OMP[6])算法或特征符号搜索算法(feature-sign search)[7]来求解。

1.2　稀疏表示框架下的多尺度字典学习

现有的多尺度字典学习算法可大致分为基于解析的方法和基于学习的方法两类。由于后者通常具有良好的数据自适应性,与前者相比具有明显的优势,因此本节将重点关注基于学习的方法。现有的多尺度字典的应用领域主要为图像复原,文献[8,9]均利用四叉树和 K-SVD[10]算法来将不同尺度的特征融入一个全局字典中。Qin 等人[11]为每个类别及每个尺度的特征学习一个字典,通过组合的方式形成各尺度的全局字典,但其采用的字典学习算法和特征编码方法为传统的 K-means 和向量量化。因此,该方法为了获得好的效果,需要将字典规模取为一个较大的值,之后再依据分类效果进行消除冗余视觉词汇的处理。

2　本文算法

如前所述,文献[4]虽然通过空间金字塔和特征各维最大汇总在图像表示中融入了局部特征之间的空间位置关系,但这只是在词袋模型的最后阶段进行的。考虑到图像的视觉信息往往分布在多个尺度中,而场景中有的对象结构尺度较大,有的对象结构尺度较小,如果在最初的特征提取阶段从每个采样点的多尺度邻域中提取特征,则与传统的只从单一尺度(通常为 16×16 像素)图像块中提取特征的方法相比,有可能增加局部特征所携带的上下文信息,从而进一步提升词袋模型的描述能力。

本文以图 1 为例来说明这一点。图 1 中的两幅图像分别来自 inside city 类和 forest 类。如果按照固定尺度(白色方框大小)从二者中抽取局部特征,则在图示位置中很难区分白色方框对应的图像块来自哪个类别,但如果扩大特征提取的尺度(如黑色方框所示),则由于局部特征中融入了更多的上下文信息,局部特征自身的鉴别性得到了增强,因此通过将不同尺度的局部特征融入字典学习过程,则有希望增强字典的区分能力。

图 1　多尺度特征的作用示例

2.1　多尺度过完备字典学习

本文提出的多尺度字典学习算法流程如图 2 所示。在训练阶段,首先依据等间隔稠密采样从训练图像集中提取多个尺度上的特征(128-D SIFT[12]描述符),并经随机抽样形成多尺度特征集 $X = [X_1, \cdots, X_c]$。其中:X_i 表示尺度 i 上的特征集,c 为所考虑的尺度级数。然后依据式(1),分别将对应各尺度的特征集作为输入,得到对应各尺度的过完备基(或字典)$B_1, \cdots, B_c$(该过程如图 2 所示)。由于有了过完备和冗余表示理论[13]知,只要基满足过完备条件,利用稀疏编码这个描述能力极强的非线性编码工具,并不需要进行像文献[11]那样的字典去冗余处理。

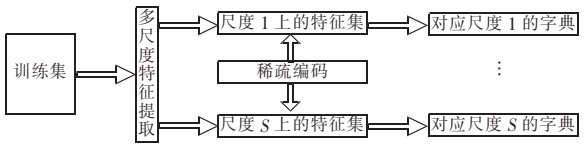

图 2　本文提出的多尺度字典学习算法

2.2　基于多尺度稀疏表示的图像表示方法

按照 2.1 节所述求得各尺度上的过完备字典后,本文提出针对不同尺度的特征,使用与其尺度对应的字典利用式(3)对其进行编码;然后按照 SPM[2]和文献[4]提出的特征各维最大汇总(max pooling)对各尺度上的特征编码结果分别进行汇总;最后将不同尺度上汇总的特征串接,作为对图像最终描述的全局向量。整个流程如图 3 所示。可以看出,按照这种方式汇总得到的全局向量,即刻画了图像在不同尺度下的特征,而且融入了不同尺度特征之间的空间关系。

2.3　本算法流程概述

利用 2.2 节所提出的图像表示方法将每幅图像概括为一

个体现了多尺度和空间信息的全局特征后，便可进行分类器的训练和决策。图 4 展示了本文提出的基于多尺度稀疏表示的场景分类算法流程。

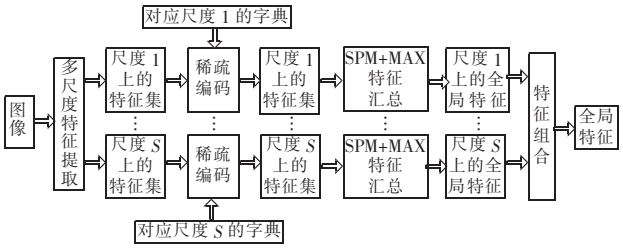

图 3　本文提出的多尺度字典学习算法

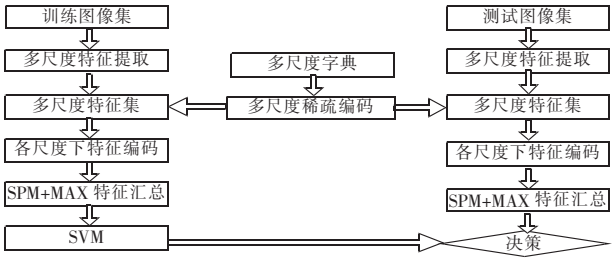

图 4　基于多尺度稀疏表示的场景分类算法框架

3　实验结果及分析

本章将介绍本文算法在三个常用的场景图像库 Scene-8[14]、Scene-15[15] 和 UIUC-Sports8[16] 上的分类结果。在实验中，并未利用颜色信息，而是事先将所有彩色图像均转换为灰度图像。为了减小运算量，在保持纵横比不变的条件下，将所有图像的尺寸缩放到了 300 × 300 像素以内。在所进行的实验中，共采用了 4 个不同的尺度，各尺度级别和相应的图像块尺寸如表 1 所示。

表 1　本实验所采用的尺度级别及相应的图像块尺寸

尺度级别	1	2	3	4
图像块尺寸	16 × 16	32 × 32	64 × 64	128 × 128

在对每幅图像提取局部特征时，本文采用了稠密采样的方式，即按照光栅扫描顺序，以步长 8、16、32、64 像素等间隔地从每幅图像中提取若干 16 × 16、32 × 32、64 × 64、128 × 128 像素的图像块。对于每个局部图像块，选用 128 维的 SIFT[12] 描述符作为其特征。在本实验中，将不同尺度上字典长度设定为如表 2 所示的三组参数。

表 2　本实验中采用的各尺度字典长度

尺度级别	1	2	3	4
第一组	512	256	128	64
第二组	1 024	512	128	64
第三组	1 024	512	256	128

在最后的特征汇总阶段，本文使用了 SPM[2] 及特征各维最大汇总[4] 的方法，其中空间金字塔的层数取为 0、1、2，各层上特征权重均为 1。在训练和测试 SVM 分类器时，则采用了 1 对多（one-vs-all）的方式。SVM 所使用的核函数为直方图相交核[17]（histogram intersection kernel，HIK）。为了保证结果的客观性，在每个库上独立进行了 10 次随机实验，并将平均准确率和标准差作为评价指标。

3.1　Scene-8 数据集

Scene-8[14] 数据集中共包含八个类别的 2 688 幅图像。各类别的名称及相应的样本数目如下：coast（360 幅）、forest（328 幅）、mountain（374 幅）、open country（410 幅）、highway（260 幅）、inner city（308 幅）、tall building（356 幅）、street（292 幅）。图 5 给出了该数据库所涉及的 8 个类别的一些样图。按照各文献通行的实验设置，在每次随机实验中，随机抽取 100 幅图像作为训练集，而将剩余图像作为测试集。表 3 给出了本算法在不同参数配置下与其他各类算法的最优结果比较，其中 ScSPM[4] + HIK 为将单尺度（16 × 16 像素）字典长度取为 1 024 时的分类结果。Qin[11] 中使用了四个尺度，且总字典长度为 6 304，远多于本算法的字典长度。

图 5　Scene-8 数据集的样图

表 3　各算法在 Scene-8 上的平均准确率

算法	平均准确率/%
Bosch[18]	87.80
Qin[11]	88.81
ScSPM[4] + HIK	88.10
本算法（第一组）	88.83 ± 0.39
本算法（第二组）	89.01 ± 0.65
本算法（第三组）	89.81 ± 0.61

由以上结果可以看出，本算法由于在图像表示中融入了多尺度特征，在所设置的三种参数下与同样使用稀疏编码和单尺度特征的 ScSPM[4] 相比，性能均获得较明显提升。且当字典长度增大时，一直表现出上升的趋势，其中第三组参数下所获的性能提升超过 1.7%；而与采用 K-means 学习多尺度字典的 Qin[11] 相比，在字典长度相同的情况下，性能提升超过了 1%。

3.2　Scene-15 数据集

Scene-15[15] 数据集共包含 15 个类别的 4 485 幅图像，该数据集除了 Scene-8 中的八个场景类外，还包含七个新增类别：bedroom（216 幅）、kitchen（210 幅）、living room（289 幅）、office（215 幅）、suburb（241 幅）、store（315 幅）及 industrial（311 幅）。图 6 给出了该库的一些样图。每个类别中的图像从 200 幅到 400 幅不等。在每次随机实验中，从各类抽取 100 幅图像作为训练集，而将其余图像作为测试集。表 4 给出了本算法在不同字典参数配置下与其他算法的性能比较。其中 ScSPM[4] + HIK 为将单尺度（16 × 16 像素）字典长度取为1 024 时的分类结果。

图 6　Scene-15 数据集的一些样图

表 4　各算法在 Scene-15 上的平均准确率

算法	平均准确率/%
SPM[2]	81.40
ScSPM[4] + HIK	82.32
本算法（第一组）	82.70 ± 0.43
本算法（第二组）	83.57 ± 0.57
本算法（第三组）	83.69 ± 0.52

从以上结果可以看出,本算法由于在图像表示阶段使用多尺度特征对图像内容进行了多层次描述,与其他使用单尺度特征的方法相比,与同样使用稀疏编码的 ScSPM[4] 相比,在三组不同字典长度配置下,性能均获得明显提升。随着字典长度的增大,分类准确率一直呈现出上升的趋势,其中第三组参数下与 ScSPM[4] 相比,性能提升超过了 1.3%。

3.3　UIUC-Sports8 数据集

UIUC-Sports8[16] 数据集中共包含八个类别的 1 578 幅图像,如图 7 所示。各类别名称及相应的样本数如下:badminton(200 幅)、bocce(137 幅)、croquet(236 幅)、polo(182 幅)、rock climbing(194 幅)、rowing(250 幅)、sailing(190 幅)、snowboarding(190 幅)。为与其他方法进行比较,在每次随机实验中,从每类中随机抽取 70 幅图像作为训练集,然后从剩余图像中随机抽取 60 幅作为测试集。表 5 给出了本算法与其他算法在不同尺度字典的不同长度组合下的性能比较,其中 ScSPM[4] + HIK 为将单尺度(16×16 像素)字典长度取为 1 024 时的分类结果。

图 7　UIUC-Sports8 数据集的部分样图

表 5　各算法在 Scene-15 上的平均准确率

算法	平均准确率/%
Fei-Fei[16]	73.40
ScSPM[4] + HIK	82.74
本算法(第一组)	83.66 ± 1.09
本算法(第二组)	84.63 ± 0.72
本算法(第三组)	85.49 ± 1.17

可以看出,本算法与只使用单尺度特征的方法相比,在三组不同字典长度配置下,均表现出明显优势。随着字典长度的增大,准确率一直呈现出上升的趋势,其中在第三组参数配置下,与 ScSPM[4] 相比,性能提升了近 3%。

综合以上三个标准场景库上的分类结果可以推知:不同尺度的特征具有很强的互补性,且在分类任务中能增强图像表示的区分性;所进行的分类实验充分表明本文所提出的基于多尺度稀疏表示的图像多层次描述方法的有效性。

4　结束语

本文通过在字典学习阶段融入多尺度信息,实现了对近年提出的基于稀疏编码的图像分类代表性框架 ScSPM[4] 的多尺度扩展,将局部特征和全局特征统一在一个框架之下。在所提出的基于多尺度稀疏表示的图像表示方法中,既体现了多尺度信息,又融合了不同尺度特征的空间位置关系,增强了多尺度字典的描述能力。在三个标准的场景测试集上的实验结果表明,本文提出的算法由于利用了不同尺度特征间鉴别性的互补关系,与采用单尺度特征的方法相比,性能有了显著提升。

参考文献:

[1] SIVIC J, ZISSERMAN A. Video google: a text retrieval approach to object matching in videos[C]//Proc of International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2003: 1470-1477.

[2] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]// Proc of International Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2006: 2169-2178.

[3] GEMERT J, VEENMAN C, SMEULDERS A, *et al*. Visual word ambiguity[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(7): 1271-1283.

[4] YANG Jian-chao, YU Kai, GONG Yi-hong, *et al*. Linear spatial pyramid matching using sparse coding for image classification[C]//Proc of International Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2009: 1794-1801.

[5] OLSHAUSEN B A, FIELD D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381(6583): 607-609.

[6] PATI Y C, REZAIIFAR R, KRISHNAPRASAD P S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition[C]//Proc of the 27th Asilomar Conference on Signals, System, and Computers. [S. l.]: IEEE, 1993: 40-44.

[7] LEE H, BATTLE A, RAINA R, *et al*. Efficient sparse coding algorithms[C]//Proc of Advances in Neural Information Processing Systems. 2007: 801-808.

[8] MAIRAL J, SAPIRO G, ELAD M. Multiscale sparse image representation with learned dictionaries[C]//Proc of International Conference on Image Processing. [S. l.]: IEEE Computer Society, 2007: 105-108.

[9] MAIRAL J, SAPIRO G, ELAD M. Learning multiscale sparse representations for image and video restoration[J]. Multiscale Modeling & Simulation, 2008, 7(1): 471-478.

[10] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Trans on Signal Processing, 2006, 54(11): 4311-4322.

[11] QIN Jian-zhao, YUNG N. Scene categorization with multiscale category-specific visual words[J]. SPIE Optical Engineering, 2009, 48(4): 1-13.

[12] LOWE D. Distinctive image featuers from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[13] LEWICKI M, SEJNOWSKI T. Learning overcomplete representations[J]. Neural Computation, 2000, 12(2): 337-365.

[14] OLIVA A, TORRALBA A. Modeling the shape of the scene: a holistic representation of the spatial envelop[J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[15] LI Fei-fei, PERONA P. A Bayesian hierarchical model for learning natural scene categories[C]//Proc of International Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2005: 524-531.

[16] LI Fei-fei, LI Li-jia. What, where and who? classifying events by scene and object recognition[C]//Proc of International Conference on Computer Vision. [S. l.]: IEEE Computer Society, 2007: 1-8.

[17] BARLA A, ODONE F, VERRI A. Histogram intersection kernel for image classification[C]//Proc of International Conference on Image Processing. [S. l.]: IEEE Computer Society, 2003: 513-516.

[18] BOSCH A, ZISSERMAN A. Scene classification using a hybrid generative/discriminative approach[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(4): 712-727.