指导教师：杨涛　　提交时间：2016/3/13

CVPR2015 Paper Translation

No: 01
姓名：张林江
学号：2013302536
班号：10011303

基于特征空间上下文语义场景识别系统

孙新航，蒋树强，Luis Herranz

中国科学院(CAS)计算机技术研究所智能信息处理重点实验室

中科院、北京,100190,中国

准确的场景识别仍然是一个挑战,因为它意味着推

摘　　要

在语义多项框架下的影像和图像在语义建模中作为单纯形概率。影像主题模型是通过图像标签,学会采取弱监督导致的问题场景类别的语义空间。幸运的是,每个类别都有自己的同现模式跨类别的图片是一致的。因此,发现和建模这些模式是至关重要的提高识别 peiformance 表示。在这篇文章中,我们观察到,不仅全球共生映像级别很重要,但不同地区也有不同的类别同现模式。我们利用本地上下文关系来解决发现的问题一致的同现模式和消除噪声的。我们的假设是一个不太嘈杂的语义表示,将极大地帮助分类器模型一致的共生和更好的区分场景类别。建模的一个重要优势特征语义空间,这个空间是功能独立。因此,我们可以结合多个特性和空间邻居同样的公共空间,并制定问题作为上下文相关的能量最小化。实验结果表明,利用不同类型的上下文关系持续提高了识别精度。特别是,从该方法获益更多更大的数据集,从而导致竞争非常激烈的性能。

1、简介

通常,一个场景是一个非常抽象的表示由许多少抽象语义实体局部地区(如天空、岩石、表、汽车)。理从低层视觉特征到高层场景类别。可以建模场景类别直接从低描述符，然而,所需的统计知识来推断场景类别(如海岸、山、办公室)是很难获得直接从低级视觉描述符,由于巨大的语义鸿沟。

更合理的方法是将推理的语义差距较小的两个(或更多)步骤(如功能主题、主题场景)。这是典型的中间表示局部地区的形象,中级的词汇概念或定义主题。图 la-b 两张图片显示了一个示例及其相应的地区中层的主题。这个词汇表可以定义明确,但需要标签区域与相应的主题培训特定的主题分类器。相反,主题可以建模为隐藏主题空间中的潜在的被发现在勒荷兰国际集团(ing)[7,32 岁,30 岁,17 岁,16)。主题 low-Ievel 视觉捕捉共生特性,从共病的主题和场景类别建模一个形象。

替代(预定义或隐藏)中级词汇直接学习中级主题使用场景分类标签。注意,主题仍在当地,但同一词汇称为场景类别。在本文中,我们侧重于语义多项式(SMN)表示[21]和[8]22 日,23 日,其扩展。语义多项代表某一块的概率(或图像)属于每个场景类别。作为一个概率,它位于一个概率单纯形(语义单工或语义空间)。没有当地注释是可用的,所有的影像一个图像共享相同的标签,但它们对应于不同的地区有不同的中间概念。这 weakly-supervised 学习归纳相关场景类别,共享相同的地区中层概念(如天空、道路、树木)会显示某些 SMN 概率,导致类别共病的表现(见图 lc)。我们称之为(场景)[类别共生]

作者使用这个词指代一致的上下文共生,因此理想的同现模式。在这里,我们把他们称为(场景)类别共生强调他们是高层次的类别而不是低或中层共生。我们也希望)。

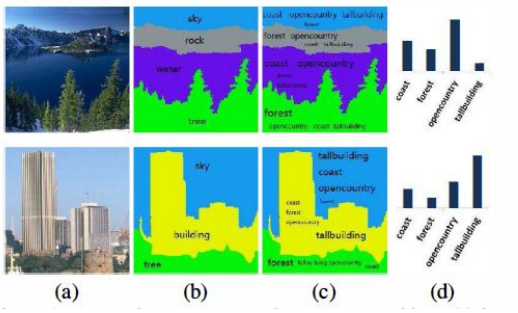
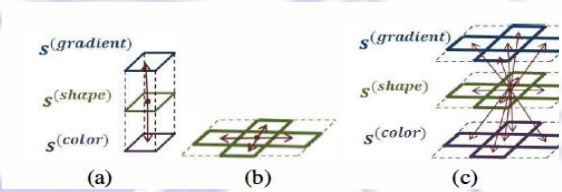


图 1 所示。场景识别类型的共生:(a)的照片 opencountry(上面一行)和 tallbuilding(底部行)类别的 15 个场景数据集,(b)地区相应的中级主题和词汇,在不同地区(c)的场景类别共存造成 weakly-supervised 学习图像分类标签,和(d)相应的图像语义多项式。

和塞·伐斯冈萨雷斯[23]表明,他的这些同现模式是一致的在同一类别的图像,所以他们可以建模和分开意外共存语义表示(即噪声)与一个合适的分类器(例如狄利克雷混合物[23],SVM[8])。他们还指出,影像 SMNs 模型可靠的同现模型太吵了,所以多个影像 SMNs 聚合到单个图像 SMN 小心保存同现模式(参见图 Id)。因此,这些只能在映像级别模型全球同现模式。

相比之下,我们想关注本地影像 SMNs 类别共生,尽可能多的类别共存取决于特定区域(参见图 Ic)。我们的动机是利用(以一种无监督的方式)的上下文关系加强一致的同现模式和快速眼动



(a)特征的背景下,空间上下文,(b)和(c)联合多元特征空间情境。

联合上下文模型加强一致的同现模式和过滤出意外的。我们表明,交付清洁 SMNs 分类器可以帮助发现内在的同现模式可以模拟一个场景类别,从而提高识别性能。剩下的纸是组织如下。部分 2 和 3 审

和相关工作和语义多项式框架。该联合上下文模型描述在第四节。实验是在第五节第六节的结论。

2、相关工作

许多方法提出中层表示使用显式的分类器。沃格尔和 Schiele[29]提出了词汇与九个地方自然场景概念模型。对象银行[13 日 38]是一个编码的语义表示的响应在不同空间位置的 pretrained 对象分类器。Classemes[l]是基于中间语义表示一组 2659 个基点的类。这些方法需要显式的中级培训,经常利用大量的外部培训数据(例如 ImageNet)勒^这些中层分类器。

密切相关,本文 Vicept[14][21]和方法扩展 SMN 框架只需要将现场图像的类别。Rasiwasia 和塞·伐斯冈萨雷斯[23]他提出一个上下文模型基于狄利克雷混合物模型上下文共生。Kwitt 等[8]提出一个区别的版本,使用 SVM 结合合适的语义空间的内核(即-测地线内核[37])。在这种情况下,语义单工被描述为一个语义歧管(SM)。这种方法扩展通过空间金字塔与粗糙的空间上下文(IO)和一个近似嵌入 NGD 内核的大规模识别算法。相比之下,我们显式地进行编码 SMN 邻近的影像和多个特性之间的关系。这些作品只在图像 SMNs 全球共生模式,虽然大多数我们提出的技术专注于当地在影像级别共存。潜在的主题模型通常使用潜在的建模

狄利克雷分配(LDA)[7,24]。然而,大多数 LDA 已被证明捕捉无关的一般规律而不是感兴趣的语义规律,由于穷人监督[24]。空间上下文可以包含模型主题的全球布局和执行本地一致性(32 岁,17)。最近,李、郭[16]提出了 patch-based 潜在的共同框架,勒^上下文表示和分类模型。大多数潜在主题模型生成,通常不能很好地扩展到大型数据集。主题模型相比,我们的方法有两个主要区别。首先,词汇的影像 SMNs 仍然是(高层)场景类别,而主题模型中层表示。第二,目标是鼓励上下文共生,然后让分类器消除了歧义的后验。

各级共生在许多场景理解系统的核心。朗等[9]提出一个功能同现矩阵用于场景分类。主题在计算机视觉模型本质上低级视觉共生。李、郭[15]提出图像分割成 superpixels,分类成预测对象类,然后利用对象共生现场类别。与这些类型的共生、弱监督学习 SMN 表示诱发一种非常特殊的共生(如类别共

存),最高的抽象。

3、框架概述

3.1 语义分析

语义歧管也是基于语义多项中层主题[23]表示。每个类别的概率分布估计从当地的视觉描述符中定义的一些视觉空间 x 图像表示为一袋当地视觉描述符 $I = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \mathbf{x}_n \in X$,人口抽样的网格与当地 N 影像。给定一个场景的词汇类别{ WI,.....W M },每个图像与一个标签这些类别。丁标签不可用,主题条件分布 Px1w(xnlw)学习通过图像标签使用弱监督。所有的影像在一个给定的图像共享相同的标签(即场景类别),我们表明,诱发类别共生。主题条件分布建模为混合物高斯模型(GMM),一个模型/场景类别。

概率 s =(Sl 的...,SM)T 和 Sw = PWlx(wlxn),可称为语义多项呢(SMN)影像 Xn[21],它位于(语义)单工。多个影像 SMNs 结合成一个单一的形象使用 voting-based SMN 的方法。首先,最可能的一类是分配给每个影像 w 然后计算得到的直方图是出现每个类别的形象噢= 1 {wn:w 获得的图像 SMN 年代。

$$s_w = \Omega_w^{\text{vot}}(I) = \frac{o_w + \beta - 1}{\sum_{w=1}^{M}(o_w + \beta - 1)} \qquad (1)$$

f3 是正则化参数。在图像类别共生 SMNs 建模使用支持向量机。注意,更一致的共存模式是在图像训练集,更好的分类器区分类别。而不是使用传统的内核(如多项式,RBF),内核为特定几何形状的设计语义使用单纯形,基于测地距离 9(s,s)= 2 arccos(\ vIS,R))vIS 表示 element-wise 平方根。一个消极的测地距离 (NGD) 内核可以定义从这个距离 kNGD(s,s)= - g(s,s)[37]。最后,一个空间金字塔表示用于粗略的编码空间上下文。

注意,使用内核限制了支持向量机的应用大型数据集分类器,由于计算成本。Kwitt 等[8]还提出一个近似的映射 NGD 内核,所以相同的框架可用于大规模场景识别与一个线性支持向量机相结合。图 3 显示了我们的场景识别框架,构建在语义歧管框架有一些差异。首先,我们的框架包括多个特性,然后我们学习 feature-specific 主题模型结合他们的语义空间。其次,结合之前影像 SMN 年代 SMN 成一个单一的形象,我们的过程影像 SMNs 使用联合特征空间上下文模型。最后,我们使用一个不同的近似的 NGD 内核基于低维投影特征空间[4]。

3.2 在语义特征组合空间

而不是单一类型的视觉特性,我们现在考虑一个集合 V 互补的(在我们的实验 V = {梯度,形状、颜色})。每个特性 v E 生成一组当地视觉描述符 ltv)= {十四),.....我和,x{我(1).....我(IYI)}代表图像的所有特性。现在我们假设我们学习 feature-specific 主题模型在相同的独立 pxev)IW(x 方法在单一特征的情况下。因此,我们可以定义 feature-specific 影像 SMN 的影像 n 和特性 v s,s 图 4 显示了一个示例有三个 SMNs feature-dependent 影像。在这个图我们可以观察到某些地区如何吵着比其他人在吗一些特性。我们还可以观察到特定的模式类别(类别共生),在(interfeature 特性邻国之间的关系)和影像(空间关系)。注意所有 SMNs 躺在相同的(语义)和空间所有代表一个概率,所以我们可以结合使用概率模型。特别是,我们获得 multifeature 从几个 feature-dependent SMN SMNs 作为代表 SMN 接近他们。一个合适的选择的概率分布是 KullbackLeibler 最小化(吉隆坡)发散[23]

$$\mathbf{s}_n = \underset{\hat{\mathbf{s}}_n}{argmin} \sum_{v \in V} KL(\mathbf{s}_n^{(v)} || \hat{\mathbf{s}}_n) \qquad (2)$$

这将导致

$$s_{nw} = \frac{\exp(\frac{1}{|V|}\sum_{v \in V}\log(s_{nw}^{(v)}))}{\sum_{w \in W}\exp(\frac{1}{|V|}\sum_{v \in V}\log(s_{nw}^{(v)}))} \qquad (3)$$

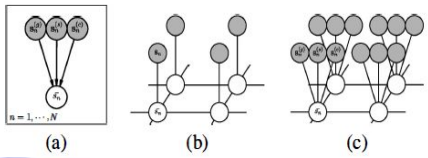
4、特征空间上下文模型

4.1 全球模型

利用空间的背景下,我们认为的关系邻国之间的影像。featuredependent 相反 SMNs,我们可以使用类似的方法使用,但这里每个影像并不是独立的。这里

我们求助于无向的模型。

我们首先制定去噪影像的问题 SMNs 使用马尔可夫随机场(MRF)4-connectivity 网格(参见图 5 b)。



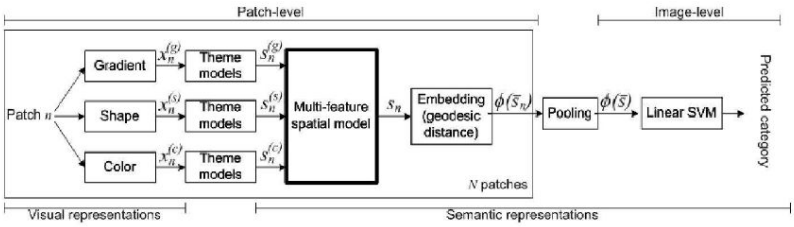


Figure 3. Overview of the recognition framework with the proposed methods highlighted.

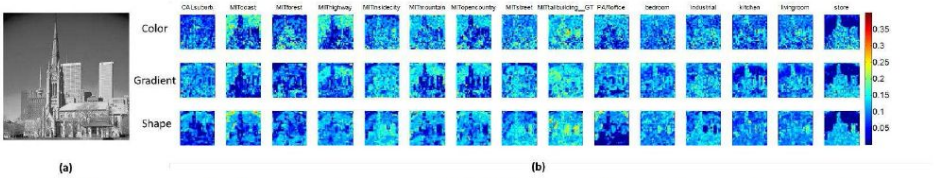


Figure 4. Patch SMNs: (a) image of the *15 scenes* dataset (category: *MITtallbuilding*), and (b) probability maps illustrating each component of the patch SMNs. Each row corresponds to SMNs obtained for a different visual descriptor.

29.698 厘米

图 5。上下文模型:(a)特征组合,(b)四连接空间网格模型和(c)特征空间网格模型。

一个功能,目标是 maxnlllze 在观察 SMNs 联合概率和去噪 SMNs 集定义为 P(的,.....SN,Sl,...SN)= iexp(- e 的,.....SN,Sl,...SN)),Z 是分区规范化的概率函数。因此,这个问题相当于减少全球能源的网络建模为

$$E\left(\bar{\mathbf{s}_1},\ldots\bar{\mathbf{s}_N},\mathbf{s}_1,\ldots\mathbf{s}_N\right) = \sum_n g(\bar{\mathbf{s}_n},\mathbf{s}_n) + \alpha \sum_{\{n,n'\}} g(\bar{\mathbf{s}_n},\bar{\mathbf{s}_{n'}}) \quad (4)$$

s -;是影像的未知去噪 SMN n(相比之下原 sn)和{ n,n ' }代表双连接影像。我们模型的能量之间的距离 SMNs。单形概率是一个合适的选择测地距离 g(s,s)[37]。我们选择在 KL 散度(2)使用因为 KL 散度是不对称的,和语义歧管框架已被证明有效的[8]。

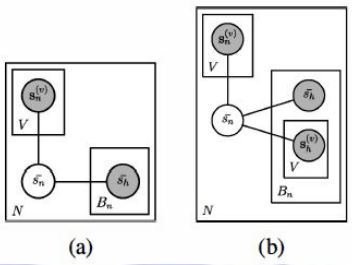


图 6。本地块上下文模型(联合特征空间):(一)只特性从目标影像,和(b)特性所有的影像的社区目标影像

既是 feature-dependent SMN 年代和去噪 SMNs 在同一个空间中,这个模型可以很容易地扩展多个功能使用模型在图 5 c。相应的能量

$$E\left(\bar{\mathbf{s}_1},\ldots\bar{\mathbf{s}_N},\mathbf{s}_1^{(1)},\ldots\mathbf{s}_N^{(1)},\mathbf{s}_1^{(|V|)},\ldots\mathbf{s}_N^{(|V|)}\right) = \sum_n \sum_{v\in V} g\left(\bar{\mathbf{s}_n},\mathbf{s}_n^{(v)}\right) + \alpha \sum_{\{n,n'\}} g\left(\bar{\mathbf{s}_n},\bar{\mathbf{s}_{n'}}\right) \quad (5)$$

为了解决优化问题,我们采取的迭代有条件的模式(ICM)算法[2],该循环在不同的影像最小化能量相关用一个变量保持其他变量节点固定。它可以被视为坐标态梯度下降法。这个算法收敛于一个局部最大值的概率。其他算法可以使用,比如图削减,但他们的扩展到更大的社区没有成对派系是困难的,计算更加昂贵和他们不导致当地制定 ICM。

4. 2 本地模型

ICM 的算法更新每个影像的价值减少当地相关的能量,保持固定其他影像变量的价值。现在我们可以定义社区 Bn 作为影像的邻居集 n。图 5 b 的情况下,Bn 包含四个邻居。现在我们可以用该模型 N 独立 patch-centred 子图(见图 6,所有 h(h i - n)考虑观察到的特定影像 n)

$$E\left(\bar{\mathbf{s}_n};\phi_n\right) = \frac{1}{|V|}\sum_{v\in V} g\left(\bar{\mathbf{s}_n},\mathbf{s}_n^{(v)}\right) + \alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n} g\left(\bar{\mathbf{s}_n},\bar{\mathbf{s}_h}\right) \quad (6)$$

¢n = {sSMN 年代特征空间附近的影像 n。为了方便我们也正常的大小社区 IBnl 和新特性。包括大社区的全球模型之前部分会导致一些因素不再成对,问题的复杂性显著增加。然而,通过使用这个局部逼

近我们可以很容易地包括更大的社区。我们也考虑一个扩展的背景下,不仅认为 feature-dependent SMNs 从目标影像,但也从邻居(图形模型显示在图 6 b)。

$$E(\bar{s_n};\phi_n)=\frac{1}{|V|}\sum_{v\in V}g(\bar{s_n},s_n^{(v)})+\alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n}g(\bar{s_n},\bar{s_h})+\beta\frac{1}{|B_n||V|}\sum_{\{n,h\},h\in B_n}\sum_{v\in V}g(\bar{s_n},s_h^{(v)}) \quad (7)$$

now with $\phi_n=\left\{s_n^{(v)}|\forall v\in V\right\}\bigcup\left\{s_h^{(v)}|\forall h\in B_n,\forall v\in V\right\}$.

最后,我们包括额外的能量惩罚太平坦 SMNs,这将导致不提供信息的影像:

$$E'(\bar{s_n};\phi_n)=E(\bar{s_n};\phi_n)+\lambda H(\bar{s_n}) \quad (8)$$

当 $H(\mathbf{s})=-\sum_{w=1}^{M}s_w\log(s_w)$ 的熵。ICM 算法同样的想法后,我们循环在每个影像的影像最小化(8)n。这个问题使用梯度下降法可以解决。相对应的梯度影像 n

$$\frac{\partial E'(\bar{s_n};\phi_n)}{\partial s_{nw}}=\frac{1}{|V|}\sum_{v\in V}f\left(s_{\bar{n}w},s_{nw}^{(v)}\right)+\alpha\frac{1}{|B_n|}\sum_{\{n,h\},h\in B_n}f(s_{\bar{n}w},s_{\bar{h}w})+\beta\frac{1}{|B_n||V|}\sum_{\{n,h\},h\in B_n}\sum_{v\in V}f\left(s_{\bar{n}w},s_{hw}^{(v)}\right)-\gamma(1+\log(s_{\bar{n}w_n})) \quad (9)$$

where

$$f(x,y)=\frac{\partial g(x,y)}{\partial x}=-\frac{\sqrt{y}}{2\sqrt{x}\sqrt{1-(\sqrt{x}\sqrt{y})^2}}$$

5、实验

在本节中,我们评估不同的上下文模型前面描述的在不同的数据集,比较相关的工作。

5. 1 实验装置

数据集 提出的方法在三个评估小的数据集。15 个场景(7,10)包含 4485 张图片在 15 个场景类别。LabeIMe[18]由 8 户外场景分类,共有 2600 张图片。UIUCSports[12)1585 年由图像标记为 8 复杂

运动场景类别。在之前的设置后工作中,我们使用 100 年,100 年和 70 年的训练图像,分别。我们也评估拟议的方法大规模数据集,包括 MIT67[20]SUN397[34]。15620 室内 MIT67 包含 15620 张图片场景类。SUN397[34]由 397 个类别、总共 108762 张图片。对于 MIT67 室内 SUN397,培训/测试提供的配置原来的作者。

视觉和语义特征 我们使用三种内核描述符[3]的局部描述符,包括梯度,(LBP)形状和颜色。所有本地视觉描述符在常规提取 16 x 16 像素密度网格(步骤 8 像素)。与 512 年为主题,我们训练 gmm 混合物每个主题模型。

我们也将描述符使用与四层空间金字塔(lO)(1 x 1,2 x 2,3 x 3,4 x 4)对支持向量机分类。

基线。我们比较我们的方法相同框架没有上下文模型,相当于空间金字塔语义歧管(SPMSM)[8]使用不同的近似嵌入[4]。我们评估独立在同一视觉特性。

变化的方法我们评估四个变化的上下文模型:

•特征上下文(MF):多个特征的总和在语义空间使用(3),相应的图 2 中的上下文和模型如图 5 所示。

•空间上下文(S):单一特性相邻关系(参见图 2 b)。通过最小化(6)当只使用一个特性。

•特征空间上下文(MFS):结合目标影像的多个特性和相邻关系(例如图2 a 和 b)的组合。通过最小化(6)的特征。

•扩展特征空间上下文(电磁辐射):包括多个特性从附近的影像(见图 2)。通过最小化(6)和相应的模型在图 6 b。

5. 2 社区规模和熵的影响正则化

两个关键参数的大小空间社区和熵正则化的影响同现模式。我们评估他们在 15

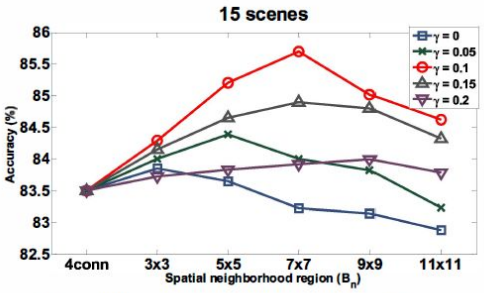


Figure 8. Region size and sparse parameter evaluation

场景数据集,使用电磁辐射和修复 f3 为 1。结果如图 8 所示。我们评估不同的社区,包括 4 -

空间邻域如图 2 b 和连接性 c、密度和其他社区的大小(3 x3 LxL 影像对应于 8 邻居)。我们可以观察到大社区可以有效地强化一致的模式和过滤意外的。但是,太大的社区不能正确地捕捉当地同现模式。从我们的实验中,一个好的权衡是 7 x7 影像。

我们也评价熵正则化的影响不同,从 0 到 0.2,0.05

的步骤。一般来说,性能增加,最大的约 0.1 然后减少。图 7 说明了熵的影响影像 SMN 的年代,没有惩罚的熵(= 0)我们获得了菲亚特影像 SMNs(即高熵)不适合同现建模。过低熵在影像 SMNs 也不是有用的(= 0.2)类别可能主导概率太高,很少有一致的同现模式。最好的结果这个实验是,= 0.1 和 L = 7。为其余的实验我们将使用这个配置,虽然具体的参数可能会提高性能。

5. 3 环境模型

我们评估提出的不同变化方法三个小规模数据集显示不同类型的上下文模型提高准确性。表 1 显示了分类精度,增加持续当我们模型中不同类型的上下文。结合多个特性帮助增加 1.1 - -2.5%左右在最好的单一功能。空间上下文变量而且没有收获不尽相同,小幅上涨约 1%。然而,结合既可以增加一个额外的 0.5 -1%只特征上下文。扩展 multifeature 空间环境贡献一个额外的 0.5 - -1%获得通过合并相邻的多个特性影像。的总增益扩展上下文模型在没有上下文的基线大约是 2.6 - -5.7%。

Table 1. Accuracy (%) for different context models.

Method (feature)	15 scenes	LabelMe	Sports
No context model			
Baseline (gradient)	78.9	86.5	83.9
Baseline (shape)	80.0	85.0	84.3
Baseline (color)	75.4	72.4	72.8
Spatial context (7x7 patches)			
Spatial (Gradient)	81.0	86.7	83.7
Spatial (Shape)	81.4	84.9	83.9
Spatial (Color)	76.6	72.9	73.1
Multiple feature context			
Multi-feature	82.5	88.3	85.4
Joint multi-feature spatial context (7x7 patches, $\gamma = 0.1$)			
MFS	83.5	88.9	85.9
Extended MFS	**85.7**	**89.3**	**86.9**

5. 4 相关的工作

我们使用中层语义与近期作品表示,如潜在的主题模型(12 日 30 日 16】15 日,24 日,SMN 框架的扩展(23 日 8)和其他扩展等银行[13 日 38]和对象 Classemes[1]。大多数这些方法不能使用在大规模数据集,我们单独的比较小数据集和更大的数据集。

5. 4. 1 小规模数据集

表 1 比较了两种方法的结果报告的作者相应的引用。虽然一个完全公平比较是不可能的,由于不同的实现,特性和其他参数,我们的框架至少似乎是很有竞争力的三个评估数据集。与方法基于 SMNs 进行比较特别感兴趣。注意,上下文多项式(复合材料)利用 CCO SMN 使用生成的图像水平模型,狄利克雷混合模型(DMM)。SPMSM

Table 2. Comparison with related works.

Dataset	Method	Accuracy (%)
15 scenes	SMN[23]	71.7
	LDA[24]	76.6
	CMN[23]	77.2
	ObjectBank[16]	80.9
	Kernel descriptor[3]*	82.2
	SPMSM[8]	82.5
	SR-LSR[16]	85.7
	Proposed (EMFS)	**85.7**
	Object-to-Class kernels[38]	88.8
LabelMe	Wang et al[30]	76.0
	SPMSM[8]	87.5
	Kernel descriptor[3]*	87.3
	Proposed (EMFS)	**89.3**
	SR-LSR[16]	89.8
Sports	Li and Fei-Fei[12]	73.4
	ObjectBank[13]	76.3
	SPMSM[8]	83.0
	SR-LSR[16]	83.9
	Kernel descriptor[3]*	85.2
	Object-to-Class kernels[38]	86.0
	Proposed (EMFS)	**86.9**

* Results are based on our own implementation using the code available from the authors.

利用有识别力的分类和粗糙的空间环境,实现更好的性能。该方法,利用多个特性和当地环境吗影像级别达到更好的性能比的方法。我们还与直接建模的类别从相同的底层内核描述符(连接到把它们),和一个支持向量机和空间金字塔。我们观察我们的方法,它使用一个中层表示达到更好的结果。

Table 3. Comparison on MIT67 dataset.

MIT67	Method	Acc (%)
Proposed	Baseline (gradient)	34.7
	Baseline (shape)	36.9
	Baseline (color)	26.8
	Proposed (MF)	42.4
	Proposed (MFS)	44.7
	Proposed (EMFS)	**48.2**
State-of-the-art	ObjectBank[16]	37.6
	Object-to-Class kernels[38]	39.6
	Deformable Part Models[19]	43.1
	SPMSM[8]	44.0
	Sparse Spatial Coding[11]	44.4
	Geometric Phrase Pooling[35]	46.4
	Linear Distance Coding[33]	46.7
	IFV[27]	60.8
	Discriminative parts[5]	64.0
	Places-CNN[39]	68.2
	CNNaug-SVM [25]	**69.0**

5. 4. 2 大规模数据

我们评估拟议的中等规模的方法数据集 MIT67 SUN397 大得多。结果分别见表 3 和图 4。的收益由于将远高于在不同的上下文更小的数据集,对重大收益的 11%和 15%最好的单一功能基线 MIT67

SUN397 分别的数据集。这表明,上下文关系成为更

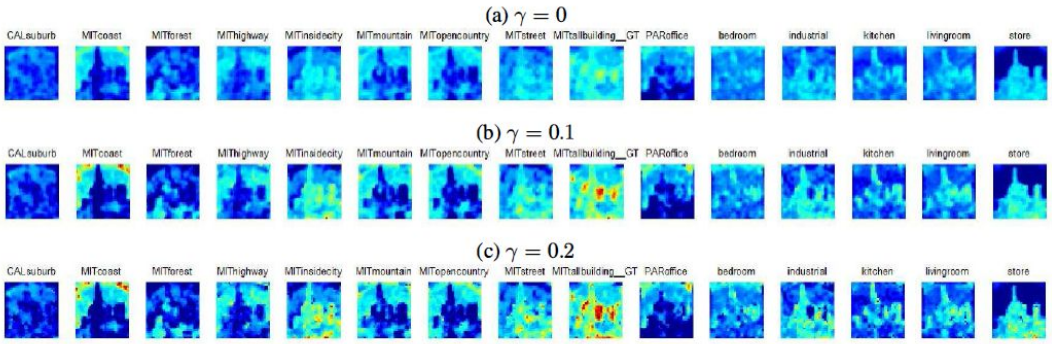


Figure 7. Effect of entropy regularization on the patch SMNs. The spatial neighborhood is 3x3 patches.

重要的重要的数字场景类别的增加,导致同现在这些大数据集模式噪声很大。利用上下文中强调代表类别共存模式可以大大有助于改善识别的性能。其他中层语义表征,如 Object-bank 和元类利用大大量的外部数据(e。g．.ImageNet)模型中层分类器。该方法优于他们没有求助于外部数据。

随着中层表示数量的方法可以训练了这些数据集是有限的,作为参考我们也与近期作品基于编码 bag-of-words 框架[35]33 岁,11 日,费舍尔向量[27],矿业歧视部分[5]和卷积神经网络(CNN)(6,39 岁,25)。我们实现稍微更好的性能比有限责任公司而不是费舍尔向量。不过要注意,比的结果[27]使用一个密集的网格采样地方特色导致更高的空间特性。矿业区别的部分也达到更好的性能在 MIT67。这个数据集包含室内场景部分原因表示可以与许多对象实现良好的性能。甚至作为一个纯粹的 scene-level 表示,我们的方法仍然达到竞争性能在这两个数据集。我们实现专业化。

Table 4. Comparison on SUN397 dataset.

SUN397	Method	Accuracy (%)
Proposed	Baseline (gradient)	25.4
	Baseline (shape)	23.2
	Baseline (color)	18.2
	Proposed (MF)	30.4
	Proposed (MFS)	34.9
	Proposed (EMFS)	**40.7**
State-of-the-art	SUN (HOG)[34]	27.2
	SPMSM[8]	28.2
	Meta-classes[1]	36.8
	SUN(MKL)[34]	38.0
	CNN (Decaf)[6]	40.9
	IFV[27]	47.2
	Places-CNN**[39]**	**54.3**

曼斯 CNN 功能学上 ImageNet[6]但不是地方[39],因为这个数据集是 scene-centric 因此更多合适。注意,在 CNN 相比,我们不使用任何外部数据。

6、结论

中间语义空间是非常有用的识别复杂的场景。与话题模型利用低,中层功能共生,我们关注特殊类型的模式导致的学习当地的主题模型与弱监督。利用这些模式正常(即场景类别共生)可以提高识别性能。我们扩展语义歧管框架[8]包括一个上下文模型集成多个特性和邻近的影像。我们利用属性在多个语义单纯形是一种常见的空间特性和相邻的影像可以自然地集成。一个联合上下文模型利用这些关系在这个框架提高性能的关键。在具体来说,大型数据集提出获益更多上下文模型,类的数量和高有用的类别共生模式是更微妙的和藏在嘈杂的模式。利用当地的空间和特征可以帮助发现一致的关系模式和过滤掉噪声模式,使事情变得更加容易的分类器可以专注于建模这些模式。

承认 这项工作是支持的中国国家基础研究计划(973 计划):2012 年 cb316400,部分国家自然科学中国的基础:61322212 和 61322212,部分由国家高科技发展计划(863 计划)中国:2014 aa015202,部分关键技术下的中国研发项目批准号 2012 bah18b02,部分和中科院总统的国际联谊计划:2011 YIGB05。这项工作也由 1319 年联想优秀青年科学家项目(S)。

参考文献

[1] A. Bergamo and L. Torresani. Classemes and other classifierbasedfeatures for efficient object categorization. In IEEETrans. on Pattern Anal. and Mach. Intell. , 2014.

[2] J. Besag. On the statistical analysis of dirtypictures. Journalof the Royal Statistical Society. Series B,

48(3):259-302,
1986.

[3] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visualrecognition. In NIPS, 2010.

[4] L. Bo and C. Sminchisescu. Efficient match kernel betweensets of features for visual recognition. In NIPS, 2009.

[5] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visualelement discovery as discriminative mode seeking. In NIPS,pages 494-502. 2013.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In ICML,2014.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model forlearning natural scene categories. In CVPR, 2005.

[8] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognitionon the semantic manifold. In ECCV, 2012.

[9] H. Lang, Y. Xi, J. Hu, L. Du, and H. Ling. Scene classificationby feature co-occurrence matrix. In Workshop on SceneUnderstanding for Autonomous System, ACCV, 2014.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags offeatures: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006.

[11] G. Leivas Oliveira, E. Nascimento, A. Wilson Vieira, and M. Montenegro Campos. Sparse spatial coding: A novel approach to visual recognition. IEEE Trans. on Image Process. 23(6):2719-2731, June 2014.

[12] L. Li and L. Fei-Fei. What, where and who? classifyingevents by scene and object recognition. In ICCV, 2007.

[13] L. Li, H.Su, E. Xing, and L. Fei-Fei. Object bank: A highlevel image representation for scene classification and semantic feature sparsification. In NIPS, 2010.

[14] L. Li, S. Jiang, and Q. Huang. Learning hierarchical semantic description via mixed-norm regularization for image understanding. IEEE Trans. Multimedia,14(5),2012.

[15] X. Li and Y. Guo. An object co-occurrence assisted hierarchicalmodel for scene understanding. In British Machine Vision Conference, pages 1-11,2012.

[16] X. Li and Y. Guo. Latent semantic representation learning for scene classification. In ICML, 2014.

[17] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In CVPR, 2012.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Com put. Vision, 42(3):145-175, May 2001.

[19] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In ICCV, 2011.

[20] A. Quattoni and A. Torralba. Recognizing indoor scenes. In CVPR,2009.

[21] N. Rasiwasia and N. Vasconcelos. Bridging the gap: Query by semantic example. IEEE Trans. on Multimedia, 9(5):923-938,2007.

[22] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. In CVPR, pages 1889-1895,2009.

[23] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. IEEE Trans. on Pattern Anal. And Mach. Intell. , 34(5):902-917, 2012.

[24] N. Rasiwasia and N. Vasconcelos. Latent dirichlet allocation models for image classification. IEEE Trans. on Pattern Anal. and Mach. Intell. , 35(11):2665-2679, 2013.

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition.In CVPR, 2014.

[26] J. Sanchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In Neural Comput.,2011.

[27] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. Int. J. Com put. Vision, 105(3):222-245,2013.

[28] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In ECCV, 2010.

[29] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. Int. J. Comput. Vision, 72(2):133-157, Apr. 2007.

[30] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In CVPR, pages 1903-1910,2009.

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In CVPR, 2010.

[32] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In NIPS, 2007.

[33] Z. Wang, J. Feng, S. Yan, and H. Xi. Linear distance

coding for image classification. IEEE Trans. on Image Process., 22(2):537-548, Feb 2013.

[34] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba.SUN database: Largescale scene recognition from abbey to zoo. In CVPR, 2010.

[35] L. Xie, Q. Tian, M. Wang, and B. Zhang. Spatial pooling of heterogeneous features for image classification. IEEE Trans. on Image Process. , 23(5):1994-2008, May 2014.

[36] J. Yang, K. Yu, Y. Gong, and T. S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In CVPR, 2009.

[37] D. Zhang, X. Chen, and W. S. Lee. Text classification with kernels on the multinomial manifold. In RDIR, pages 266- 273,2005.

[38] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. IEEE Trans. on Image Process. 23(8):3241-3253, Aug 2014.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. 1320 Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, NIPS, pages 487-495,2014.