Focus + Context 语义表征的场景图像分割

吴 绿[1],张馨月[2],唐 茉[2],王 梓[3],王永安[4]
(1. 武汉理工大学宽带无线通信与传感器网络湖北省重点实验室,湖北武汉 430070;
2. 武汉大学资源与环境科学学院,湖北武汉 430079; 3. 中国人民解放军 95028 部队,湖北武汉 430070;
4. 中国电子科技集团公司第五十四研究所,河北石家庄 050081)

摘 要: 场景图像分割一直是机器视觉学习中较为复杂的重难点问题.本文在机器视觉注意力机制学习方法的基础上,融合人类对事物个体的认知,提出场景对象的 Focus + Context 语义表征,将对象类别信息带入图像底层特征学习中,运用概率统计理论,在抽象层上建模局部区域对象,再联合上下文语义信息推理全局与局部区域对象之间的关系,以实现类内焦点对象(Focus)突出的场景语义分割.实验验证,基于 Focus + Context 的语义表征和建模能够增加对象的识别率,尤其是在小样本环境下,所提出的方法能极大地简化场景的理解.

关键词: 场景图像分割; Focus + Context; 语义表征; 主题模型
中图分类号: TP391.7 **文献标识码:** A **文章编号:** 0372-2112 (2021)03-0596-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200161

Focus + Context Semantic Representation in Scene Segmentation

WU Lü[1],ZHANG Xin-yue[2],TANG Mo[2],WANG Zi[3],WANG Yong-an[4]
(1. *Key Laboratory of Broadband Wireless Communication and Sensor Networks*, *Wuhan University of Technology*, *Wuhan*, *Hubei* 430070, *China*;
2. *School of Resource and Environmental Sciences*, *Wuhan University*, *Wuhan*, *Hubei* 430079, *China*;
3. *A Unit of the Chinese People's Liberation Army*, *Wuhan*, *Hubei* 430070, *China*;
4. *The* 54*th Research Institute of China Electronics Technology Group Corporation*, *Shijiazhuang*, *Hebei* 050081, *China*)

Abstract: Scene segmentation has always been a key and complicated problem in machine learning. In order to understand the scene and recognize the objects more accurately, this paper adopts human attention mechanism, takes the category semantic information into consideration and merges it into the image feature learning. The Focus + Context semantic representation is proposed, where the context describes the relationship between the focus and different objects in the scene, and the focus shared among the same category are composed of similar clusters. The probabilistic topic model is used to compute the local features as well as their semantic information. The experimental results show that the Focus + Context method increases the recognition rate of the scene objects, and specially, the proposed method, in a local and global understanding way, can simplify the scene recognition greatly under a small sample size.

Key words: scene segmentation; Focus + Context; semantic representation; topic model

1 引言

场景图像分割是计算机视觉领域研究的重难点问题之一,它是诸如智能驾驶[1]、人机交互[2]、目标追踪[3]等应用实现的重要技术环节.由于对象不同位置所传递的语义不尽相同,所以图像区域的准确划分是对多义性场景语义准确表达的一个重要参考因素.现阶段,图像分割方法主要分为基于图像底层特征的分类分析[4],以及基于深度卷积模型的语义分割方法[5].基于特征的分割是对图像中具有共同语义属性的特征分类;语义分割也属于分类问题,它将图像分割成几组具有某种特定含义的像素区域,然后学习出每个区域的语义类别,进而获得一副具有像素标注的分割图像.所以,无论是基于图像特征还是基于图像像素的分类,如果能区分出最具代表性的场景对象,并以这个对象为中心展开相关区域对象的特征描述和语义识别,那么场景内容即可得到有效表达;然而,这个焦点对象的

收稿日期:2020-02-20;修回日期:2020-08-11;责任编辑:孙瑶
基金项目:"十三五"国家重点研发计划(No. 2017YFB0503500)

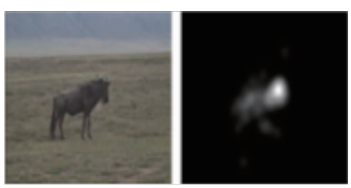
学习和确定,在小样本条件下,往往很难获取.

人类可以用眼睛轻易感知外界事物,比如墙上挂着的可能是一张照片而不是一辆汽车,沙发上伸展的可变形物体更有可能是猫而不是老虎. 机器视觉的发展一直在试图模拟人类视觉的认知过程,经过多年的探索和研究,模拟人类视觉注意力机制发展出的三种学习方式为研究者提供了进一步探索的空间(如图 1):

(1)基于生物学原理(人眼)的注意力预测机制[6](图 1(*a*)),旨在预测出图像中的注视点,这个注视点与眼动数据密切相关,有可能是自下而上与任务无关的,也可能是自上而下与当前任务相关的,其中自上而下有意识的注意力是有预定目的、依赖任务的、主动有意识地聚焦于某一对象的注意力.

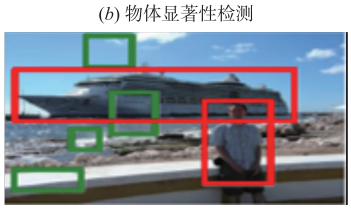
(2)基于机器学习的显著物体检测[7](图 1(*b*)),是一种自下而上的、通过计算一种或多种图像特征进而得到显著区域的方法,这种自下而上无意识的注意力称为基于显著性的注意力.

(3)基于对象窗口检测的度量方法[8](图 1(*c*)),是在众多滑动窗中选择需要对象所在的矩形窗.



(*a*) 注意力预测

(*b*) 物体显著性检测



(*c*) 滑动窗检测

图1　视觉注意力机制的三种表现形式

为了从图像中找到令人感兴趣的区域或者对象,建立在视觉注意力机制上的、作为计算机视觉应用处理第一步的视觉显著性检测技术,它更关注图像中感兴趣区域或有吸引力的对象,经过多年的发展,已取得了巨大成功. Chen 等人将眼睛注视流和语义流的输出融合到初始分割模块中以预测显著性[9]. Wang 等人在生物视觉相关知识指导下,通过在前馈传递中学习粗略的注视图,并利用长短记忆卷积网络从连续较浅层中结合多级特征逐个像素迭代来推断显著对象[10]. Chen 等人在自上而下的途径中嵌入反向注意模块,使用更深层次的输出来补充强调非目标区域内容以指导残差显著性学习[11]. 然而,当前大多数显著性检测技术依然无法应对背景复杂多变,区域显著但对象不突出以及检测对象过于单一的情况,特别是未知类别条件下显著区域(对象)和全局对象的空间属性以及语义解释还有待进一步提高.

场景上下文是一种可以将事物关联在一起学习的具有强大功能的信息载体,比如上下文信息可以有效分辨具有相似外观"云和烟"的语义特征描述,或是分割形成不规则区域之间的相互依赖性也可以得到解释等. 上下文在人类生活环境的统计属性里提供关键信息,有助于更快、更准确地解决感知推理任务,这种关于上下文信息在图像分割和对象识别中的重要性已经得到了广泛的认可. 在采用上下文解析图像内容的方法中,Zhou 等人采用金字塔池化模块将上下文信息编码为多个空间级别,从而可在每个空间位置上揭示更多的全局信息[12]. Wang 等人通过在视频中寻找潜在目标以及目标之间的时空关系来提取语义上下文[13]. Mottaghi 等人提出采用一种基于主题的可变形模型,它既可以捕获候选区域检测的局部上下文信息又可以在场景层级上捕获全局上下文信息[14]. 另外,以一种反向视角省略长时间学习过程的文献[15],对全局和局部的上下文语义标签信息使用非参数方法,可从全局、局部和像素三个角度更快更好地理解场景主题. 雷涛等人从空间邻域信息,直方图信息以及维度加权三个方面采用模糊聚类方法对图像分割做了详细阐述,以全方位学习场景内容[16]. 以上方法都是在特征学习基础上利用场景上下文信息获取特征更多属性以预测未知物体出现的可能性,未引入主观认知或决策信息,是监督式方法的学习和推理. 本文联合场景焦点对象(Focus)和上下文(Context)共同描述对象语义,将视觉焦点的挖掘转换为对象的显著性检测,融合视觉主观认知对场景进行分割,属于视觉认知和机器学习领域的交叉学科,这种给予未知场景部分先验信息的学习方式类似弱监督学习方法[17].

综合而言,本文将视觉注意力机制和场景类别结合起来,融合并发展基于视觉注意力机制的场景语义分割学习方法,主要分析两个影响场景语义分割的重要因素.

(1)类内关系:假设同类场景对象语义的一致性描述,即同类场景共享一种 Focus + Context 表征形式,并从焦点对象出现的区域推断其他图像中出现相似区域应该含有相似或相同的对象.

(2)焦点对象的显著性:从两个方面阐述焦点对象的视觉显著性,一是依据人类主观认知推断语义表征

下焦点对象承载的本体信息;二是焦点对象对应特征所在区域与剩余区域之间的视觉差异性.

2 场景语义表征

表征是知识在个体心理的反映和存在方式[18]. 它借鉴人类视觉的认知过程,在保持信息全局可见的同时对局部兴趣区域进行详细探索. 一般要求这种表征既能较好的描述全局信息,亦能表达场景所包含的局部信息. 遵循这条思路,构建场景类别的 Focus + Context 层级表征形式,在突出场景类别本体语义描述时,将用户关注的局部焦点与全局上下文对象描述在不同层级上,结合概率图模型[19]的生成式方法对场景进行语义区域分割,其中,Focus 表示同类场景中具有相同语义信息的焦点(显著对象),它表示类别标签中的本体信息,可以是场景中的一个或多个具有紧密连接关系的对象; Context 指除 Focus 以外的其他视觉对象,可以从不同视角、不同空间分布突出 Focus 位置和语义的重要性. 以海景图片为例,定义 Context = {大海,树木,天空,其它},Focus = {大海},图 2 描述了 Focus + Context 的语义表征形式,这种粗略的语义赋值,若能计算出各对象在图像中的位置,并将其关联起来,也足够表达一张海景所传递的主要内容.

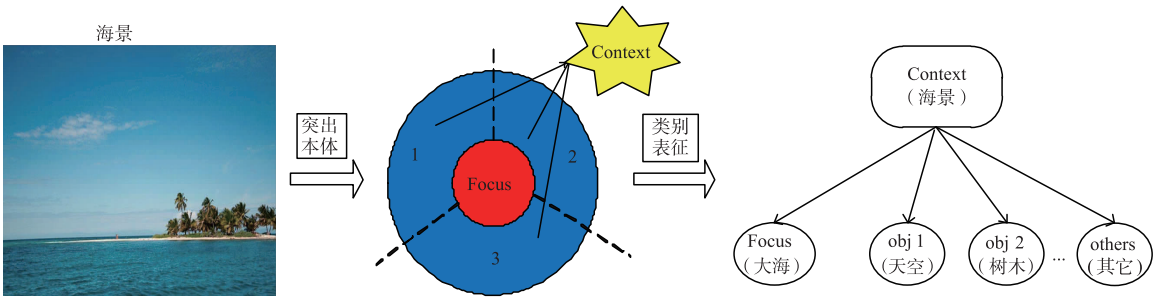


图2 Focus+Context 海景表征示意图

3 焦点对象突出的场景结构建模

语义信息是对场景内容的抽象描述,由于场景多区域、多对象的复杂组成结构,在定义场景类别本体语义描述后,需要自上而下将语义信息映射到图像特征上;在上层语义和底层符号的对接上,既要探索不同特征聚类后的语义信息,又要辨别多个语义对象之间的位置关系,最后在场景对象的空间属性上可视化对象语义描述. 遵循这一思路,首先抽象出场景类别表征信息,以概率主题模型为方法原型,利用底层特征位置的高斯分布描述,建立焦点突出的 Focus + Context 结构主题模型(图 3),即“特征-主题-对象-场景”;其次,利用同类图像共享相似特征的方法[20],分析场景类内分割区域是否含有相似或相同的对象,并以此推断其他图像中是否出现相似区域;最后,根据模型中焦点对象特征出现的高频特性以及不同主题特征在不同区域出现的不同频次,从场景全局和局部上定位 Focus + Context 语义表征对象,并在抽象层上突出以焦点所在位置为

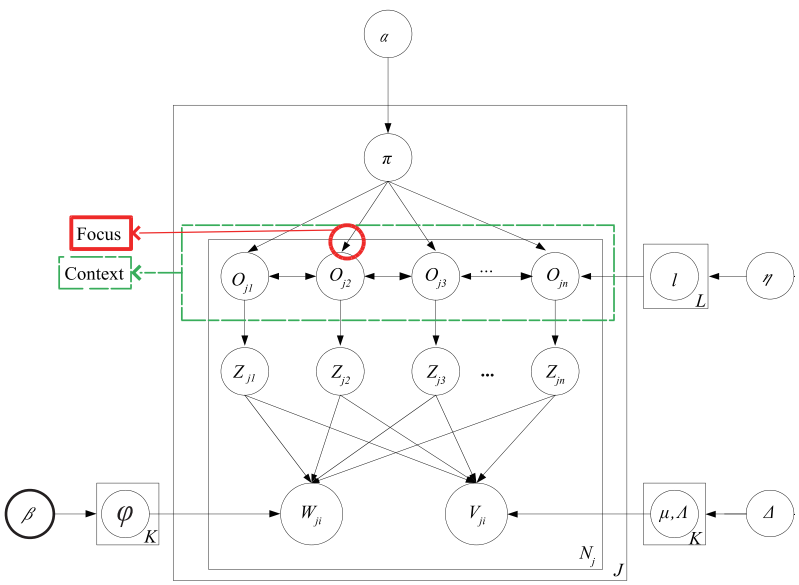


图3 Focus+Context场景图像结构主题模型

显著区域的联合语义分割.

3.1　模型参数的数学符号描述

Focus + Context 场景对象层级结构主题模型(图 3)中,$H=\{\alpha,\beta,\eta,\Delta\}$ 表示超参数集合;$\{O,Z,W,V\}$ 分别对应{对象,主题,特征,特征位置};l 表示场景类别,$l\in\{1,2,\cdots,L\}$,L 是场景类别数量;j 表示图像,$j\in\{1,2,\cdots,J\}$,J 表示场景图像数;i 表示图像特征个数,$i\in\{1,2,\cdots,N\}$,N 表示特征数目;k 表示主题,$k\in\{1,2,\cdots,K\}$,K 代表主题个数. 在先验信息的分配上,赋予对象 O 狄利克雷超参数 α,同时也赋予图像底层特征 W 狄利克雷超参数 β,以构成具有对称先验信息的层级主题模型. π 是多元分布对象的最大似然估计,φ 是特征的多元分布. 在已知类别标签 l 和 SIFT 特征 w_{ji} 条件下,假设场景与以一定比例生成的每一个对象的离散分布特征 w_{ji} 关联,对象 O_{ji} 语义先从标签集合 l 中取样,与对象对应的主题 Z_{ji} 也从离散分布的对象中取样得到,数学描述如式(1)和式(2).

$$O_{ji}\sim l_j \tag{1}$$

$$Z_{ji}\sim\pi_{o_{ji}},\ i=1,2,\cdots,N \tag{2}$$

由于二维数字图像可以用一个满秩的协方差矩阵描述,故采用正态分布来拟合特征的位置信息. 根据对象和主题 Z_{ji} 的一一对应关系,每一个描述主题特征的 w_{ji} 都独立取样,其对应位置 v_{ji} 的分布描述如式(3)和式(4)所示.

$$w_{ji}\sim\varphi_{z_{ji}} \tag{3}$$

$$v_{ji}\sim\text{Normal}(\mu_{z_{ji}},\Delta_{z_{ji}}) \tag{4}$$

3.2　对象和主题的联合学习

场景对象一般由多个主题构成,在模型学习中,对象和主题是两个隐变量,在特征按一定规则形成的场景中不直接显现,所以将两者作为一个整体进行推理学习,对象和主题联合学习的后验分布由式(5)描述.

$$\begin{aligned}&p(O_{ji},Z_{ji}\mid O_{\backslash ji},Z_{\backslash ji},w,v,H)\\&\propto p(O_{ji}\mid O_{\backslash ji},h)\cdot p(Z_{ji}\mid Z_{\backslash ji},O_{ji},h)\\&\cdot p(w_{ji}\mid w_{\backslash ji},z,h)\cdot p(v_{ji}\mid v_{\backslash ji},w_{ji},z,h)\end{aligned} \tag{5}$$

为了增加计算速度,使算法尽快收敛,设计 Blocked Gibbs 取样算法求解式(5),从成对出现的对象和主题的任意分配中取样,带入类别信息,自顶向下计算每个对象语义所对应的特征及特征出现的位置,直至遍历所有对象,取样算法如算法 1 所示.

算法 1　场景中对象和主题联合学习的取样算法描述

在给定的场景 l_j 序列中,对以下变量依次取样计算

输入:场景 j 的 SIFT 特征

输出:场景各对象 (o_{ji},z_{ji}) 所属特征类别以及其对应的位置信息

for $d\in\{1,2,\cdots,D\}$

(1)随机抽样任意序列 $\tau(\bullet)$ 的图像 SIFT 特征 $w_{ji},i\in\{1,2,\cdots,N\}$

(2)从图像特征中对对象和主题 (o_{ji},z_{ji}) 进行序列取样,针对成对出现的对象和主题,计算其似然函数

$$f_{l,k}(w_{ji}=w,v_{ji})=\left(\frac{C_{kw}+\frac{\lambda}{w}}{\sum_{w'}C_{kw'}+\lambda}\right)\bullet N(v_{ji};\mu_k,\Lambda_k)$$

对服从多元分布的对象和主题取样

$$p(o_{ji},z_{ji})\sim\frac{1}{Z_i}\sum_{d=1}^{D}\sum_{k=1}^{K}(M_{jl_d}+\lambda/L)\bullet\left(\frac{N_{lk}+\frac{\alpha}{k}}{\sum_{w'}N_{lk'}+\alpha}\right)\bullet f_{lk}(w_{ji}=w,v_{ji})\bullet\delta(o_{ji},d)\bullet\delta(z_{ji},k)$$

其中,$Z_i=\sum_{k=1}^{K}(N_{lk}+\alpha/k)\bullet f_{lk}(w_{ji},v_{ji})$

为对象和主题加入特征计数描述

$$\begin{aligned}&M_{jl_d}\leftarrow M_{jl_d}+1\quad d=o_{ji}\\&N_{l,k}\leftarrow N_{l,k}+1\quad k=z_{ji}\\&C_{kw}\leftarrow C_{kw}+1\quad w=w_{ji}\\&(\mu_k,\Lambda_k)\leftarrow(\mu_k,\Lambda_k)\oplus(v_{ji}\bullet\delta(o_{ji},d))\end{aligned}$$

(3)对主题进行取样,利用对称的狄利克雷先验分布计算对象标签 d 对应特征的概率

$$\begin{aligned}&\pi_{lk}\sim\text{Dir}(N_{l1}+\alpha/K,\cdots,N_{lK}+\alpha/K)\\&\phi_k\sim\text{Dir}(C_{k1}+\lambda/W,\cdots,C_{kW}+\lambda/W)\\&p(w_{ji},v_{ji}\mid o_j=d)=\sum_{k=1}^{K}\pi_{lk}\phi_k(w_{ji})\bullet N(v_{ji};\mu_k,\Lambda_k)\end{aligned}$$

end for

(4)更新对象语义标签 d 对应的位置

$$(\mu_k,\Lambda_k)\leftarrow(\mu_k,\Lambda_k)\oplus(v_{ji}\bullet\delta(o_{ji},d+1))$$

在算法 1 中,d 表示类别 l 下各对象的语义标签,比如在由{建筑,天空,街道,汽车}四个对象构成的街景中,l 表示街景,d 表示{建筑,天空,街道,汽车}中的某一个对象,$\tau(\cdot)$ 表示图像特征序列. 计算特征出现的概率是一个词频的统计过程,特别是在计算主题参数 Z_{ji} 时,主要采用条件概率的边缘计算后得到由有限个主题描述对象出现的概率,即每一个对象都由不同的主题组成且每个主题对应不同的位置,计算描述如算法 1 中的(3)所示,最后,逐一更新 d 值,直至收敛.

在场景语义表述和分类的方法中,基于对象语义特征的场景建模方法一直都备受关注,因为对象的描述直接反映了场景的全局信息,同时场景分割区域的类别信息可以直接输入到图像库进行管理和索引,并且这种方法也比较容易推广到多类场景的分类问题中.

4　实验分析与讨论

实验参数设置:利用最大稳定极值区域检测算法 MSER (Maximally Stable Extremal Regions)[21] 和局部特征提取方法 SIFT(Scale-Invariant Feature Transform)[22] 计算图像特征,每一张词典大小设为 $W=300$,类别信息服从标准均匀分布,代表聚类速率的超参数 $\alpha=5$,描述特征位置信息的高斯分布先验 $\Delta=6$,满足多元分布形式的狄利克雷先验 $\eta_k=\text{Dir}(\lambda)$,$\lambda=0.1$. 依据对象的

语义标注信息,将每一个对象和其对应的主题作为一个整体,即 $p(o,z)$,采用 Blocked Gibbs 取样算法进行计算,计算迭代次数的设置同参考文献[23].

实验对象:在 2 个图像集 PASCAL VOC 2012 ACTION 以及 LabelMe[24] 上展开场景语义分割实验.

4.1 PASCAL VOC 2012 ACTION 实验分析与讨论

以 PASCAL VOC 2012 Challenge Action 作为实验数据,该数据子集包含 10 个动作类(Running, Jumping, Phoning, Playing Music Instrument, Reading, Riding-bike, Riding-horse, Taking-photo, Using-computer, Walking),由于图像数据库还在扩增中,选择样本数量超过 50 的类别作为实验数据,对模型进行训练和测试,示例如图 4 所示.

表 1 是实验中涉及到的有效样本类别、数目、相关实验参数及分类结果. 每类图片训练数目占图片总数的一半,码本大小统一设为 300;在语义表征中,通常选择常用对象名称描述,比如骑马类场景,定义常用的四个对象{马,人,树,地面},这四个常用对象能够满足骑马类场景的内容描述,如果继续在骑马类场景中增加更多的对象,如天空、山峦等,不仅会带来计算复杂度的增加,还会产生更多的语义歧义性. 因为对象数目划分的越多,涉及到底层特征聚类选择的准确度会降低,从而导致图像语义特征分类更模糊,使得由下至上的对象被正确识别和定位的概率减少,所有这些因素在整个模型的学习中都是环环紧扣. 此外,分类准确率的好坏还受训练数目多少以及焦点在图像空间中占比的影响. 在表 1 中,Using-computer 和 Taking-photo 这两类动作的分类准确率也进一步证明了这个结论.

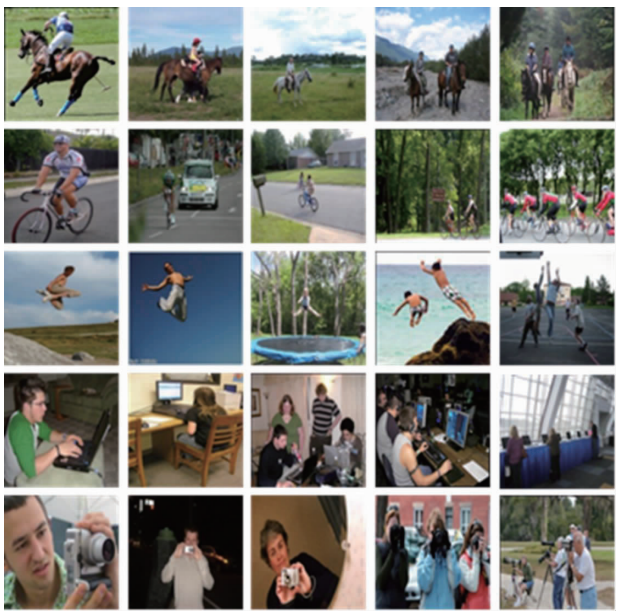


图4 PASCAL VOC2012 Action 样本示例

表 1 PASCAL VOC **2012** Action **样本参数及分类准确率**

类别名称	图片数目	训练数目	码本大小	Focus + Context 语义表征	分类准确率
Riding-horse	116	60	300	{马,人,树,地面}	88.24%
Riding-bike	51	25	300	{车,人,道路,绿树}	65.22%
Jumping	65	35	300	{人,天空,地面}	55.26%
Using-computer	81	40	300	{电脑,人,桌}	22.32%
Taking-photo	90	45	300	{相机,人,其它}	18.75%

图 5 反应了小样本数据下 5 类动作的分类正确率,采用焦点对象最大似然函数的后验概率值作为分类依据,直方图上的数据表示此类场景中焦点对象被正确分类的准确率. 在计算中,Taking-photo 和 Using-computer 的分类准确率很低,除了上一段描述的部分因素外,还有一个主要原因在于图像类内相似性差,以致很难计算出焦点对象的共同特征.

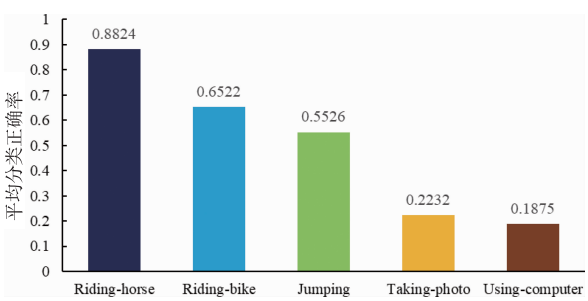
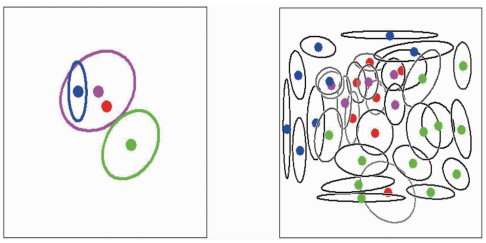


图5 PASCAL VOC 2012 Action 焦点对象分类正确率

图 6 是 60 张 Riding-horse 图像经训练后得到的反应整体属性的全局对象相对位置坐标及对象组成主题的局部位置分布图,部分原图可参见图 4 第一行;定义 Context = {马,人,树,地面},分别用粉色、红色、蓝色和绿色表示,不同颜色描述的语义区域共同构成场景上下文信息. 由于场景是由多个兴趣区域组成,对描述同一对象的区域取均值和方差作为对象可能出现位置的判断标准,如图 6(a)中圆点及其对应的椭圆. 另外,从抽象层描述角度而言,骑马类场景主要反应了人与马之间的共同合作,推断作为此类场景的焦点对象,即 Focus = {(人 & 马)},对应全局坐标图 6(a)中粉色中心点和椭圆. 椭圆邻接或重叠的地方反应了场景多对象之间的相互解释性. 这种特性在图 6(b)中有更为直观的视觉表现,粉色和红色的椭圆密不可分,反应了人

和马所在区域的密不可分,这也间接体现了对象的相互依赖性. 图 6(*b*)则是对图 6(*a*)一个更深入的特征位置分布的描述,通过计算兴趣区域后验概率的均值和方差得到.



(*a*) 全局对象相对位置坐标　　(*b*) 不同对象局部位置估计

图6　Riding-horse全局对象及其组成主题估计位置分布图

图 7 是 Riding-horse 测试集场景语义分割与对象检测的可视图. 图中第 1 行表示原图,第 2 行是第一行场景的 Context 语义分割图,第 3 行是在第 2 行基础上对不同对象可能出现的位置描述,第 4 行是检测到的最能反应类别信息的焦点对象.

从第 2 行到第 4 行,整体描述了一个 Focus + Context 的场景内容表达,图中的语义上下文在第 2 行由不同颜色描述,每一种颜色代表一种语义标签,空间上下文则反应在第 2 行和第 3 行的对象区域位置描述上. 另外,从第 4 行对象主题的可视化描述上,每一张图像的焦点对象都被检测到,这也间接体现了弱监督学习方式下语义对象识别的有效性.

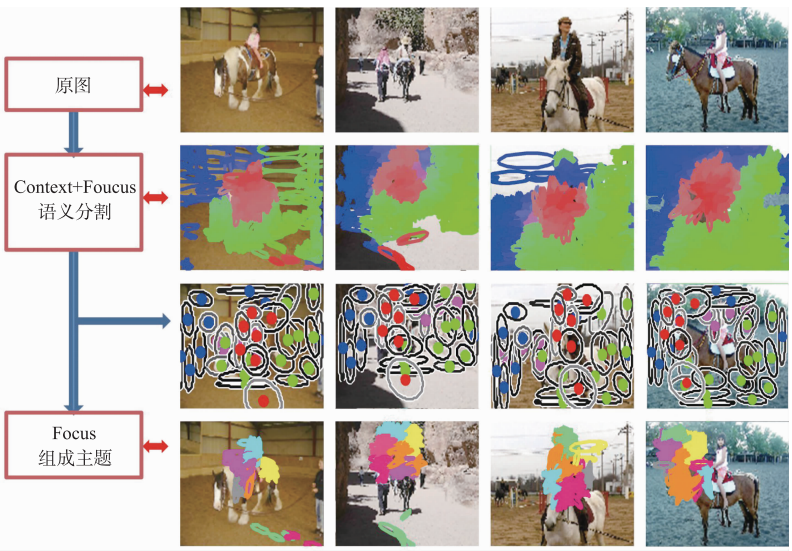


图7　Riding-horse类的Focus+Context 描述图

4.2　LabelMe 实验分析与讨论

表 2 描述 LabelMe 各类图像的平均分类准确率,其中用于训练的图像数占总数的一半,第 5 列给出对象名称作为已知信息,排在最前面的加粗符代表此类场景的焦点,它也是第 6 列采用最大后验概率分类的主要依据.

图 8 描述已知街景类别的街景分割示意图. Context = {街道,汽车,建筑,树木},Focus = {街道},蓝色代表街道,红色代表汽车,粉色代表建筑,绿色代表树木. 每一行图像的含义同图 6 的描述. 不一样的地方在第 2 行,因为街道和树木的分界非常不明显,并且建筑占据的面积远远大于树木,这些因素导致算法在做分割时将树木和道路作为一个整体划分.

表 2　LabelMe 样本参数及分类准确率

类别名称	图片数目	训练数目	码本大小	Focus + Context 语义表征	分类准确率
coast	360	180	300	{大海,天空,海岸}	89.12%
forest	328	164	300	{树,土壤}	98.32%
highway	260	130	300	{道路,天空,树木}	88.64%
insidecity	308	154	300	{建筑,天空,道路}	86.15%
mountain	374	187	300	{山石,天空,水面}	87.54%
opencountry	410	205	300	{农田,房屋,汽车,道路}	80.37%
street	292	146	300	{街道,汽车,建筑,树木}	79.96%
tallbuilding	356	178	300	{高楼,天空,道路}	88.22%

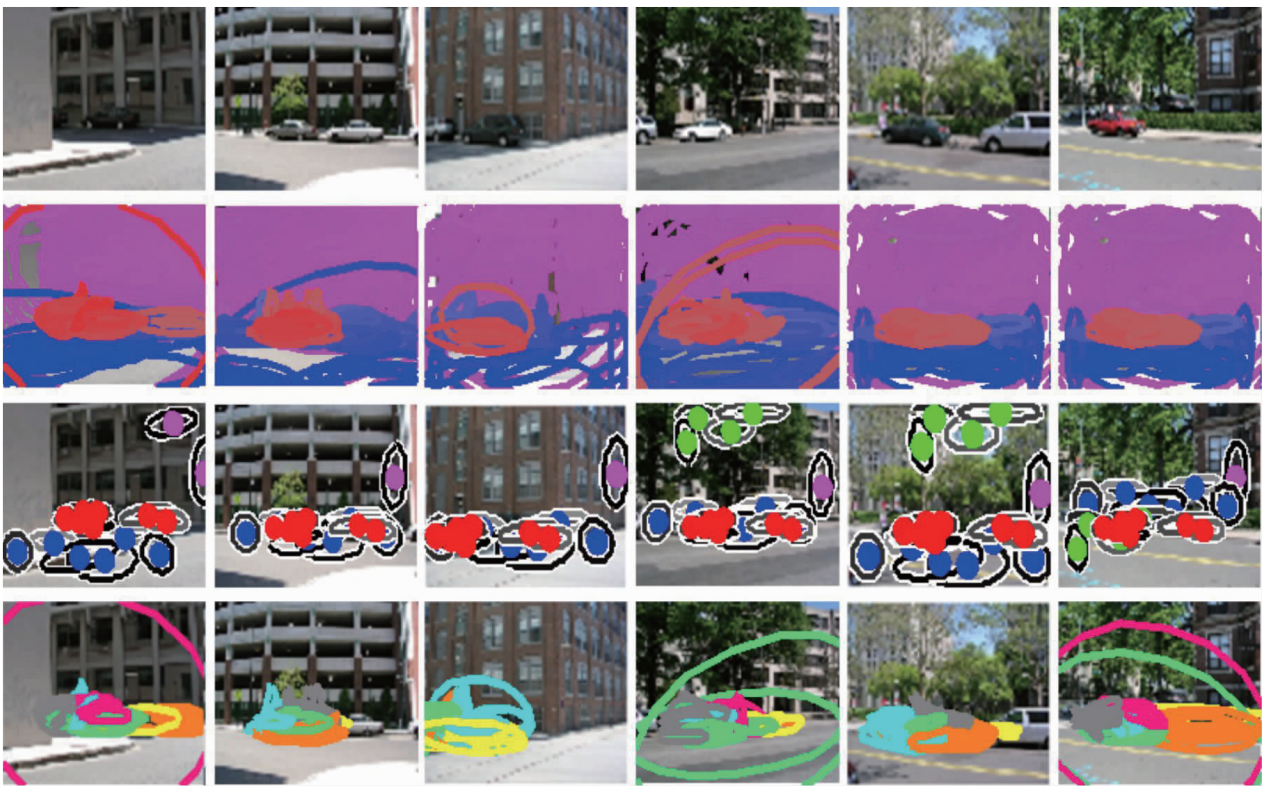


图8 街景类Focus+Context描述示意图

图 9 是与相似方法在 LabelMe 上分类准确率的比较结果,其中 sLDA(Supervised Latent Dirichlet Allocation)是同时用于分类和标注的主题模型[25];MMLDA(Max-Margin Latent Dirichlet Allocation)将判别式图像分类方法融入到监督式的 LDA(Latent Dirichlet Allocation)模型中,利用最大边界变化值对主题模型表示的图像向量进行分类[26];supDocNADE(Supervised Extension of DocNADE)[27,28]将多种属性的数据(文本和图像)一起输入到主题模型进行训练,利用 DocNADE(Document Neural Autoregressive Distribution Estimator)[29]神经网络自回归的文本分析方法,在测试阶段输入图像数据进行学习和推理;DiscLDA(Discriminative Variation on Latent Dirichlet Allocation)系列模型是 Niu 等人提出的融入辨别方法的监督式主题模型[23].

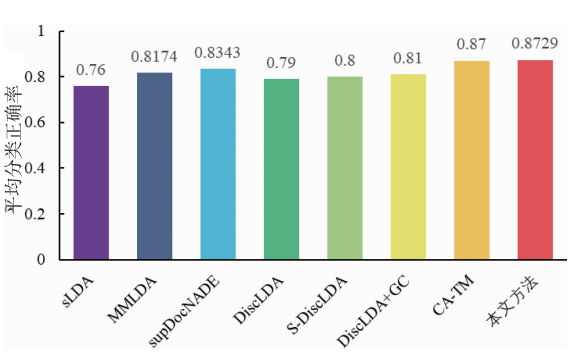


图9 同类方法在LabelMe上的分类性能比较

在与同类方法的比较中,本文提出的方法无论在形式和内容上都有很大的不同,分类准确率的提高主要依赖于明确焦点对象的先验语义信息在改进主题模型中的有效计算.

5 结论

本文提出了 Focus + Context 语义表征方法,通过建模融合类别信息的底层特征主题层级模型展开,依据人类对事物的认知,提出 Focus + Context 的语义表征方法,从全局和局部两个角度推理对象语义类别信息;采用数据自驱动的非参数学习方法,联合图像底层视觉特征,构建自上而下的场景分割主题模型,以在抽象层上实现局部焦点对象突出的、区别于其他区域的场景对象语义层级理解. 此外,本文还在 PASCAL VOC 2012 Action 以及 LabelMe 图像集上验证了提出方法在场景分割上的有效性和依赖焦点对象分类性能上的优越性.

为了进一步完善 Focus + Context 类别表征的图像分割理论,将认知心理学与图像分割算法中的相关理论相结合,拓展基于概率主题建模图像分割算法的应用领域,给出了面向图像分割的下一步发展方向:

(1)概率主题模型的学习和推理. 根据应用需求给定模型的表达后,如何从给定数据中学习模型的参数以及如何快速、准确的进行模型推断都是至关重要的,

不同的方法在学习效率、准确性上各有差异. 研究快速、准确、有效的学习和推理算法,是实现概率图模型实时计算的一个重要瓶颈.

(2)深度学习与概率图模型的联合学习. 因为概率图模型能够从数学理论上很好的解释事物本质间的联系,而具有参数自学习功能的深度学习正好可以弥补概率图模型不能很好进行网络化学习的缺陷,所以联合两者进行图像特征的描述、主题的挖掘和场景对象的识别是图像分割研究方向的又一个重要工作.

参考文献

[1] BADUEA C, GUIDOLINIA R, RAPHAEL C, et al. Self-driving Cars: A Survey[EB/OL]. https://arxiv. org/abs/1901. 04407v1, 2019-01-14.

[2] PORCHERON M, FISCHER J E, REEVES S, et al. Voice interfaces in everyday life[A]. Proceedings of the CHI Conference on Human Factors in Computing Systems[C]. Montreal QC, Canada: ACM, 2018. 1 – 12.

[3] WANG Q, ZHANG L, BERTINETTO L. Fast online object tracking and segmentation: a unifying approach[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition [C]. Long Beach, USA: IEEE, 2019. 1328 – 1338.

[4] 田艳玲,张维桐,张锲石,等. 图像场景分类技术综述[J]. 电子学报,2019,47(4):915 – 926.
TIAN Yan-ling, ZHANG Wei-tong, ZHANG Qie-shi, et al. Review on image scene classification technology[J]. Acta Electronica Sinica, 2019, 47(4): 915 – 926. (in Chinese)

[5] 罗会兰,张云. 基于深度网络的图像语义分割综述[J]. 电子学报,2019,47(10):2211 – 2220.
LUO Hui-lan, ZHANG Yun. A survey of image semantic segmentation based on deep network[J]. Acta Electronica Sinica, 2019, 47(10): 2211 – 2220. (in Chinese)

[6] HOCK H S, GREGORY P G, WHITEHURST R. Contextual relations: The influence of familiarity, physical plausibility, and belongingness[J]. Attention Perception & Psychophysics, 1974, 16(1): 4 – 8.

[7] 钱晓亮,白臻,陈渊,等. 协同视觉显著性检测方法综述[J]. 电子学报,2019,47(6):1352 – 1365.
QIAN Xiao-liang, BAI Zhen, CHEN Yuan, et al. A review of co-saliency detection[J]. Acta Electronica Sinica, 2019, 47(6): 1352 – 1365. (in Chinese)

[8] LIU T, YUAN Z, SUN J, et al. Learning to detect a salient object[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(2): 353 – 367.

[9] CHEN X W, ZHENG A L, LI J, et al. Look, perceive and segment: finding the salient objects in images via two-stream fixation semantic CNNs[A]. IEEE International Conference on Computer Vision[C]. Venice, Italy: IEEE, 2017. 1050 – 1058.

[10] WANG W G, SHEN J B, DONG X P, et al. Salient object detection driven by fixation prediction[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, Utah: IEEE, 2018. 1711 – 1720.

[11] CHEN S H, TAN X L, WANG B, et al. Reverse attention for salient object detection[A]. 2018 European Conference on Computer Vision[C]. Munich, Germany: Springer, 2018. 236 – 252.

[12] ZHOU Y Z, SUN X Y, ZHA Z J, et al. Context-reinforced semantic segmentation[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Long Beach, CA, USA: IEEE, 2019. 4041 – 4050.

[13] WANG T H. Context propagation from proposals for semantic video object segmentation[A]. The 25th IEEE International Conference on Image Processing[C]. Athens: IEEE, 2018. 256 – 260.

[14] MOTTAGHI R, CHEN X J, LIU X B, et al. The role of context for object detection and semantic segmentation in the wild[A]. 2014 IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, OH: IEEE, 2014. 891 – 898.

[15] BANSAL A, SHEIKHY, RAMANAN D. Shapes and context: In-the-wild image synthesis & manipulation[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Long Beach, USA: IEEE, 2019. 2312 – 2321.

[16] 雷涛,张肖,加小红,等. 基于模糊聚类的图像分割研究进展[J]. 电子学报,2019,47(8):1776 – 1791.
LEI Tao, ZHANG Xiao, JIA Xiao-hong, et al. Research progress on image segmentation based on fuzzy clustering[J]. Acta Electronica Sinica, 2019, 47(8): 1776 – 1791. (in Chinese)

[17] ZHOU Z H. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44 – 53.

[18] 杨盛春. 知识表征研究述评[J]. 图书情报导刊,2012,22(19):145 – 147.
YANG Sheng-chun. A review of the research on knowledge tepresentation[J]. Tech Information Development & Economy, 2012, 22(19): 145 – 147. (in Chinese)

[19] BLEI D M, NG A Y, JORDAN M I, et al. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993 – 1022.

[20] WU L, LIU Q. Weakly supervised object co-localization via zharing parts based on a joint bayesian model[J]. Symmetry, 2018, 10(5): 142.

[21] MATAS J, CHUMO, URBAN M, et al. Robust wide base-

line stereo from maximally stable extremal regions[J]. Image and Vision Computing,2004,22(10):761 – 767.
[22] VEDALDI A,FULKERSON B. An Open and Portable Library of Computer Vision Algorithms[EB/OL]. http://www. vlfeat. org/,2018.
[23] NIU Z X,HUA G,GAO X B,et al. Spatial-discLDA for visual recognition[A]. 2011 IEEE Conference on Computer Vision and Pattern Recognition[C]. Providence,RI:IEEE,2011. 1769 – 1776.
[24] LabelMe. LabelMe 数据库[DB/OL]. http://labelme2.csail. mit. edu/Release3. 0/index. php,2020.
[25] WANG C,BLEI D M,LI F F. Simultaneous image classification and annotation[A]. 2009 IEEE Conference on Computer Vision and Pattern Recognition[C]. Miami,FL:IEEE,2009. 1903 – 1910.
[26] WANG Y,Mori G. Max-margin latent dirichlet allocation for image classification and annotation[J]. Lecture Notes in Computer Science,2011,1674(1):39 – 48.
[27] ZHENG Y,ZHANG Y J,LAROCHELLE H. Topic modeling of multimodal data:An autoregressive approach[A]. 2014 IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus,OH:IEEE,2014. 1370 – 1377.
[28] ZHENG Y,ZHANG Y J,LAROCHELLE H. A deep and autoregressive approach for topic modeling of multimodal data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,38(6):1056 – 1069.
[29] LAROCHELLE H,LAULY S. A neural autoregressive topic model[A]. Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2[C]. New York,USA:ACM,2012. 2708 – 2716.

作者简介

吴 绿 女,1983 年 1 月出生,湖北潜江人. 武汉理工大学通信与信息系统专业博士,现任武汉理工大学信息学院讲师. 主要研究方向为机器视觉、模式识别.
E-mail:wulv@ whut. edu. cn

张馨月 女,1994 年 7 月出生,吉林双辽人. 武汉大学资源与环境科学学院博士研究生.