一种双层条件随机场的场景解析方法

李艳丽 周 忠 吴 威

(北京航空航天大学虚拟现实技术与系统国家重点实验室 北京 100191)

摘 要 现有的场景解析方法主要依赖于先验模型,由于先验模型难于全面表示物体的各种细节部分,使得场景解析后的物体不够精细.针对这个问题,该文引入了局部颜色模型,提出了一种结合先验和局部颜色模型的双层条件随机场的场景解析方法.首先以超像素为结点构建一个条件随机场,根据颜色、梯度、纹理和几何等外观特征训练出的先验模型粗略解析场景,进而提取场景中每个物体的局部颜色模型;然后构建一个以像素点为结点的条件随机场,通过 EM(Expectation-Maximization)迭代法更新物体的局部颜色模型来指导优化场景解析.实验结果表明,相比于以往单纯利用先验模型的场景解析方法,该方法能有效地保持场景细节、提高解析精度.

关键词 场景解析;先验模型;局部颜色模型;条件随机场;EM 迭代

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2013.01898

Scene Parsing Based on A Two-Level Conditional Random Field

LI Yan-Li ZHOU Zhong WU Wei

(*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191)

Abstract Recently, there have been increasing interests in semantic scene parsing, which are mainly based on prior models. However, as prior models only consider the common merits within objects, they ignore the inner color coherence lied in the single object and suffer from detail loss. In this paper, we utilize color cue as the inner information, and present an approach to incorporating local color model with prior models under a two-level conditional random field to preserve scene parsing details. More specially, objects of the scene are first roughly extracted using prior models within a superpixel-based conditional random field, in which prior models are acquired by supervised learning using color, gradient, texture and geometric cues, and then a local color model is built for each object based on the initial parsing result. Combining the local color model and prior models, we employ an EM(Expectation-Maximization) scheme for iterative refinement within a pixel-based conditional random field. Experimental evaluations with state-of-the-art methods verify that our approach is able to preserve details and achieve better performance.

Keywords scene parsing; prior model; local color model; conditional random field; Expectation-Maximization

1 引 言

场景解析是识别和分割图像内各个物体的技术,例如从街景图像中分解出天空、道路、行人和车辆等,广泛应用于内容检索、智能导航和视频监控等领域,是计算机视觉的研究热点问题之一.

场景解析法往往建立在马尔可夫随机场下以保证邻居结点的标识一致性. Lafferty 等人[1]在 2001 年提出的条件随机场解决了其他马尔可夫模型难以

收稿日期:2011-10-17;最终修改稿收到日期:2013-07-13.本课题得到国家自然科学基金(611701880)、国家"八六三"高技术研究发展计划项目基金(2012AA011803)及教育部博士点基金(20121102130004)资助.**李艳丽**,女,1982 年生,博士研究生,主要研究方向为图像处理、模式识别. E-mail: liyanli725@gmail.com. **周 忠**,男,1978 年生,博士,副教授,中国计算机学会(CCF)高级会员,主要研究方向为增强虚拟环境和自然现象建模. **吴 威**,男,1961 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为虚拟现实和分布式系统.

避免的标识偏置问题. 2004 年，He 等人[2]首次实现了条件随机场下的场景解析，此后研究人员提出了各种基于像素点或超像素的条件随机场下的场景解析法. 然而，以像素点为计算单元的方法需要获取大量特征，存在计算效率低的问题[2-3]；以超像素为计算单元的方法受超像素的影响无法完整保留物体边缘[4-5]. 为了更精细地解析场景，最近几年研究人员提出了多层超像素下的场景解析方法[6-9]以及结合像素点和超像素的高阶条件随机场下的场景解析方法[10-12]. 前者将场景解析描述为多层超像素的识别和组合问题. 由于多层超像素可获取更多分割粒度的聚类块，此类方法有效地避免了单层超像素的欠分割问题. 后者在二阶条件随机场上增加了一个由像素点和超像素搭建的高阶项，从而将其扩展为高阶条件随机场. 由于此类方法中的高阶项是对像素点的软聚类，相比于基于超像素的方法对像素点的硬聚类，解析出的物体边缘更精细.

上述方法都是依赖先验信息来分离场景中的物体，即从大量训练数据统计出的先验模型指导场景解析. 然而，先验模型只反映了物体间的共性特征，难以全面表达物体内的细节特性. 我们注意到在图像分割领域还存在交互式的前景分割算法[13-15]. 交互式分割方法根据物体颜色具有内敛性的特点，依赖局部颜色模型来分离前景和背景. 局部颜色模型代表了图像中物体内特有的属性，跟先验模型具有互补性. 与上述方法不同，本文从另一个角度来提高场景解析精度，即将局部颜色模型引入到场景解析中，提出结合先验和局部颜色模型的场景解析方法. 考虑到计算效率，首先以超像素为计算单元获取先验模型来粗略地解析场景，然后以像素为计算单元获取局部颜色模型进一步精细地解析场景. 本文的主要贡献点在于提出了一种结合像素点和超像素的双层条件随机场，该条件随机场利用先验模型和局部颜色模型的互补性来保留场景细节. 将该方法与基于像素点的方法[3]、基于超像素的方法[5]以及结合像素点和超像素的方法[10]进行了实验比较，结果表明本文方法引入局部颜色模型后使得场景解析结果更精细.

本文第 2 节介绍场景解析方面的相关工作；第 3 节描述引入局部颜色模型的双层条件随机场；第 4 节给出双层条件随机场下的场景解析；第 5 节为实验结果和比较讨论；最后进行总结.

2　相关工作

场景解析问题早在 20 世纪 70 年代就已提出[16]，直到近年随着底层算法的成熟才成为计算机视觉的研究热点. 现有的场景解析方法基本上可分为基于像素点的方法、基于超像素的方法以及结合像素点和超像素的方法.

基于像素点的方法以像素点为计算单元. He 等人[2]最早提出了条件随机场下基于像素点的场景解析法，首先由神经网络训练像素点的颜色特征获取先验模型，然后求解一个条件随机场下的全局能量函数完成场景解析. 由于像素点的局部特征不包涵物体的全局统计信息，此类方法的准确性较差. 另外，以像素点为计算结点的计算量大，此类方法的效率也偏低. 基于超像素的方法以超像素为计算单元，其中的超像素是根据图像底层颜色信息聚类的像素块. Yang 等人[4]最早提出了基于超像素的条件随机场下的场景解析方法. 相比于基于像素点的方法，基于超像素的方法计算结点少，因此效率较高. 然而每个超像素算法都存在欠分割或过度分割问题. 欠分割无法完整保留物体边缘，而过度分割提取不到有价值的全局特征，此类方法解析出的场景比较粗糙.

为了完整的保留物体轮廓，Caroline 等人[6]、Stephen 等人[7]、Kumar 等人[8]和 Cheny 等人[9]提出了多层超像素下的场景解析方法. 其中，Caroline 等人[6]用超像素算法将图像分割为 18 层，以超像素的交叉区域为计算单元在一个条件随机场下解析场景；Stephen 等人[7]获取了 3 层超像素空间，将场景解析描述为分割块选取和能量优化两个问题，用梯度下降法迭代优化；Kumar 等人[8]则将场景解析描述为超像素选取的整形规划问题；类似的，Cheny 等人[9]提出了一种增加上下文约束的整形规划式的场景解析方法. 此类方法能精细解析场景主要建立在两个假设基础上：(1) 算法能选取最佳分割块组合；(2) 该组合内的分割块边缘能完整表示物体轮廓. 然而这两种假设不一定总成立.

为了避免超像素的欠分割问题，近年研究人员提出了结合像素点和超像素的高阶条件随机场下的场景解析方法[10-12]. Kohli 等人[10]最早把高阶条件随机场引入到场景解析中，通过将每个分割块内的像素点聚类成一个高阶项，求解一个以像素点为计算单元的高阶能量来解析场景. 基于该工作，Kohli 等人[11]又提出了一种泛化的高阶条件随机场，任何一种高阶条件随机场下的场景解析方法都可看成其参数调整后的特例. 在此基础上，他们又提出增加共存性约束的高阶条件随机场下的场景解析方法[12]. 相比于基于超像素的方法对像素点的硬聚类，此类方法对像素点进行了软聚类，因此使得物体边缘更

平滑,解析的场景更精细.

本文方法也属于结合像素点和超像素的方法,跟 Kohili 法[10-12]的不同在于:(1) Kohili 法建立在一个高阶条件随机场下,本文方法建立在一个双层的二阶条件随机场下;(2) Kohili 法的贡献点在于引入了像素点和超像素之间的高阶项以避免超像素导致的欠分割问题,本文贡献点在于引入了局部颜色模型以避免先验模型导致的统计特性偏差问题.

3 引入局部颜色模型的双层条件随机场

3.1 场景解析问题的数学描述

给定一幅图像 I,场景解析是对其中的像素点或超像素 $X=\{i\}$ 做自动标识 $L=\{l_i\}$ 的问题.其中,l_i 代表了标识类别,如天空、道路、树木、行人和车辆等等.由于图像中存在空间马尔可夫性,在以像素点或超像素为结点构建的图结构 $G=\langle X,Y\rangle$(Y 为相邻结点组成的边集合)中,现有方法一般将结点标识定义为一个二阶条件随机场下的能量函数[2-5]:

$$E(L)=\sum_{i\in X}\phi(l_i\mid I,\theta_\phi)+\lambda\sum_{(i,j)\in Y}\varphi(l_i,l_j\mid I,\theta_\varphi)\quad(1)$$

其中,$\phi(\cdot)$为数据项,该项根据结点特征计算单结点的标识误差,从而使得结点赋值最佳标识;$\varphi(\cdot,\cdot)$为平滑项,该项根据结点特征差计算相邻结点的标识误差,从而使得相邻结点赋值一致性的标识;θ_ϕ 和 θ_φ 是数据项和平滑项中的参数;λ 为权重,用来平衡数据项和平滑项的比重,是经过实验分析得到视觉满意结果下选定的经验值.场景解析结果为最小能量下的标识:

$$L^*=\arg\min_L(E(L)).$$

3.2 双层条件随机场的建立

基于上小节描述的单层条件随机场,我们设计了一个两层条件随机场的方法来解析场景.首先以超像素为结点构建一个条件随机场,在先验模型约束下初步解析场景;然后以像素点为结点再构建一个条件随机场,在先验和局部颜色模型联合约束下对场景再次解析.

考虑到计算效率,初次场景解析以超像素为计算结点.图 1 为一个建立在超像素上的条件随机场示意图,其中的边连接了两个相邻的超像素.由于特定物体往往出现在特定场景下,本阶段从两个角度对场景解析,首先根据图像的全局特征识别出场景类别,即判断图像是属于街景还是室内场景等,然后在场景识别基础上根据分割块的外观特征解析出场景中的物体类别,即判断每个分割块是属于天空还是道路等等.相应的,所依赖的先验模型包括全局场景先验模型和局部分割块先验模型.两者都是以监督学习方式从先验数据训练获取的,在整个场景解析过程中保持不变.

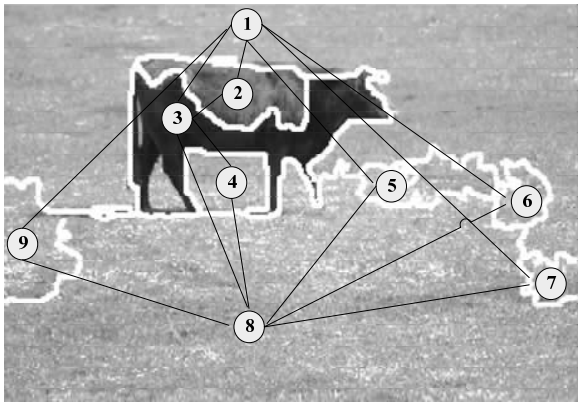


图 1 以超像素为结点的条件随机场示意图

在全局场景和局部分割块先验模型约束下的条件随机场能量函数定义为

$$E(L^s,S)=\sum_{i\in X_s}(\Psi(S\mid I,\theta_\psi^s)+\lambda_k\phi_s(l_i^s,S\mid I,\theta_\phi^s))+\lambda_s\sum_{(i,j)\in Y_s}\varphi_s(l_i^s,l_j^s,S\mid I,\theta_\varphi^s)\quad(2)$$

其中,X_S 为超像素结点集合,Y_S 为边集合,S 为场景标识,$L^s=\{l_i^s\}$ 为分割块标识,$\Psi(\cdot)$为全局场景先验模型约束下的数据项,$\phi_s(\cdot)$为局部分割块先验模型约束下的数据项,$\varphi_s(\cdot,\cdot)$为相邻分割块间的平滑项,θ_ψ^s、θ_ϕ^s 和 θ_φ^s 分别为全局场景先验模型、局部分割块先验模型和平滑项中的参数,λ_k 和 λ_s 为权重.场景解析结果为最小能量下的联合标识:

$$((L^s)^*,S^*)=\arg\min_{(L^s,S)}E(L^s,S).$$

由于超像素导致的欠分割以及物体外观多样性,该解析结果往往比较粗糙.在此基础上,我们引入局部颜色模型,以像素点为计算结点对场景进一步精细解析.图 2 为一个建立在像素点上的条件随机场的示意图.其中,条件随机场的结点为像素点,边连接了四连通的邻居结点.

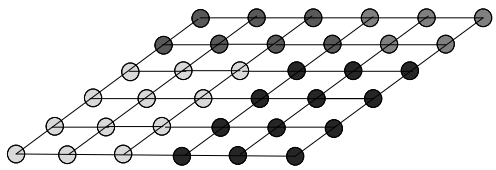


图 2 以像素点为结点的条件随机场示意图

局部颜色模型是根据当前解析结果统计出来的,在解析过程中可迭代更新,其初始值来源于初次解析结果.在先验和局部颜色模型联合约束下的条件随机场的能量函数为

$$E(L^p)=\sum_{i\in X_p}(\phi_s(l_i^p\mid I)+\lambda_c\phi_p(l_i^p\mid I,\theta_\phi^p))+\lambda_p\sum_{(i,j)\in Y_p}\varphi_p(l_i^p,l_j^p\mid I)\quad(3)$$

其中，X_p为像素结点集合，Y_p为边集合，$L^p=\{l_i^p\}$为像素点标识；$\phi_s(\cdot)$是根据初次解析结果得到的先验数据项，$\phi_p(\cdot)$为局部颜色模型约束下的数据项，$\varphi_p(\cdot,\cdot)$为相邻结点间的平滑项；λ_p和λ_c为权重. 最终的场景解析结果为该条件随机场下最小能量对应的标识：

$$(L^p)^*=\arg\min_{L^p}E(L^p).$$

4　基于双层条件随机场的场景解析

我们在上一节引入了局部颜色模型，建立了双层条件随机场，在此基础上设计一个具体的三步骤场景解析法. 首先，根据全局场景先验模型识别场景类别；其次，以超像素为计算结点依赖先验模型完成场景的初次解析；最后，提取物体的局部颜色模型，在以像素点为计算结点、局部颜色模型约束下的条件随机场中用 EM 迭代法[17]优化解析.

4.1　场景全局语义识别

全局场景先验模型结合了 3 种全局特征：Gist 特征[18]、颜色直方图和图像的缩放图. 其中，Gist 特征是一种梯度直方图描述子，通过将原图切分为 4×4 个小图，统计每个小图 3 通道(RGB 空间)上 20 维梯度直方图获取一个 960 维描述子；颜色直方图是一种全局颜色直方图描述子，统计原图像 3 通道(RGB 空间) 上 8 维颜色直方图获取一个 24 维描述子；缩放图是原图 16×16 尺寸 3 通道(RGB 空间)的小图，一共 768 维. 因此从每幅图像可提取一个 1752 维的全局特征描述子.

在训练阶段，提取图像的全局特征描述子用随机森林分类器[19]进行训练，将获取的分类器模型(模型参数为θ_ψ^s)作为全局场景先验模型；在解析阶段，同样提取图像的全局特征，通过全局场景先验模型预测属于每类场景的概率 $p(S\mid I,\theta_\psi^s)$. 定义式(2)中的场景数据项为

$$\Psi(S\mid I,\theta_\psi^s)=-\log p(S\mid I,\theta_\psi^s).$$

为了验证该步骤的有效性，我们对 MSRCv2 数据集进行了测试. 该数据集包括 20 类场景，每类场景大约有 30 幅图像，一共是 591 幅图像. 我们从每类中随机选取 6 幅图像，一共 120 幅图像作为测试数据，其它 471 幅用作训练数据来获取分类器模型. 假设图像 I 的前 a 个最大可能标识下的场景标识集合为 $ST(a,I)$，则对于测试集 $ISet$，定义其场景识别率为前 a 个最大可能标识中存在正确标识的概率，即

$$FT(a)=\frac{\sum_{I\in ISet}\delta(lb(I)\in ST(a,I))}{Num(ISet)},$$

其中，$lb(I)$为图像 I 的真实场景标识，$\delta(\cdot)$为 dirac delta 函数，$\delta(\text{true})=1$，$\delta(\text{false})=0$，$Num(ISet)$为训练数据中的图像个数. 表 1 为上述 MSRCv2 数据集的场景识别率.

表 1　全局场景先验模型下 MSRCv2 数据集的识别率

场景识别数 a	识别精度 $FT(a)/\%$	场景识别数 a	识别精度 $FT(a)/\%$
1	80	3	95
2	91	4	100

从表 1 中可以看出，该方法能高概率地识别出场景类别，前 $a=4$ 个最大可能标识下的识别率高达 100%，证实了由随机森林分类器结合三类全局特征训练出的全局场景先验模型可有效识别场景类别. 在实验中，我们选出前 $a=4$ 个最大可能标识为候选场景.

4.2　场景的初次解析

局部超像素先验模型建立在 Graph-Based 超像素法[20]基础上，并结合了 4 类外观特征，即颜色、梯度、纹理和几何特征. 其中，颜色特征包括 HSV 颜色空间的均值(3 维)、HSV 颜色空间的亮度分量直方图(5 维)和饱和度分量直方图(5 维)，一共是 13 维颜色描述子；梯度特征建立在稠密 SIFT 特征[21]上，首先均匀采样出图像集的 SIFT 特征，然后用 K-Means[22]将 SIFT 特征聚类成 100 个中心，并根据聚类中心量化 SIFT 特征得到 SIFT 量化子，最后统计分割块中 SIFT 量化子的直方图得到一个 100 维梯度描述子；纹理特征是由 48 个 49×49 维纹理滤波子①对图像滤波生成的一个 48 维纹理描述子；几何特征包括：分割块的中心点(2 维)、水平方向的最高和最低位置(2 维)、垂直方向的最高和最低位置(2 维)、分割块包围盒的长宽比(1 维)，一共是 7 维几何描述子. 因此从每个分割块中可提取一个 168 维的外观特征描述子.

由于我们在场景识别阶段已经对场景分类，下面对每类场景独立训练以避免不同场景中外观相似物体出现混淆. 由随机森林分类器[19]对先验数据的分割块特征进行训练，以训练出的分类器模型为分割块外观先验模型，其中的模型参数为 $\theta_\phi^s=\{\theta_{\phi,k}^s\}$，$k=1,\cdots,K$，$K$ 为从训练数据中获知的场景标识数.

定义式(2)的分割块数据项为

$$\phi_s(l_i^s,S=k\mid I,\theta_\phi^s)=-\log p(l_i^s\mid I,\theta_{\phi,k}^s),$$

① http://www.robots.ox.ac.uk/~vgg/research/texclass/

其中，$p(l_i^s|I,\theta_{\phi,k}^s)$是由分类器模型$\theta_{\phi,k}^s$预测的分割块标识概率.

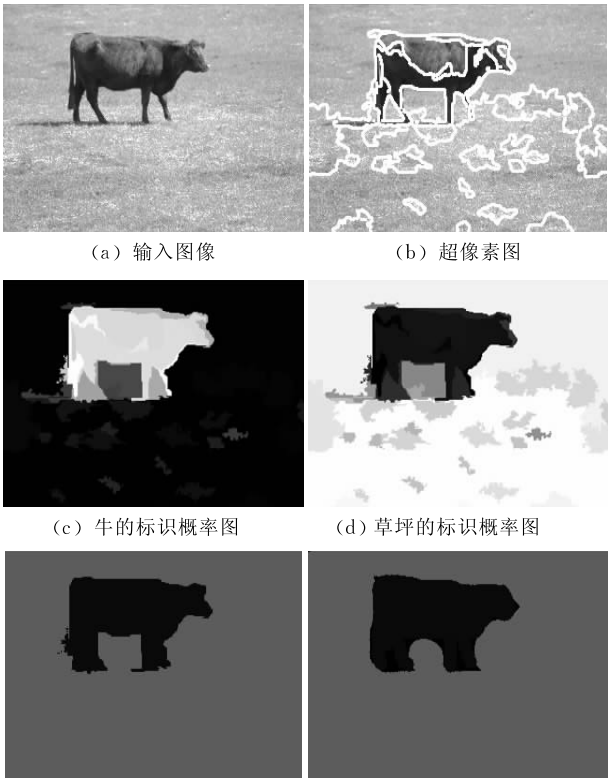
定义式(2)的平滑项为

$$\varphi_s(l_i^s,l_j^s,S=k|I,\theta_\varphi^s)=w(l_i^s,l_j^s,k)\delta(l_i^s\neq l_j^s),$$

其中，$\delta(\cdot)$为 dirac delta 函数，$w(l_i^s,l_j^s,k)$为相邻项的边缘权重，我们同样以监督学习方式训练出的平滑项先验模型计算该值. 平滑项先验模型所依赖的边缘特征为相邻超像素外观特征差的绝对值，边缘标识为根据标准解析图计算的 $\delta(\cdot)$值. 由随机森林分类器[19]结合边缘特征和边缘标识进行训练，获取 K 组二值分类器模型作为平滑项先验模型(模型参数为 $\theta_\varphi^s=\{\theta_{\varphi,k}^s\},k=1,\cdots,K$). 解析阶段的边缘权重为

$$w(l_i^s,l_j^s,k)=-\log p(\delta(l_i^s\neq l_j^s)|I,\theta_{\varphi,k}^s).$$

将式(2)中的权重经验值设定为 $\lambda_k=0.8,\lambda_s=1.2$，最后由最大流/最小割算法[23]优化求解式(2)完成场景的初次解析. 图 3 展示了 MSRCv2 中某个场景(图 3(a))的初次解析结果. 其中，图 3(b)为原图的超像素图，图 3(c)、(d)分别为牛和草的标识概率图(黑色为低概率，白色为高概率)，图 3(e)为初次解析结果，图 3(f)为标准解析图.



(a) 输入图像 (b) 超像素图

(c) 牛的标识概率图 (d) 草坪的标识概率图

(e) 初次解析结果 (f) 标准解析图

图 3 以像素点为结点的条件随机场示意图

4.3 场景的再次解析

从图 3(e)可以看出，基于先验模型的场景解析虽然可以粗略识别和分割出物体，但无法保留场景中的一些细节如牛腿、牛尾巴等，主要原因是超像素算法导致的欠分割和物体外观的多样性. 借鉴于 GrabCut[14]，我们引入局部颜色模型对场景进一步精细解析. 与 GrabCut 不同的是：(1) GrabCut 的输入为交互式的前景包围盒或前景/背景笔划，本文方法输入为初次解析结果，因此是全自动的；(2) GrabCut 仅分割前景和背景两个物体，本文方法将其扩展为多物体的分割；(3) GrabCut 在分割过程中仅利用了局部颜色模型，而本文方法结合了先验模型.

在再次解析阶段用 EM 迭代法[17]不断更新局部颜色模型参数来指导场景解析，实验表明经过 2 或 3 次迭代后场景解析结果趋于稳定，因此将 EM 迭代次数的经验值设定为 3.

M-步骤. 用于更新局部颜色模型的参数. 用 RGB 空间中的颜色高斯混合模型表示局部颜色模型，其参数为 $\theta_\phi^p=\{(\mu_m^n,\Sigma_m^n,w_m^n)\},m=1,\cdots,M,n=1,\cdots,N$. 其中，$N$ 为场景中的物体数目，M 为颜色高斯混合模型的簇数(考虑到场景物体一般有 5 部分以下不同颜色的部件组成，将 M 值固定为 5)，μ_m^n为每簇颜色的均值，Σ_m^n为颜色协方差，w_m^n为相应的权重.

在初次解析后我们根据解析结果估计出场景中的物体数目 N. 在迭代解析过程中，首先根据当前解析结果将同类物体的像素颜色值汇总，然后用 K-Means 聚类法[22]将每类物体颜色聚为 M 簇，最后统计每个聚类簇的均值、方差和权重来更新模型.

E-步骤. 用于优化解析结果. 在 3.2 节式(3)描述了该阶段条件随机场下的能量函数. 其中的数据项包括先验模型数据项 $\phi_s(l_i^p|I)$和局部颜色模型数据项 $\phi_p(l_i^p|I,\theta_\phi^p)$. 先验模型数据项 $\phi_s(l_i^p|I)$来源于初次解析结果，即将初次解析过程中计算出的超像素先验概率传递到块内的各个像素点上. 局部颜色模型数据项定义为

$$\phi_p(l_i^p=n|I,\theta_\phi^p)=\min_m(\phi_{p,m}(l_i^p=n|I,\theta_\phi^p));$$

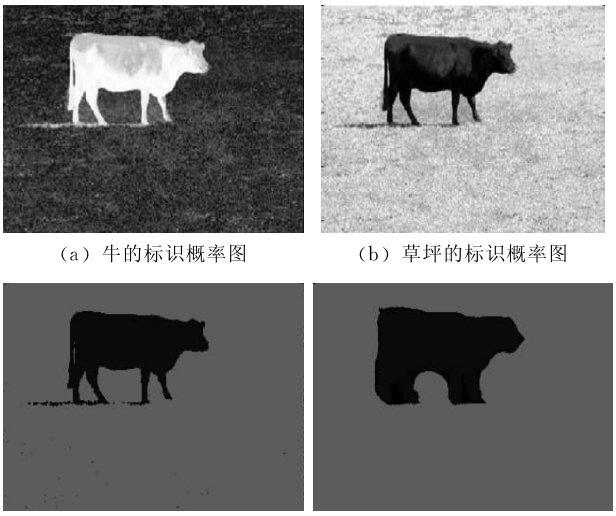
$$\phi_{p,m}(l_i^p=n|I,\theta_\phi^p)=-\log(w_m^nG(l_i^p|\mu_m^n,\Sigma_m^n)).$$

其中，$G(\cdot)$为正态分布. 平滑项定义为

$$\varphi_p(l_i^p,l_j^p|I)=\exp(-d(i,j)/2\beta)\delta(l_i^p\neq l_j^p).$$

其中，$d(i,j)=\|I_i-I_j\|$，I_i为像素点 i 的颜色，β 为图像内所有相邻像素颜色差的数学期望值，$\delta(\cdot)$为 dirac delta 函数，将权重 λ_c、λ_p的经验值设为 2.0、50. 最后，由最大流/最小割算法[23]求解式(3)对场景再次解析. 图 4 为图 3(a)的再次解析结果. 其中，图 4(a)、(b)分别为牛和草坪的标识概率图(黑色为低概率，白色为高概率)，图 4(c)为再次解析结果. 从图 4 中可以看出，相比初次解析结果(图 3(e))，

再次解析结果更精细，完整地保留了边缘细节，如牛尾巴、牛腿处等.



（a）牛的标识概率图　（b）草坪的标识概率图

（c）再次解析结果　（d）标准解析图

图 4　本文方法的再次场景解析结果

算法 1.　本场景解析法的算法流程.

输入：训练图像及其场景标识和标准解析图，测试图像

输出：测试图像的解析图

训练过程：

步骤 1. 训练出全局场景先验模型

① 提取训练图像的全局特征和场景标识；

② 由随机森林分类器训练获取全局场景先验模型.

步骤 2. 训练出局部分割块先验模型

③ 提取分割块的外观特征和分割块标识；

④ 由随机森林对不同场景标识的测试图像独立训练，获取局部分割块以及平滑项先验模型.

解析过程：

步骤 1. 计算场景标识

提取测试图像的全局特征，计算场景标识概率，选前 4 个最大可能标识为候选场景.

步骤 2. 场景的初次解析

提取分割块的外观特征，计算式(2)中的数据项和平滑项，由最小割算法求解式(2)实现场景的初次解析.

步骤 3. 用 EM 法迭代解析场景

M-步骤. 由当前解析结果更新局部颜色模型；

E-步骤. 根据先验和局部颜色模型计算式(3)中的数据项，用最小割法求解式(3)对场景再次解析.

5　实验结果和讨论

我们用两组公共的数据集 MSRCv1 和 MSRCv2① 对本文方法进行测试. MSRCv1 是由 13 种物体（建筑物、草、树、牛、马、羊、天空、山、飞机、水、人脸、汽车、自行车）组成的 240 幅图像，MSRCv2 是由 23 种物体（建筑物、草、树、牛、马、羊、天空、山、飞机、水、人脸、汽车、自行车、花、路牌、鸟、书、椅子、道路、猫、狗、人体、船）组成的 591 幅图像. 这两组图像集均有标准解析图，每幅图像大小为 320×213 左右. 在训练阶段，我们根据场景标识将 MSRCv1 分为 8 类，将 MSRCv2 分为 20 类. 从每类场景中随机选取 20% 的图像为测试数据，剩余图像为训练数据.

本文方法在两层条件随机场下解析场景，图 5 为部分图像的初次和再次解析结果. 如图 5 所示，虽然初次解析阶段可以粗略地提取物体，然而由于超像素算法导致的欠分割使得物体边缘轮廓比较粗糙. 经过再次解析后，场景被分割的更精细，物体的边缘细节得以完整保留.

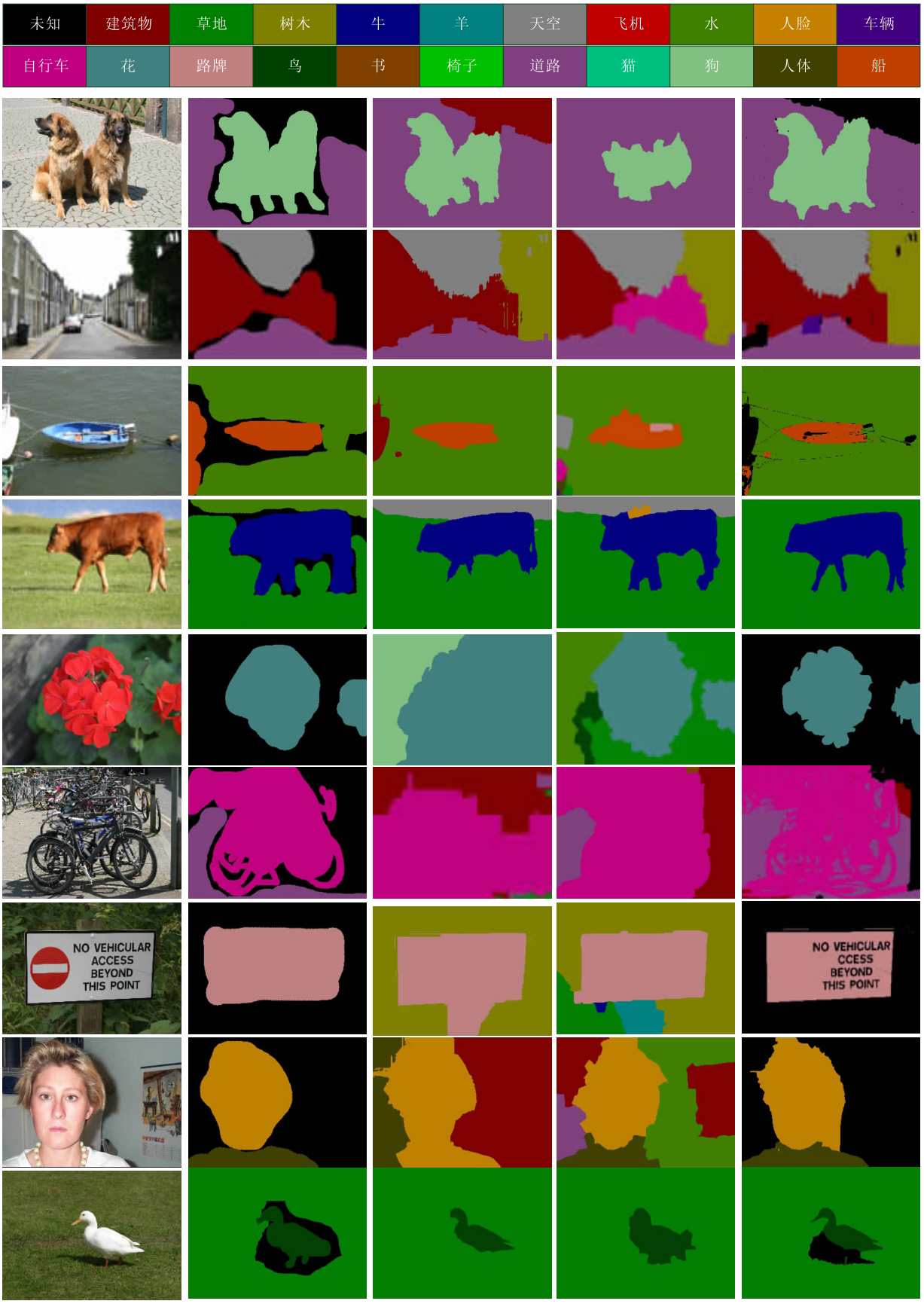
（a）原图　（b）初次解析　（c）再次解析　（d）标准解析图

图 5　本文方法的初次和再次解析结果比较

将本文方法与近年的 Texton[3]、STAIR[5] 和 ALE[10] 方法进行了比较. 这 3 个方法都是利用先验模型建立在条件随机场下的场景解析法. 其中，Texton[3] 建立在像素点基础上，用 Joint Boosting 分类器对纹理、颜色、位置等特征进行监督学习；STAIR[5] 建立在超像素基础上，用 Discrete Boosting 分类器对纹理、形状、位置等特征进行监督学习；ALE[10] 结合了像素点和超像素，用随机森林分类器对纹理、梯度等特征进行监督学习，然后在一个高阶条件随机场下解析场景.

本文方法和 STAIR[5]、ALE[10] 的部分结果如图 6

① http://research.microsoft.com/en-us/projects/objectclassrecognition/



(a) 输入图像 (b) 标准解析图 (c) STAIR[5]结果 (d) ALE[10]结果 (e) 本方法结果

图 6 本文方法和 STAIR[5]、ALE[10]的解析结果比较

所示.从图 6 比较效果上看,STAIR[5] 仅在超像素级别上进行解析,超像素算法存在的欠分割问题使得大部分物体的边界轮廓比较粗糙. ALE[10] 通过对像素点的软聚类使得物体边缘比较平滑,然而仍然受超像素和先验模型的影响,解析精度不够高.相比于 ALE[10],本文引入局部颜色模型在双层条件随机场下的方法能更精细地解析场景,较好保留边缘细节,例如车轴、牛腿、鸭嘴等.

我们用相同的训练和测试数据在一个 2.99 GHz CPU、2.0 GB 内存的台式机上量化比较了这 4 个方法,得到平均像素级、物体级的 F-Measure 值如表 2、表 3 所示,表中加粗数字为该组的最优值.其中,F-Measure $=2\times Pre\times Rec/(Pre+Rec)$,$Pre$ 为正确解析点在解析结果图所占的比例,Rec 为正确解析点在标准解析图所占的比例.如表 2、表 3 所示,Texton[3] 的精度最低,STAIR[5] 和 ALE[10] 的精度相当,本文方法的精度明显优于以上三者.

表 2　本文方法和 Texton[3]、STAIR[5]、ALE[10] 的平均像素级分割精度(%)和性能比较

方法的性能	MSRCv1			MSRCv2		
	精度/%	训练时间/h	解析时间/s	精度/%	训练时间/h	解析时间/s
Texton	45.3	6.80	169.00	43.40	14.70	169.00
STAIR	72.6	0.83	**6.80**	61.30	1.90	**6.80**
ALE	71.3	2.80	43.00	61.70	4.90	43.00
本文方法	**81.8**	**0.24**	13.30	**80.00**	**0.53**	13.30

表 3　本文方法和 Texton[3]、STAIR[5]、ALE[10] 的物体级分割精度比较

方法	精度/%																					
	建筑物	草地	树木	牛	羊	天空	飞机	水	人脸	汽车	自行车	花瓣	路牌	鸟	书架	椅子	道路	猫	狗	人体	船	平均值
Texton	23	26	47	65	71	38	67	42	73	37	82	34	63	0	84	17	33	0	56	42	0	42.8
STAIR	50	71	67	69	82	81	88	52	**89**	57	89	94	82	60	**96**	59	75	62	57	76	37	71.0
ALE	53	71	66	86	93	**86**	**89**	75	83	**89**	**93**	**96**	71	**87**	95	63	61	67	83	73	40	77.1
本文方法	**73**	**72**	**71**	**89**	**94**	82	73	**83**	87	82	92	90	**88**	86	94	**83**	**84**	**86**	**94**	**90**	**66**	**83.7**

从计算效率上分析,无论解析时间还是训练时间,Texton[3] 的效率都是最低的,主要原因是该方法直接在像素级层次解析场景,需要提取每个像素点的外观特征用于监督学习,因此计算量比较大.相比而言,STAIR[5]、ALE[10] 和本文方法初次解析都是建立在超像素级别上,因此训练和解析的时间较低.其次,相比 STAIR[5] 和本文方法,ALE[10] 的训练和解析效率比较低,这是因为 ALE[10] 共构建了 6 层的超像素层,多层超像素涉及结点过多,计算量偏大,而本文方法初次解析和 STAIR 都建立在单层超像素场景解析的基础上.相比于 STAIR[5],本文方法为了保留场景细节,引入局部颜色模型对场景的再次解析增加了计算量,因此解析效率稍低于 STAIR[5].然而本文方法训练效率高于 STAIR[5],我们通过补充实验找出其原因是我们采用的随机森林分类器比 STAIR[5] 的 Boosting 分类器计算效率更高.

6　总　结

针对目前场景解析方法中存在的解析不精细问题,本文引入了局部颜色模型,提出了双层条件随机场的场景解析方法.首先根据先验模型在以超像素为结点的条件随机场下粗略解析出场景中的物体;然后提取图像中每个物体的局部颜色模型,在以像素点为结点条件随机场下再次精细解析场景.实验表明,相比于以往单纯基于先验模型的场景解析方法,本文方法能更精细地解析场景,保留了物体的边缘细节.

本文方法还存在两点局限性:(1) 局部颜色模型是基于颜色信息来提取物体,因此本文方法主要适用于物体间颜色差异较大的场景解析;(2) 本文方法初次解析属于监督学习的方式,要求训练数据包括像素级的标准解析图,而手动获取标准解析图的工作量大,因此该方法的数据规模较小.为扩展应用范围,可进一步考虑如何提取带有颜色、纹理、梯度和深度等信息的局部外观模型,结合先验模型和局部外观模型以弱监督方式来精细解析场景.此外,目前视频场景解析方法中也存在解析不精细的问题,还可以考虑如何结合先验模型和局部外观模型来提高视频场景解析的精度.

参　考　文　献

[1] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the International Conference on Machine Learning. Williamstown, USA, 2001: 282-289

[2] He X, Zemel R S, Carreira-Perpinan M A. Multiscale conditional random fields for image labeling//Proceedings of the IEEE Computer Vision and Pattern Recognition. Washington, USA, 2004: 695-702

[3] Shotton J, Winn J, Rother C, Criminisi A. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation//Proceedings of the European Conference on Computer Vision. Graz, Austria, 2006: 1-15

[4] Yang L, Meer P, Foran D J. Multiple class segmentation using a unified framework over mean-shift patches//Proceedings of the IEEE Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[5] Gould S, Rodgers J, Cohen D, Elidan G, Koller D. Multi-class segmentation with relative location prior. International Journal of Computer Vision, 2008, 80(3): 300-316

[6] Caroline P, Cordelia S, Martial H. Object recognition by integrating multiple image segmentations//Proceedings of the European Conference on Computer Vision. Marseille, France, 2008: 481-494

[7] Stephen G, Richard F, Daphne K. Decomposing a scene into geometric and semantically consistent regions//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1-8

[8] Kumar M P, Koller D. Efficiently selecting regions for scene understanding//Proceedings of the IEEE Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3217-3224

[9] Cheny X, Jainy A, Guptax A, Davis S. Piecing together the segmentation jigsaw using context//Proceedings of the IEEE Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 2001-2008

[10] Kohli P, Ladicky L, Torr P H S. Robust higher order potentials for enforcing label consistency//Proceedings of the IEEE Computer Vision and Pattern Recognition. Anchorage, Alaska, USA, 2008: 1-8

[11] Ladicky L, Russell C, Kohli P. Associative hierarchical CRFs for object class image segmentation//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 739-746

[12] Ladicky L, Russell C, Kohli P, Torr P H S. Graph cut based inference with co-occurrence statistics//Proceedings of the European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 239-253

[13] Boykov Y Y, Jolly M. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images//Proceedings of the IEEE International Conference on Computer Vision. Vancouver, Canada, 2001: 105-112

[14] Rother C, Kolmogorov V, Blake A, "Grabcut": Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics, 2004, 23(3): 309-314

[15] Wang J, Cohen M. Image and video matting: A survey. Foundations and Trends in Computer Graphics and Vision, 2007, 3(2): 97-175

[16] Ohta Y, Kanade T, Sakai T. An analysis system for scenes containing objects with substructures//Proceedings of International Joint Conference on Pattern Recognitions. Kyoto, Japan, 1978: 752-754

[17] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977, 39(1): 1-38

[18] Oliva A, Torralba A. Building the gist of a scene: The role of global image features in recognition. Visual Perception, Progress in Brain Research, 2006, 155: 23-36

[19] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32

[20] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation. International Journal of Computer Vision, 2004, 59(2): 167-181

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

[22] Ding C, He X. K-means clustering via principal component analysis//Proceedings of International Conference on Machine Learning. Banff, Alberta, Canada, 2004: 225-232

[23] Kolmogorov V, Zabih R. What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 147-159

LI Yan-Li, born in 1982, Ph. D. candidate. Her research interests include image processing and pattern recognition.

ZHOU Zhong, born in 1978, Ph. D., associate professor. His current research interests include augmented virtual environment and natural phenomena simulation.

WU Wei, born in 1961, Ph. D., professor. His current research interests focus on virtual reality and distributed system.

Background

The problem of scene parsing has a long history in computer vision dating back to the 1970's, which is the core technology of image understanding, content based image retrieval and object recognition. The goal is to decompose the scene into semantically labeled objects, i. e., assign every pixel of the image with an object class label. State-of-the-art approaches all focus on exploiting prior models to guide scene parsing. The prior models, defined at the pixel or super-pixel level, are typically trained by supervised learning using image cues such as color, location and texture etc. Since the appearances for a single category are various in different conditions, and the prior models only encode the common traits of the category, those approaches suffer from detail loss.

In the field of segmentation, interactive figure-ground segmentation methods preserve the ability of precisely separating the figures from the background, which are mainly based on local color model. Motivated by the interactive figure-ground segmentation methods, we introduce the local color model into scene parsing and present a hierarchical two-level CRF approach to preserve details for scene parsing. Given a test image, we first obtain initial pixel labels in a super-pixel based CRF with constraint of prior models, and then extract the local color model for each object. Combining the prior and local color models, we iteratively refine scene parsing with an EM method in a pixel based CRF. Experimental results show that the two-level CRF approach is superior to state-of-the-art methods on benchmark data sets, verifying that our approach is able to preserve details and achieve better performance.

This work is supported by the National Natural Science Foundation of China under Grant of 61170188, the National High Technology Research and Development Program (863) of China under Grand of 2012AA011803, and the Ph. D. Programs Foundation of Ministry of Education under Grant of 20121102130004.