



结合 CRF 与 ShapeBM 形状先验的图像标记

王浩, 郭立君, 王亚东, 张荣

(宁波大学信息科学与工程学院, 浙江 宁波 315211)

摘要: 条件随机场(CRF)是一种强大的图像标记模型, 适合描述图像相邻区域间(例如超像素)的相互作用。然而, CRF 没有考虑标记对象的全局约束。对象的整体形状可以作为对象标记的一种全局约束, 利用形状玻尔兹曼机(ShapeBM)在建模对象的整体形状方面的优势, 提出了一种 CRF 与 ShapeBM 相结合的标记模型。标记模型建立在超像素的基础上, 并通过 pooling 技术在 CRF 的超像素层与 ShapeBM 的输入层间建立对应关系, 增强了 CRF 与 ShapeBM 结合的有效性, 提高了标记准确率。在 Penn-Fudan Pedestrians 数据集和 Caltech-UCSD Birds 200 数据集上的实验结果表明, 联合模型明显地改善了标记结果。

关键词: 条件随机场; 形状玻尔兹曼机; 联合模型; 超像素; 图像标记

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017004

CRF combined with ShapeBM shape priors for image labeling

WANG Hao, GUO Lijun, WANG Yadong, ZHANG Rong

School of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China

Abstract: Conditional random field (CRF) is a powerful model for image labeling, it is particularly well-suited to model local interactions among adjacent regions (e.g. superpixels). However, CRF doesn't consider the global constraint of objects. The overall shape of the object is used as a global constraint, the ShapeBM can be taken advantage of modeling the global shape of object, and then a new labeling model that combined the above two types of models was presented. The combination of CRF and ShapeBM was based on the superpixels, through the pooling technology was used to establish the corresponding relationship between the CRF superpixel layer and the ShapeBM input layer. It enhanced the effectiveness of the combination of CRF and ShapeBM and improved the accuracy of the labeling. The experiments on the Penn-Fudan Pedestrians dataset and Caltech-UCSD Birds 200 dataset demonstrate that the model is more effective and efficient than others.

Key words: CRF, ShapeBM, join model, superpixels, image labeling

收稿日期: 2016-07-24; 修回日期: 2016-09-30

通信作者: 郭立君, guolijun@nbu.edu.cn

基金项目: 国家自然科学基金资助项目(No.61175026); 浙江省自然科学基金资助项目(No.LY17F030002); 宁波市自然科学基金资助项目(No.2014A610031); 浙江省“信息与通信工程”重中之重学科开放基金资助项目(No.xkx1516, No.xkx1521)

Foundation Items: The National Natural Science Foundation of China(No.61175026), Zhejiang Provincial Natural Science Foundation of China(No.LY17F030002), Ningbo Municipal Natural Science Foundation(No.2014A610031), Open Research Fund of Zhejiang First-Forermost Key Subject-Information and Communications Engineering of China(No.xkx1516, No.xkx1521)



1 引言

图像分割与区域标记是计算机视觉中的核心技术,目的是把图像区域分割成相关的若干部分。现如今很多视觉中的高层应用都依赖于准确的图像分割结果或者标记结果,如对象识别、场景分析等。由于存在遮挡、阴影、目标与背景特征相似等问题使得图像分割与标记一直是计算机视觉中最富有挑战性的任务之一。

在图像标记中将图像标记成对象(前景)与背景部分,已有研究工作表明,CRF 常常被用于图像标记任务。在图像的标记任务中,图像的局部约束关系一般表示为表观特征的局部一致性(相邻的区域更可能具有相同标记)。CRF 是一个概率图模型,能够有效地描述图像的局部约束关系,而对要标记对象的全局约束关系的描述具有局限性。例如,当对象部分区域边界模糊时,CRF 仅仅利用图像的局部约束关系,很可能将表观特征相似的背景错标记成对象部分,或者反之。在这种情况下,引入对象的全局约束来补充 CRF 在对象标记上的局限性。近年的研究表明,受限的玻尔兹曼机(RBM)和它的一些扩展模型,例如深度玻尔兹曼机(DBM)等,常常被用于目标对象的形状建模,特别是 Eslami 等人提出的 ShapeBM (形状玻尔兹曼机)在捕捉目标对象的形状方面有很好的表现,而对象的形状信息可以作为对象全局约束的一种表示。本文提出了一种 CRF 结合 ShapeBM 的标记模型,有效结合了 CRF 模型(增强相邻节点之间的局部约束)与 ShapeBM 模型(获得目标对象的整体形状信息作为对象的全局约束)的优点,获得较好的标记效果。这个模型有效地权衡了以下 3 个目标。

- 区域标记结果应该与图像的底层特征一致。
- 完整的图像标记结果应该符合通过训练数据所学到的形状先验。
- 结合局部与全局约束使得区域标记结果遵守对象边界。

其中,第一个目标由 CRF 部分完成,第二个目标通过训练 ShapeBM 完成,而第三个目标由 CRF 与 ShapeBM 联合完成。在 Penn-Fudan Pedestrians 数据集和 Caltech-UCSD Birds 200 数据集上评估了所提出的模型的标记效果,经过实验论证发现提出的模型比 CRF 和 ShapeBM 等模型效果好。

本文的主要工作总结如下。

- 所提出的模型有效地结合了 CRF 和 ShapeBM 模型,保持了图像局部约束与对象全局约束的一致性。
- 针对联合模型求解较难的问题,给出了有效的训练和推断算法。
- 联合模型在对象标记的结果上要明显优于基本的 CRF 模型与 ShapeBM 模型。

2 相关工作

Lafferty 等人在 2001 年基于隐马尔可夫模型和最大熵模型的理论基础提出了 CRF^[1],是一种概率图模型,最早应用于自然语言处理,近年来已成功应用于图像分割^[2-5]。He 等^[3]用 CRF 分割静态图像,通过引入隐随机变量加入更多的上下文知识以便进行图像分割。Zhang 等^[6]将超像素方法引入 CRF 模型中,通过一种各向异性扩散算法将图像过分割成超像素区域,然后将区域作为 CRF 图模型中的节点,通过参数估计获得给定测试图像的最优标记。上述方法均只利用图像局部约束信息,缺乏标记对象的全局约束信息,导致对象边界模糊的情况,分割效果不理想。针对上述情况,引入对象全局约束信息来补充 CRF 的局限很有必要。对象的形状约束信息可以作为对象全局约束信息的一种表现形式,最近很多文献都采用了 RBM^[7]或者其扩展模型来获取对象的形状约束信息。Salakhutdinov 等^[8]在 RBM 的基础上提出 DBM,是一个多层的 RBM 模型,通过多层的 RBM 进行特征提取获得目标对象的高层语义特征如形状、姿态等。Eslami 等^[9]提出了 ShapeBM,是一个两层的 DBM,第一层将输入单元分成 4 部分,每一部分与其相对应的部分隐单元连接,而第二层与 RBM 相同采用全连接,ShapeBM 在对象形状建模上取得较好效果。近年来有许多方法也在图像的局部约束和对象的全局约束相结合上做了很多工作。Kae 等^[10,11]提出通过 RBM 模型学习人脸形状先验特征作为对象全局约束,并借助于 CRF 分割框架能够融入其他分割线索的特性,将所学到的形状先验融入 CRF 分割框架中,获得较好的分割与标记效果。Chen 等^[12]通过 ShapeBM 学到的对象形状先验结合到 Cremers 等^[13]提出的变分分割模型中,通过求其能量函数最小达到分割的效果。

联合模型同样根据 CRF 分割框架能够融入多种分割线索的特性,通过 ShapeBM 学习对象的形状先验,并以能量项的形式融入 CRF 分割框架中,通过有效训练和推断算法,得到最终标记结果。所提出的模型尽管与上

述介绍的局部与全局约束相结合的方法相似,但是联合模型与它们有一些区别。首次提出 CRF 与 ShapeBM 结合的联合标记模型,该联合模型采用超像素标签代替像素标签,有效地利用了超像素边界通常与真实对象边界一致的特性,提高了联合模型的分割效果并且降低了联合模型计算的复杂性,并且通过 pooling 技术解决了图像的超像素个数与 ShapeBM 输入节点不对应问题,使得 CRF 与 ShapeBM 有效结合并保持了图像局部约束与对象全局约束的一致性。

3 模型建立

提出 CRF 与 ShapeBM 结合的联合模型,该模型有效地结合了图像局部约束与对象全局约束,其中用 CRF 描述图像的局部约束,用 ShapeBM 描述对象形状作为对象的全局约束,首先分别介绍基本 CRF 模型、ShapeBM 模型,在此基础上详述提出的 CRF 与 ShapeBM 相结合的联合模型。

使用如下定义。

- $S^{(I)}$ 表示图像 I 的超像素个数。
- $V^{(I)} = \{1, \dots, S(I)\}$ 表示第 I 幅图像的超像素节点的集合。
- $\mathcal{E}^{(I)} = \{(i, j), i, j \in V^{(I)}\}$ 表示相邻超像素节点组成的边的集合。
- $V^{(I)} = \{X_V^{(I)}, X_E^{(I)}\}$ 表示节点特征与边特征的集合,其中 $X_V^{(I)}$ 表示节点特征的集合即 $\{x_s^{\text{node}} \in R^{D_h}, S \in V\}$, D_h 表示节点特征的维度。 $X_E^{(I)}$ 表示相邻节点连接边的边特征集合即 $\{x_{ij}^{\text{edge}} \in R^{D_e}, (i, j) \in \mathcal{E}\}$, D_e 表示边特征的维度。

- $Y^{(I)} = \{y_s \in \{0, 1\}^L, S \in V, \sum_{s \in V} y_{sl} = 1\}$ 表示图像超像素节点

点标签的集合,其中 L 表示类标签数,本文的实验中 L 为 2, y_s 中 0 表示背景,1 表示对象(前景)。即本文中的联合模型完成图像前景背景二分类标记问题。

3.1 CRF

CRF 是一个强大的判别型模型,并在文本分析、图像分割^[2-4,14]等领域有广泛应用。CRF 的条件分布与其能量函数定义如下:

$$P_{\text{crf}}(Y|X) \propto \exp(-E_{\text{crf}}(Y, X)) \quad (1)$$

$$E_{\text{crf}}(Y, X) = E_{\text{node}}(Y, X_V) + E_{\text{edge}}(Y, X_E) \quad (2)$$

其中, $E_{\text{node}}(Y, X_V)$ 表示 CRF 中的一阶势,表示超像素属于对象或属于背景的似然。 $E_{\text{edge}}(Y, X_E)$ 表示二阶势,主要决定了两个超像素之间是如何相互作用的,体现对象的局部约束,主要作用是去除锯齿、平滑边界。通过求解训练集 $V^{(I)} = \{Y^{(I)}, X^{(I)}\}_{I=1}^L$ 的最大条件似然概率来训练其模型参数:

$$\max \sum_{I=1}^L \lg P_{\text{crf}}(Y^{(I)} | X^{(I)}) \quad (3)$$

对于该条件似然概率可以通过 LBP (loopy belief propagation)^[15]或者结合了标准优化算法 LBFGS^[16]的平均场估计^[17](mean-field approximation)计算。

3.2 ShapeBM

针对对象形状信息的获取,近些年 RBM 及其扩展模型在这方面应用较多,本文所采用的模型是 RBM 的一种扩展模型 ShapeBM,已有文献证明它相对于 RBM 能够更加有效地提取对象整体形状信息。RBM 可以被视为一个无向图,具有一个可见层、一个隐藏层,其层间全连接层内无连接,结构如图 1(a)所示。ShapeBM 是一个 3 层模型,具有一个可见层和两个隐层,ShapeBM 具体结构如图 1(b,c)所示,由图 1(c)可以看出,ShapeBM 的可见层与第一层隐层相比 RBM 有些变化,主要是可见层分成了 4 部分区域,并且每部分区域之间有部分重叠,区域重叠使得 ShapeBM 能够更好地训练对象边界的连续性。另外,可见层中每部分可见单元与其相对应的部分隐层单元的连接权值共享(如图 1(c)中阴影部分可见单元与阴影部分隐层单元的连接权值可以共享作为其他颜色部分的连接权)。但其可见单元偏置不共享。权值共享使得 ShapeBM 第一层的参数相对于全连接方式要缩小 16 倍,即 ShapeBM 对训练集数量要求变小。ShapeBM 两个隐层的隐单元间的连接方式与 RBM 相同,采用全连接。

ShapeBM 能通过图模型的多层结构对对象的整体形状信息进行提取,使其最高层隐单元含有一些形状信息,通过实验结果验证了这一点。在 Penn-Fudan Pedestrians 数据集和 Caltech-UCSD Birds 200 数据集上通过一定数量的训练集训练,得到了 ShapeBM 最高层隐单元所含有的形状信息,并通过重构得出可视化的结果,图 2 列举出部分代表性重构结果。由其结果可以看出,ShapeBM 能够得到对象的整体形状,但可能会丢失一些细节,这也表明,能够对同一类对象的形状进行抽象描述。由此可见,ShapeBM



能够抽象出对象形状作为对象全局约束融合到联合标记模型中。

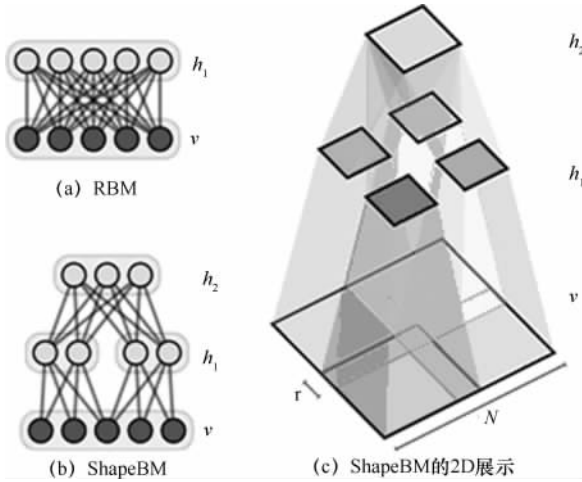


图1 ShapeBM 结构

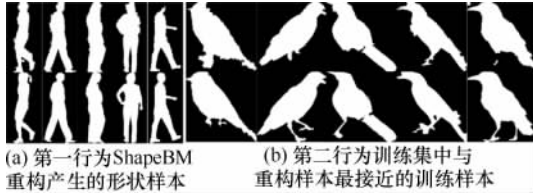


图2 部分代表性重构结果

在本文中,假设 ShapeBM 有 R^2 个可见单元 $y_r \in \{0,1\}^L$, 第一层隐层 h_1 有 K 个隐单元, 第二层隐层 h_2 有 M 个隐单元。它们的联合分布可以有如下定义:

$$P_{\text{ShapeBM}}(Y, h_1, h_2) \propto \exp(-E_{\text{ShapeBM}}(Y, h_1, h_2)) \quad (4)$$

$$E_{\text{ShapeBM}}(Y, h_1, h_2) = -\sum_{r=1}^{R^2} \sum_{l=1}^L \sum_{k=1}^K y_{rl} W_{rlk}^1 h_k^1 - \sum_{k=1}^K b_k h_k^1 - \sum_{r=1}^{R^2} \sum_{l=1}^L a_{rl} y_{rl} - \sum_{k=1}^K \sum_{m=1}^M h_k^1 W_{km}^2 h_m^2 - \sum_{m=1}^M c_m h_m^2 \quad (5)$$

其中, $W^1 \in R^{R^2 \times K}$ 是可见单元与第一层隐单元之间的连接权重, $W^2 \in R^{K \times M}$ 表示第一层隐层 h_1 与第二层隐层 h_2 之间的连接权重。 b_k 表示第一层隐节点的偏置值, a_{rl} 表示可见节点的偏置值, c_m 表示第二层隐节点的偏置值。训练 ShapeBM 分为两步, 首先采用 DBM 分层贪婪预训练的方式去预训练 ShapeBM 的每一层, 预训练第一个 RBM 时其连接方式要遵守 ShapeBM 的连接方式, 通过预训练的方式得到初始化参数 $\theta = \{W^1, W^2, a, b, c\}$ 。将得到的初始化参数 $\theta = \{W^1, W^2, a, b, c\}$ 作为 ShapeBM 的初始参数, 然后采用随机梯度下降 (stochastic gradient descent) 法最大化如下对数函数来微调

参数 θ :

$$\max_{\theta} \sum_{i=1}^I \lg \sum_{h_1, h_2} (P_{\text{ShapeBM}}(Y^{(i)}, h_1, h_2)) \quad (6)$$

由于参数的梯度不易直接求解, 采用平均场估计 (mean-field approximation) 来近似计算。

3.3 联合模型

在图像标记中, 表观特征的局部一致性 (相邻的区域更可能具有相同标记) 和全局一致性约束 (对象的整体形状) 都很重要。一方面, CRF 在通过二阶势 (边势能) 建模局部一致性方面具有优势; 另一方面, ShapeBM 能通过隐层单元描述对象全局形状结构。本文从两者优点出发, 采用有效的方式融合了 CRF 与 ShapeBM 得到联合标记模型, 使其既有局部约束又有全局约束即对象形状信息。

3.3.1 基于超像素的 CRF 与 ShapeBM 结合方式

联合模型中 CRF 与 ShapeBM 的结合是基于超像素的, 基于超像素的结合有以下两点好处: 一, 超像素的边界通常为真实对象边界, 提高了联合模型的分割效果; 二, 采用超像素代替像素, 降低了计算的复杂性。然而, 采用超像素也给 CRF 与 ShapeBM 的结合带来一定困难。由于不同图像的超像素个数不固定, 而 ShapeBM 要求输入层的可见层节点数量是固定的, 这造成超像素个数与 ShapeBM 的输入节点不对应, 因此不能简单地将超像素标签层作为 ShapeBM 的输入与其隐节点直接相连, 需要对其进行处理。为了建立超像素标签层与 ShapeBM 隐节点之间的连接, 在超像素标签层与 ShapeBM 隐层之间引入一个固定大小的虚拟标签层, 将每个超像素标签节点通过 pooling 方式映射到 $R \times R$ 个虚拟标签层节点, 每个虚拟可见节点都是一个小方格, 如图 3 所示, 上面 3 层是 ShapeBM 部分, 包括虚拟可见节点层以及两层隐藏节点层。下两层是 CRF 部分, 包括特征层以及超像素标签层。其中 P_{rs}^0 表示 pooling 中的映射概率用 $\bar{y}_r$ 来表示一个虚拟节点的标签。重新定义 ShapeBM 的能量函数, 其函数定义如下:

$$E_{\text{ShapeBM}}(Y, h_1, h_2) = -\sum_{r=1}^{R^2} \sum_{l=1}^L \sum_{k=1}^K \bar{y}_{rl} W_{rlk}^1 h_k^1 - \sum_{k=1}^K b_k h_k^1 - \sum_{r=1}^{R^2} \sum_{l=1}^L a_{rl} \bar{y}_{rl} - \sum_{k=1}^K \sum_{m=1}^M h_k^1 W_{km}^2 h_m^2 - \sum_{m=1}^M c_m h_m^2 \quad (7)$$

其中, 虚拟节点标签 $\bar{y}_{rl} = \sum_{s=1}^S p_{rs} y_{sl}$, p_{rs} 表示每个虚拟节点小方格由哪几个超像素部分组成的概率矩阵。 p_{rs} 的定义

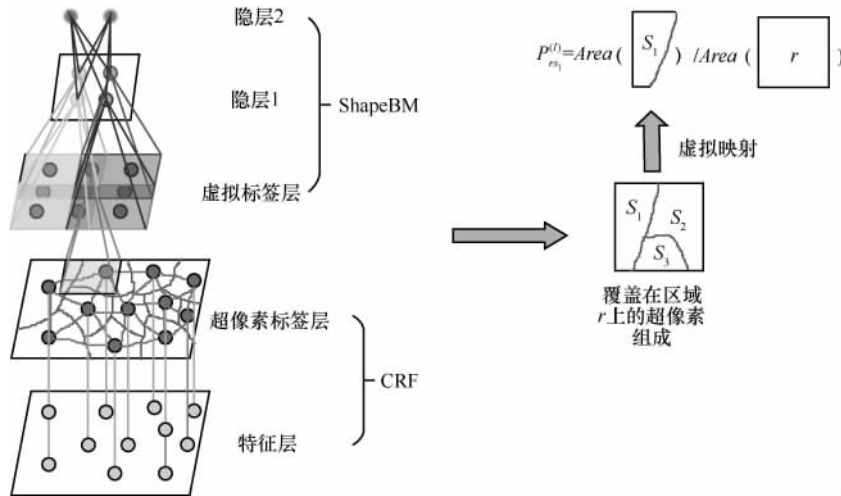


图 3 联合模型

如下:

$$p_{rs} = \frac{Area(Region(s) \cap Aegion(r))}{Area(Region(r))} \quad (8)$$

其中, $Region(s)$ 和 $Region(r)$ 表示对应超像素 s 覆盖区域和标签层第 r 个节点在超像素层投影覆盖的区域。

另外, CRF 部分采用的是基本 CRF 的一个变形叫做空间条件随机场 (SCRf) [10], 它利用了对对象空间分布的特征, 将对象的空间依赖关系引入 CRF。例如, 在空间上人体的头部在躯干的上方, 而躯干在腿部的上方。具体做法将图像规则成 $N \times N$ 的虚拟小格 (注意这里的虚拟小格的大小与上文 ShapeBM 虚拟输入中的 $R \times R$ 的虚拟小格不同), 这时一个超像素可能跨越多个虚拟小格, 这样每个超像素落在不同小格的部分都能够学习到不同的节点特征权重。本文定义 SCRf 的一阶势函数如下:

$$E_{\text{node}}(Y, X_V) = - \sum_{S \in V} \sum_{l=1}^L y_{sl} \sum_{n=1}^{N^2} p_{sn} \sum_{d=1}^{D_n} \alpha_{nld} x_{sd} \quad (9)$$

其中, N_{sd} 表示第 S 个超像素的节点特征, α_{nld} 为每个超像素落在不同虚拟方格部分的节点权重, p_{sn} 表示每个超像素由哪几个虚拟小方格组成 ($\sum_{n=1}^{N^2} p_{sn} = 1$)。求解类似于上文 p_{rs} 采用面积比, 区别在于 p_{sn} 表示每个小方格在一个超像素中的面积占比。

SCRf 二阶势函数定义如下:

$$E_{\text{edge}}(Y, X_E) = - \sum_{(i,j) \in E} \sum_{l=1}^L \sum_{c=1}^{D_c} y_{il} y_{jl} \beta_{ll'e} x_{ije} \quad (10)$$

其中, x_{ije} 表示边特征, 用来衡量相邻超像素之间的相

似度。 $\beta_{ll'e}$ 为相邻超像素间边特征的对应权重。

有了 SCRf 和 ShapeBM 的具体形式, 而本文联合模型是在超像素的基础上结合 SCRf 与 ShapeBM 模型。因此, 可以得到联合模型的具体形式。具体地, 在给定的超像素特征 X 下, 超像素标签集 Y 的条件分布以及能量函数定义如下:

$$P_{\text{join}}(Y/X) \propto \sum_{h_1, h_2} \exp(-E_{\text{join}}(Y, X, h_1, h_2)) \quad (11)$$

$$E_{\text{join}}(Y, X, h_1, h_2) = E_{\text{scrif}}(Y, X) + E_{\text{ShapeBM}}(Y, h_1, h_2) \quad (12)$$

从上述定义可以看出, 联合模型的结合在形式上仅仅是结合了 SCRf 和 ShapeBM 的能量函数。但是, 实际上, 本文提出的基于超像素的 pooling 方式在 SCRf 和 ShapeBM 的结合上起着至关重要的作用。

3.3.2 联合模型的训练与推断

关于联合模型的训练, 原则上, 可以直接通过最大化条件似然概率来一起训练模型的参数 $\{W^1, W^2, a, b, c, \alpha, \beta\}$ 。但是实际上, 会通过单独预训练 SCRf 和 ShapeBM 来提供一组参数作为联合模型的初始参数, 然后采用随机梯度下降法 (stochastic gradient descent) 最大化联合模型对数似然函数来调整参数 $\{W^1, W^2, a, b, c, \alpha, \beta\}$ 。模型的整个训练步骤见算法 1。在算法 1 的步骤 3 采用随机梯度下降法最大化条件似然过程中, 由于参数的梯度不易直接求解, 本文采用 CD-PercLoss [18] 方法来近似估计参数梯度, 而在 CD-PercLoss 算法反向计算过程中所推断出的超像素标签不仅仅与 ShapeBM 的隐节点有关, 还与 CRF 有关, 因此超像素的标签和隐节点的联合推断是必需的。本文采用平均场估计的方法来解决该联合推断问题。



具体地,发现分布 $G(Y, h_1, h_2; \mu, \gamma) = \prod_{s \in V} G(y_s) \prod_{m=1}^M G(h_m^2)$, 其中 $G(y_s=l) \triangleq \mu_{sl}$ 和 $G(h_m^2=1) \triangleq \gamma_m$ 能够使 $KL(G(Y, h_1, h_2; \mu, \gamma) \| P(Y, h_1, h_2 | X))$ 达到最小。即可以通过平均场估计的方法循环迭代更新 μ_{sl} 和 γ_m 来使模型的能量达到最小,从而获得较优的标记结果。平均场具体推断步骤见算法 2。

算法 1 训练联合模型

(1) 通过最大化空间 CRF 模型 (SCRF) 的条件似然来预训练参数 $\{\alpha, \beta\}$ (见式 (1)、式 (2)、式 (9) 和式 (10))。

(2) 通过最大化 ShapeBM 模型的条件似然 $\lg \sum_{h_1, h_2} P_{\text{ShapeBM}}(Y, h_1, h_2, |X_V)$ 来预训练参数 $\theta = (W^1, W^2, a, b, c)$ 。其中:

$$P_{\text{ShapeBM}}(Y, h_1, h_2 | X_V) \propto \exp(-E_{\text{node}}(Y, X_V; \alpha) - E_{\text{ShapeBM}}(Y, h_1, h_2; \theta))。$$

(3) 通过最大化联合模型的条件似然来训练参数 $\{W^1, W^2, a, b, c, \alpha, \beta\}$ (见式 (11))。

算法 2 平均场推论

(1) 初始化 $\mu^{(0)}$ 和 $\gamma^{(0)}$,

$$\mu^{(0)} = \frac{\exp(f_{sl}^{\text{node}})}{\sum_{l'} \exp(f_{sl'}^{\text{node}})}$$

$$\gamma^{(0)} = \text{sigmoid}\left(\sum_{k=1}^K h_k^{(0)} W_{km}^2 + c_m\right)。$$

$$\text{其中, } f_{sl}^{\text{node}}(X_V, \{p_{rs}\}, \alpha) = \sum_{n,d} p_{sn} x_{sd} \alpha_{ndl},$$

$$h_k^{(0)} = \text{sigmoid}\left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl}^{(0)}\right) W_{rlk}^1 + \sum_{m=1}^M h_m^2 W_{km}^2 + b_k\right),$$

$$h_m^2 = \text{sigmoid}\left(\sum_{k=1}^K h_k^1 W_{km}^2 + c_m\right),$$

$$h_k^1 = \text{sigmoid}\left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl}^{(0)}\right) W_{rlk}^1 + b_k\right)。$$

(2) for $t=0:\text{maxiter}$ (或者直到收敛) do

更新 $\mu^{(t+1)}$,

$$f_{sl}^{\text{ShapeBM}}(\mu, \gamma; \{p_{rs}\}, W^1, W^2, a, b) =$$

$$\sum_{r,k} p_{rs} \left(W_{rlk}^1 \left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl} \right) W_{rlk}^1 + \sum_m W_{km}^2 \gamma_m + b_k \right) + a_{rl} \right)。$$

$$\text{其中, } f_{sl}^{\text{edge}}(\mu; X_e, \varepsilon, \beta) = \sum_{j \in \{s, l\}} \sum_{l' \in e} \mu_{jl'} \beta_{ll'e} x_{sje},$$

$$f_{sl}^{\text{ShapeBM}}(\mu, \gamma; \{p_{rs}\}, W^1, W^2, a, b)$$

$$= \sum_{r,k} p_{rs} \left(W_{rlk}^1 \left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl} \right) W_{rlk}^1 + \sum_m W_{km}^2 \gamma_m + b_k \right) + a_{rl} \right)。$$

更新 $\gamma^{(t+1)}$,

$$\gamma_m^{(t+1)} = \text{sigmoid}\left(\sum_{k=1}^K (h_k^{(t+1)}) W_{km}^2 + c_m\right)。$$

$$\text{其中, } (h_k^{(t+1)}) = \text{sigmoid}\left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl}^{(t+1)}\right) W_{rlk}^1 + \sum_{m=1}^M (h_m^2)^{(t+1)} W_{km}^2 + b_k\right),$$

$$(h_m^2)^{(t+1)} = \text{sigmoid}\left(\sum_{k=1}^K (h_k^{(t+1)}) W_{km}^2 + c_m\right),$$

$$(h_k^1)^{(t+1)} = \text{sigmoid}\left(\sum_{r,l} \left(\sum_s p_{rs} \mu_{sl}^{(t+1)}\right) W_{rlk}^1 + b_k\right)$$

end for

4 实验与分析

4.1 数据集

本文在 Penn-Fudan Pedestrians 数据集^[19]和 Caltech-UCSD Birds 200 数据集^[20]上评估了所提出的联合模型的标记效果。Penn-Fudan Pedestrians 数据集共有 170 张图像, 每张图像包含至少一个行人, 每个被标记的行人都有 ground-truth。根据 Penn-Fudan Pedestrians 数据集的行人检测框标记信息提取出行人检测框部分的图像, 得到 423 张单个行人图像并将图像大小统一为 128 dpi×256 dpi 像素。为了增加训练和测试样本, 对 Penn-Fudan Pedestrians 数据集提取的 423 张行人图像进行镜像对称复制形成 846 张图像的数据集, 其中 500 张图像作为训练集, 346 张图像作为测试集。

Caltech-UCSD Birds 200 数据集包含 200 类鸟类的 6 033 张图像, 图像拥有粗糙的分割掩膜, 因为其掩膜精确度不够, Yang 等人^[21]手动标记出该数据集中图像的准确掩膜, 采用 Yang 等人手动标记的准确掩膜作为该数据集标准结果。另外本文根据数据集提供的检测框标记对 6 033 张图像分别提取出其中检测框部分的图像并将图像大小统

一为 128 dpi×128 dpi 像素。其中 3 033 张图像作为训练图像,3 000 张图像作为测试图像。

4.2 仿真与分析

本文对两个数据集中的每张图像都采用 SLIC^[22]方法进行超像素分割(如图 4 所示),并基于标准分割掩膜对每个超像素进行了标记作为 ground-truth。图 4 (a)是 Penn-Fudan Pedestrians 数据集和 Caltech-UCSD Birds 200 数据集中根据标记信息提取的单张图像,图 4(b)是通过 SLIC 方法分割获得的超像素图片,图 4(c)为 ground truth。

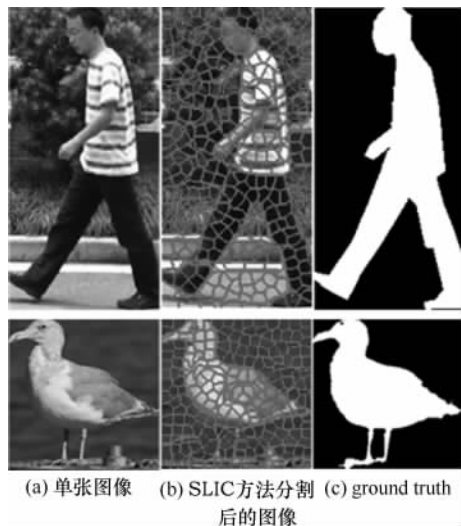


图 4 超像素分割

对每个超像素使用如下的节点特征。

- 颜色:在 LAB 空间中采用 K-means 产生的 64 位归一化颜色直方图。
 - 纹理:采用参考文献[23]产生的 64 位归一化纹理直方图。
 - 位置:每张图像规则成 8 dpi×8 dpi 的方格,超像素落在每个方格的概率直方图。
- 相邻超像素之间采用如下的边特征。
- 采用参考文献[24]边界的 PB 值之和。
 - 平均颜色直方图之间的欧式距离。
 - 采用参考文献[25]计算纹理直方图之间的卡方距离。

评估模型的标记能力时采用了 4 种不同的模型:标准的 CRF、SCRF、ShapeBM 以及本文联合模型。本文在 CPU 为 Intel Xeon E5-2650 2.60 GHz,内存为 128 GB 的计算机上进行实验。

本文正确率的衡量标准如下:

$$avg_truth = 1 - \frac{\sum_{i=1}^I Y_i OR(GT^{(i)})}{\sum_{i=1}^I GT^{(i)}} \quad (13)$$

式(13)主要表示标记正确的超像素个数和总的超像素个数的比例。其中 Y_i 表示第 i 张测试图像的超像素标记结果,OR 为异或操作, $GT^{(i)}$ 为第 i 张测试图像超像素标记的 ground-truth, I 表示测试图像的数量。

针对 Penn-Fudan Pedestrians 数据集,经过多次实验证明,本文设置参数 $K=500, M=200, R=30, N=23$ 具有较优的效果。通过 10 次交叉实验,每次实验的 500 张训练图像和 346 张测试图像都不完全相同,通过式(13)得到每次实验的准确率并通过平均得到本文最终实验准确率 87.90%。在相同实验情况下,相对于标准 CRF、SCRF、ShapeBM 在超像素标记上准确率都有提高,具体对比结果见表 1。表 1 中的错误率减少的比例表示模型相较于标准的 CRF 错误率减少程度。还给出了 Penn-Fudan Pedestrians 数据集中具有代表性的部分可视化分割结果,如图 5 所示。图 5(a)表示联合模型相较于其他模型分割效果较好的部分对比结果,图 5(b)表示联合模型分割效果不理想的部分对比结果,其中第一列为原图,第二列为 CRF 模型分割结果,第三列为 SCRF 模型分割结果,第四列为本文联合模型分割结果,第五列为 ground-truth。联合模型结合了对对象的整体形状先验,所以相对于其他的两个未增加对象整体形状信息的模型,联合模型能够调整明显违背对象形状信息的错误标记,如图 5(a)中第 1 行,行人两腿中间有一块错误标记,该错误标记明显违背了对象形状信息,联合模型通过形状信息将该部分错误标记调整过来。另外,如图 5(a)中第 2 行,行人的脸部和行人的手臂处都有一部分缺少,而联合模型能够通过形状信息将其填充完整。然而,联合模型有时在部分样本上分割效果不理想,如图 5(b)中第 2 行,由于图片分辨率较低以及光照影响较强,联合模型在处理光照变化存在一定局限性,所以导致图像分割失败,但是其他模型在这种情况下分割效果也不理想。如图 5(b)中第 5 行,图像中人体对象的腿部部分,联合模型将背景中的腿部部分错误估计为人体对象形状的一部分造成错误分割。

针对 Caltech-UCSD Birds 200 数据集,经过多次实验证明,设置参数 $K=400, M=100, R=32, N=21$ 具有较优的效果。通过 10 次交叉实验,每次实验的 3 033 张训练图像和



表 1 本文方法与其他方法在 Penn-Fudan Pedestrians 数据集上标记准确率对比

方法	超像素标记准确率	误识率
CRF	83.97%	参照
SCRf	86.31%	14.60%
ShapeBM	85.01%	6.49%
本文方法	87.90%	24.52%

3 000 张测试图像都不完全相同,通过式(13)得到每次实验的准确率并通过平均得到本文最终实验准确率 83.34%。在相同实验情况下,相对于标准 CRF,SCRf、ShapeBM 在超像素标记上准备率对比结果见表 2,可以看出联合模型准确率为 83.34%相对于其他方法有提高。理论上,该数据集上联合模型正确率提高的程度应该比 Penn-Fudan Pedestrians 数据集上高,因为该数据集的训练样本多,训练效果应该更好。但是实际上,从具体数据看,在该数据集上联合模型正确率提高的程度相对于

Penn-Fudan Pedestrians 数据集提高的程度反而较低,主要因为该数据集中包含了 200 类鸟类图片,各类鸟类的形状差异较大。而本文是直接将 3 033 张多类鸟类的图像统一作为训练集,使得训练集中鸟类形状多样化,所以联合模型训练出的结果并不理想。后期可以考虑将形状近似的鸟类作为训练集,训练出一个较好的结果。关于 Caltech-UCSD Birds 200 数据集的具有代表性的部分可视化分割结果如图 6 所示。

表 2 本文方法与其他方法在 Caltech-UCSD Birds 200 数据集上标记准确率对比

方法	超像素准确率	误识率
CRF	78.85%	参照
SCRf	81.93%	14.60%
ShapeBM	81.27%	11.44%
本文方法	83.34%	21.23%

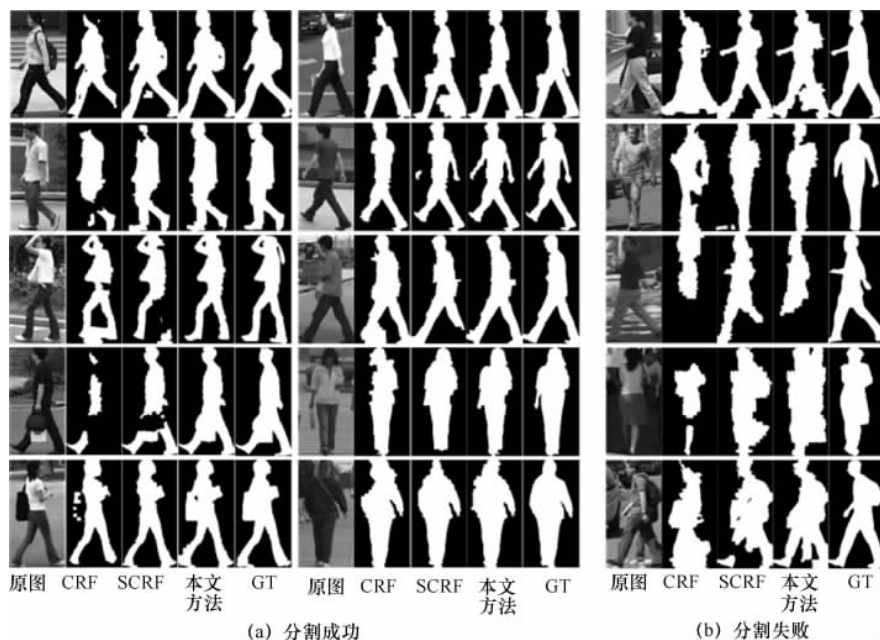


图 5 Penn-Fudan Pedestrians 数据集上的部分分割结果

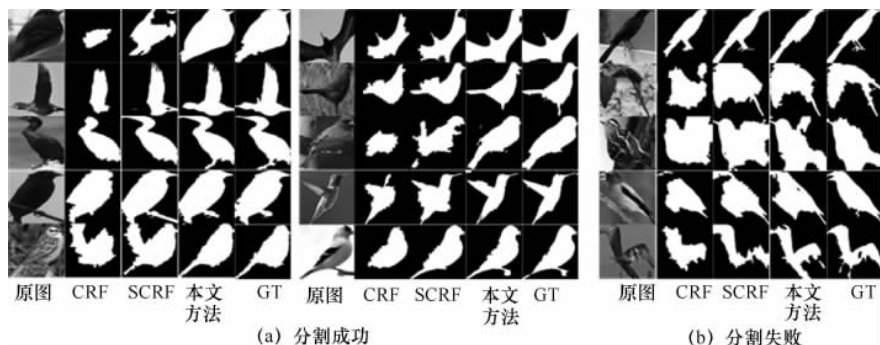


图 6 Caltech-UCSD Birds 200 数据集上的部分分割结果

对联合模型的算法复杂度进行分析。CRF 模型使用 LBP 算法预测一副图像标签的时间复杂度是 $O(EL^C)^{[28]}$, E 为图模型中边的数量, 边数与超像素个数 S 直接关联, 使用 Ncut 算法估算 E 为 $4 \times S$, C 为超像素块的大小, 即 CRF 算法复杂度为 $O(4SL \frac{m \times n}{s})$ 。ShapeBM 模型算法复杂度分为正向学习阶段时间复杂度和反向学习阶段时间复杂度, 其中正向学习的时间复杂度为 $O(N \times h_1 \times h_2 \times n)$, 反向学习的时间复杂度为 $O(N \times h_1 \times h_2 \times k_{cd_1} \times k_{cd_2})$, 则 ShapeBM 算法复杂度为 $O(N \times h_1 \times h_2 \times (k_{cd_1} + k_{cd_2} + n))$, 其中 N 表示 ShapeBM 虚拟输入节点个数, h_1 表示第一层隐层节点个数, h_2 表示第二层隐层节点个数, k_{cd_1} 、 k_{cd_2} 表示对比散度算法迭代次数, n 表示迭代次数。因为联合模型的能量函数是 SCRF 与 ShapeBM 的能量函数的叠加, 即 SCRF 与 ShapeBM 能量函数为线性关系, 故联合模型的整体复杂度为 $O(4SL \frac{m \times n}{s})$, 即联合模型在标记准确率提高的情况下, 整体复杂度并没有更复杂。

5 结束语

本文提出了一种结合 CRF 与 ShapeBM 的新标记模型, 将 ShapeBM 与 CRF 的结合建立在超像素基础上, 通过 pooling 技术克服图像超像素个数与 ShapeBM 输入不对应问题, 使得该模型有效地结合了 CRF 模型(增强相邻节点之间的局部约束)与 ShapeBM 模型(获得目标对象的整体形状信息作为对象的全局约束)的优点。在与其他方法的对比实验中验证了本文模型在标记准确性方面的优势。在未来的工作中, 可以将本文的二类标记应用到多类标记中, 可以做人体对象的部件检测、场景分析。另外, 可以在该模型中加入时间势, 在视频帧之间建立联系, 后期可将该模型扩展到视频对象分割上。

参考文献:

- [1] LAFFERTY J, MCCALLUM A, PEREIRAF C, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// ICML, June 28–July 1, 2001, Williams College, UK. New Jersey: IEEE Press, 2001.
- [2] BORENSTEIN E, SHARON E, ULLMAN S, et al. Combining top-down and bottom-up segmentation [C]//Conference on Computer Vision and Pattern Recognition, June 27–July 2, 2004, Washington, DC, USA. New Jersey: IEEE Press, 2004.
- [3] HE X, ZEMEL R S, CARREIRAPERPINAN M A, et al. Multi-scale conditional random fields for image labeling [C]// Conference on Computer Vision and Pattern Recognition, June 27–July 2, 2004, Washington, DC, USA. New Jersey: IEEE Press, 2004.
- [4] HE X, ZEMEL R S, RAY D, et al. Learning and incorporating top-down cues in image segmentation [C]// European Conference on Computer Vision, May 7–13, 2006, Graz, Austria. New Jersey: IEEE Press, 2006: 338–351.
- [5] SHOTTON J, WINN J, ROTHER C, et al. Texton boost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context [J]. International Journal of Computer Vision, 2009, 81(1): 2–23.
- [6] ZHANG L. A unified probabilistic graphical model and its application to image segmentation [J]. Rensselaer Polytechnic Institute, 2009(3).
- [7] SMOLENSKY P. Information processing in dynamical systems: foundations of harmony theory [M]. Cambridge: MIT Press, 1986(1): 194–281.
- [8] SALAKHUTDINOV R, HINTON G E. Deep Boltzmann machines [J]. Journal of Machine Learning Research, 2009, 5(2): 1967–2006.
- [9] ESLAMI S M, HEES N, WILLIAMS C K, et al. The shape Boltzmann machine: a strong model of object shape [J]. International Journal of Computer Vision, 2014, 107(2): 155–176.
- [10] KAE A, SOHN K, LEE H, et al. Augmenting CRFs with Boltzmann machine shape priors for image labeling [C]// Conference on Computer Vision and Pattern Recognition, June 23–28, 2013, Portland, Oregon, USA. New Jersey: IEEE Press, 2013.
- [11] KAE A, MARLIN B M, LEARNEDMILLER E G, et al. The shape-time random field for semantic video labeling [C]// Conference on Computer Vision and Pattern Recognition, June 23–28, 2014, Columbus, OH, USA. New Jersey: IEEE Press, 2014.
- [12] CHEN F, YU H, HU R, et al. Deep learning shape priors for object segmentation [C]//Conference on Computer Vision and Pattern Recognition, June 23–28, 2013, Portland, Oregon, USA. New Jersey: IEEE Press, 2013.
- [13] CREMERS D, SCHMIDT F R, BARTHEL F, et al. Shape priors in variational image segmentation: Convexity, Lipschitz continuity and globally optimal solutions [C]//Conference on Computer Vision and Pattern Recognition, June 24–26, 2008, Anchorage, Alaska, USA. New Jersey: IEEE Press, 2008.
- [14] ARBELAEZ P, HARIHARAN B, GU C, et al. Semantic segmentation using regions and parts [C]//Conference on Computer Vision and Pattern Recognition, June 16–21, 2012, Providence, RI, USA. New Jersey: IEEE Press, 2012.
- [15] MURPHY K, WEISS Y, JORDAN M I, et al. Loopy belief



- propagation for approximate inference: an empirical study[C]// 15th Conference on Uncertainty in Artificial Intelligence, July 30–August 1, 1999, Stockholm, Sweden. New Jersey: IEEE Press, 1999.
- [16] SCHMIDT M. minFunc: unconstrained differentiable multivariate optimization in Matlab [EB/OL]. [2016-07-10]. <http://www.di.ens.fr/~mschmidt/Software/minFunc.html>.
- [17] SAUL L K, JAAKKOLA T S, JORDAN M I, et al. Mean field theory for sigmoid belief networks [J]. Journal of Artificial Intelligence Research, 1996(13).
- [18] MNIH V, LAROCHELLE H, HINTON G E, et al. Conditional restricted Boltzmann machines for structured output prediction[C]// Conference on Uncertainty in Artificial Intelligence, Aug 15–17, 2012, Catalina Island, USA. New Jersey: IEEE Press, 2012.
- [19] WANG L, SHI J, SONG G, et al. Object detection combining recognition and segmentation[C]//Asian Conference on Computer Vision, November 18–22, 2007, Tokyo, Japan. New Jersey: IEEE Press, 2007.
- [20] WELINDER P, BRANSON S, MITA T, et al. Caltech-UCSD Birds 200[J]. California Institute of Technology, 2010.
- [21] YANG J, SAFAR S, YANG M H. Max-margin Boltzmann machines for object segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 23–28, 2014, Columbus, OH, USA. New Jersey: IEEE Press, 2014.
- [22] ACHANTA R P, SHAJI A, SMITH K M, et al. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274–2282.
- [23] MALIK J, BELONGIE S, SHI J, et al. Textons, contours and regions: cue integration in image segmentation [C]//ICCV, September 20–25, 1999, Kerkyra, Corfu, Greece. New Jersey: IEEE Press, 1999.
- [24] MARTIN D R, FOWLKES C C, MALIK J, et al. Learning to detect natural image boundaries using brightness and texture[C]// Conference on Neural Information Processing Systems, December 8–13, 2003, Providence, USA. New Jersey: IEEE Press, 2003.
- [25] HUANG G B, NARAYANA M, LEARNED MILLER E G, et al. Towards unconstrained face recognition [C]//Conference on Computer Vision and Pattern Recognition, June 24–26, 2008, Anchorage, Alaska, USA. New Jersey: IEEE Press, 2008.
- [26] COHN T. Efficient inference in large conditional random Fields[M]. Berlin: Springer, 2006: 606–613.

[作者简介]



王浩(1992–),男,宁波大学信息科学与工程学院硕士生,主要研究方向为计算机视觉与模式识别。



郭立君(1970–),男,博士,宁波大学教授,主要研究方向为计算机视觉与模式识别、移动互联网及其应用。



王亚东(1990–),男,宁波大学信息科学与工程学院硕士生,主要研究方向为计算机视觉与模式识别。



张荣(1974–),女,博士,宁波大学副教授,主要研究方向为计算机视觉与信息安全。