基于分层区域合并的自然场景理解①

孙丽坤, 刘 波

(北京工业大学 计算机学院, 北京 100124)

摘 要: 针对自然场景理解问题, 利用图像中的层次结构, 提出了一种基于分层合并的图像场景理解方法. 该方法通过不断合并相邻区域, 直到合并出图像中的各个对象为止; 最终得到一个合并森林, 森林里的每棵树对应图像中的一个对象. 我们设计了一个机器学习模型来描述合并过程、一种贪心推理方法来求解最优的合并森林以及一种基于最大间隔的学习方法来训练模型中的参数, 同时采用分层聚类来进行参数的初始化. 本文方法可以看成为图像语义理解而设计的一种深度学习方法. 实验效果令人满意.

关键词: 自然场景理解; 层次结构; 森林结构; 最大间隔; 贪心推理; 聚类

Parsing Natural Scenes Based on Hierarchical Region Merge

SUN Li-Kun, LIU Bo

(College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The goal of scene understanding is to recognize what objects in the image and where the objects are located. Hierarchical structure is commonly found in the natural scene images. This structure not only can help us to identity the objects but also how the small units interact to form the whole objects. Our algorithm is based on the level structure. We merge the neighboring segments continuously until they combined into the whole object. The result is a forest which contains several trees, one tree commonly represents one object. We introduce a machine learning model to describe the merge process, greedy inference to compute the best merge trees, and the max margin to learn the parameters. We cluster the segments features to initialize the parameter. The experiment result could be accepted.

Key words: parsing natural scenes; hierarchical structure; forest structure; max margin; greedy inference; clustering

场景理解的目的是识别图像中的对象和对象所在位置. 例如我们希望计算机能从图 1 中的左上角图像中识别出树木、建筑物、人和草等对象, 并标识出这些对象分别对应图像中的哪些区域. 场景理解是现在计算机视觉领域的一个热点研究方向, 因为现在网络上图像数量成爆炸式增长, 如果计算机能自动有效地理解这些图像, 就能帮助我们高效地利用这些图像资源, 从而方便我们的生活. 但场景理解也是一个难点问题, 因为图像中经常出现的遮挡、光亮度、角度旋转、大小比例、对象的多样式等问题, 使得图像分割、预处理和特征提取都比较困难, 从而影响识别效果.

经过研究者们多年的努力, 该领域提出了许多较好的方法, 这些方法(Schmid[1], Socherd[2], Rabinovich[3], Shotton[4]等)大致可以分为两种思路: ①基于超像素(Superpixels). 图像首先被分割成多个小片段, 这些小片段称为超像素. 这类方法把超像素作为分类单元, 即模型为图像中的每个超像素计算类别. 当所有超像素的类别计算完毕时, 由同类别相邻超像素组成的对象也就识别完成. 典型的方法①是基于条件随机场[4], ②基于层次结构. 图像中的复杂对象是由一系列的简单部分组成, 而这些简单部分是由更简单的小区域组成, 这种一层一层的组成就构成了图像的层次结构. 这类方法利用图像中的层次结构, 由简单区域不断合并组成复杂区域, 直到构成一个对象或者整幅图为止.

① 基金项目:安徽省电力公司 2013 科技项目

收稿时间:2014-03-19;收到修改稿时间:2014-04-14

若把简单区域看作孩子节点, 它们合并成的复杂区域为其父亲节点, 则在不断合并的同时也在构造树结构, 并且是自底向上构造. 典型的方法是基于递归神经网络的场景理解[5], 它不断地合并相邻区域直到构成整幅图像, 其树结构的根节点为整幅图, 叶节点为图像被分割后的原始小区域. 这两种思路的区别之处在于前者把图像看作小区域的集合, 是非结构的; 后者则把图像看作是这些小区域结构化的合并结果.

方法[5]的识别效果很好, 但也存在一些缺点: ①其合并过程和分类过程是分开进行的, 它先对区域进行合并, 然后利用合并好的树结构做分类; ②由于图像中不同类别的多对象存在, 把区域合并成完整图像构造成一棵树是不必要的; ③整个合并过程用同一个区域特征参数是不合理的, 因为随着合并的不断进行, 区域也在不断增大, 同类别的大区域和小区域的特征存在一定的差异.

我们的方法是基于层次结构的, 如图 1 所示. 首先图像被分割成多个小区域, 然后提取小区域的特征, 利用特征计算区域对的分数, 选择满足条件的区域对进行合并, 直到没有符合条件的区域对合并为止. 最终的结果为一个森林, 森林里有多棵树, 每棵树都是相邻区域的合并结果. 如图 1 中合并森林中有 4 棵树, 每个三角形代表一棵树, 我们详细展示了建筑物这棵树的构造过程. 这四棵树拼接起来构成整幅图像. 此时每棵树都有其可能类别, 现只需从其可能类别中找出最可能的类别即可, 构成树结构的所有小区域所在图像中的位置即为对象位置. 在合并过程中, 我们会为每次合并后的新区域预测可能类别, 这样做的目的是: ①保证合并过程和分类过程的同时进行; ②局部决策对最终结果产生影响, 因为区域的可能类别会制约其以后的合并决策.

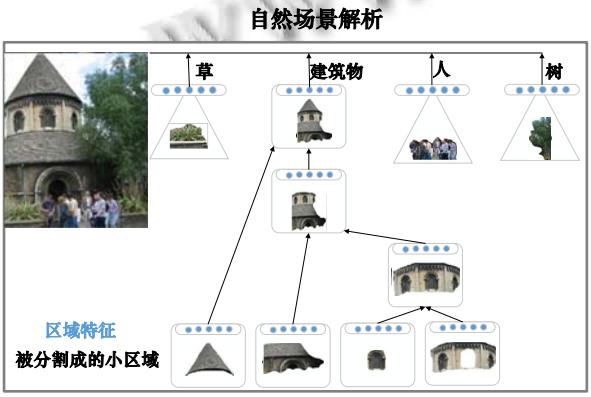

图 1 自然场景图像的解析过程

1 模型

每幅图像作为输入 x 共包括两个部分: ①一个向量集合 $Fea=\{f_1,\ldots,f_{segs}\}$, 表示该图像所有区域的特征; ②一个对称的邻接矩阵 A, 当第 i 区域和第 j 区域相邻时有 $A(i,j)=1$, 只有相邻的区域才可以合并.

一幅图像中相邻区域间的合并有多种可能, 如图 2 下方展示了左上角简单图像中灰色类别对象的四种可能合并. 所有可能的合并形成的森林及森林中每棵树的类别构成一个集合, 记为 $\kappa(x)$. 我们要做的就是在该集合中寻找那些正确的合并森林 $Y(x,l)$, 所谓正确森林就是森林里的每棵树都由相同类别的区域合并而成. $Y(x,l)$ 可以利用样本的类别标签信息 l 来构造.

为每个类别 c ($1\le c\le m$, m 是总类别数)学习一些权值 $W^c=\{W^{ci}\}$, 其中 $1\le i\le u$, u 为权值数目, 通常不同的类别所学到的权值数目 u 不同, 详见第 3 节; 所有类的权值组成参数 $W=\{W^c\}$. 我们考虑最优的合并森林以及每棵树的类别由如下最优化问题确定:

$$y^*=\arg\max_{y\in\kappa(x)} score(x,y\,|\,W) \tag{1}$$

其中, $y^*=(F^*,L^*)$ 为最优的合并森林及森林中每棵树的类别标签, 合并森林 $F^*=\{T_1,\ldots,T_n\}$, 每棵树 $T_i=\{R_1,\ldots,R_n\}$, R_i 是树中的节点区域, 由两个区域合并而成.

此最优化问题既要确定最优的合并森林, 同时也要为每次合并后区域的预测类别 $L(R_i)$.

1.1 贪心求解最优 F^*,L^*

由于相邻区域间的合并有指数数量级的可能性, 因此直接搜索空间 $\kappa(x)$ 是不可行的, 我们采用贪心的近似方法来求解该问题. 下面我们来解释整个合并及分类过程.

首先, 通过邻接矩阵 A, 找出所有的相邻区域对加入到 C 中, C 是可合并区域对的集合:

$$C=\{[a_i,a_j]: A(i,j)=1\}$$

以图 2 中的简单图像为例, 有 C={[a_1, a_2], [a_1, a_4], [a_2, a_6], [a_3, a_4], [a_4, a_6], [a_5, a_6]}, 由于 A 是对称矩阵, 故只需考虑上三角或者下三角即可.

然后为 C 中的所有区域对计算局部分数, 假设区

域对 i 和 j 合并成的超级区域特征为 $f_{(i,j)}$，其局部分数为:

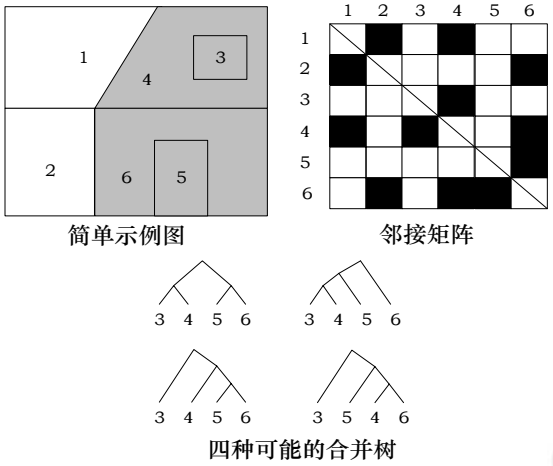

图 2　简单示例图、邻接矩阵及合并情况

$$localScore_{(i,j)}^{c}=\max_{k} W^{ck}\cdot f_{(i,j)} \tag{2}$$

其中, 局部分数为一个向量, 向量中的元素与参数 W 中的各权值一一对应. 我们利用该局部分数为合并后的新片段预测类别, 哪个类别对应的内积越大, 区域就越可能属于该类别. 这是由参数 W 的特性决定的, 在第 3 部分会阐述该特性. 通常我们会为新区域计算 r 个可能的类别, $1\le r\le m-1$, 具体个数根据情况自己选定. 在实验中我们设定 $r=3$, 即每次计算 3 个可能类别. 注意预测过程实际在合并区域完成之后, 我们只为选择合并的那一对区域做类别预测.

然后利用局部分数来计算 C 中所有区域对的分数, 有:

$$score_{(i,j)}=\max_{c}(localScore_{(i,j)}^{c}) \tag{3}$$

得到所有的分数后, 对这些分数进行降序排列, 然后由高到低依次检查是否满足合并条件, 直到找到符合条件的区域对为止. 假设现在判断区域对 a_i 和 a_j, 合并条件如下:

1) 若 a_i 和 a_j 都是叶子区域, 即从未合并过, 则合并;

2) 若有一个是超级区域, 则需判断是否满足标签一致性, 即如果先前已经确定的区域类别为 c, 则在该区域周围继续合并进来后该区域仍然属于类别 c. 假设 a_i 是超级区域, 其可能类别为 $L(a_i)$, 合并后区域 a_{ij} 的可能类别为 $L(a_{ij})$, 若 $L(a_i)\bigcap L(a_{ij})\neq\phi$, 则合并, 否则不合并;

3) 若两个都是超级区域, 且满足 $L(a_i)\bigcap L(a_j)\neq\phi$ 和 $score_{(i,j)}>(score_i+score_j)/2$ 这两个条件, 即它们有共同的可能类别, 且合并后的分数比它们分数和的平均值大则合并否则不合并. 其中 $score_i$ 和 $score_j$ 已经在之前的合并中计算并存储, 无需重新计算.

假设 a_i 和 a_j 是当前需要合并的区域对, 且它们合并后的新区域为 a_{ij}, 合并时需做如下更新:

(a) 更新特征集合 Fea, 把新区域特征 $f_{(i,j)}$ 加入到集合中, 并从 Fea 中删除 f_i 和 f_j;'

(b) 更新邻接矩阵 A, 增加新行和新列 a_{ij}, 并把与原 a_i 和 a_j 区域相邻的区域位置标 1, 从 A 中删除 a_i、a_j 行和列;

(c) 更新集合 C, 需删除掉一切与 a_i 和 a_j 相邻的区域对, 增加新区域 a_{ij} 的邻接信息, 即那些与 a_i 和 a_j 相邻的区域与 a_{ij} 相邻;

如图 2, 假设我们把区域 3 和 4 合并为区域 7, 则此时需从 C 中删除 $\{[a_3,a_4],[a_1,a_4],[a_4,a_6]\}$, 增加 $\{[a_1,a_7],[a_6,a_7]\}$.

当合并更新完毕后, 为新区域预测类别时有 2 种情况: ①、满足条件(1)和(2)的合并, 利用局部分数来预测类别, 具体过程: (a)找出各类别在 $localScore$ 中的最大值; (b)找出前 r 个最大值对应的类别作为其可能类别. ②、满足条件(3)的合并, 则新区域的类别为两个旧区域类别的交集, 即 $L(a_{ij})=L(a_i)\bigcap L(a_j)$.

此时一个合并过程结束, 不断重复以上过程直到没有符合条件的区域对为止, 整个合并过程结束.

不再合并的大区域就构成以其为根的树结构, 此时每棵树都有等于或小于 r 个的可能类别, 从中选择那个使整棵树分数最高的类别作为该树类别:

$$c_R^{*}=\max_{c}\sum_{r\in R}\max_{i} W^{ci}\cdot f_r \tag{4}$$

因为在合并过程中已经计算了各个中间分数并保存, 此时只需取出每个节点相应的类别分数做和比较即可. 与此同时, 我们把分数最高的那个类别在计算树分数时选择的那些权值加入到集合 I 中.

最后我们给出整幅图的分数, 即森林的分数:

$$score\ (image\ )=\sum_{R}\sum_{r\in R}\max\ {}_{i}W^{c_R^{*}i}\cdot f_r \tag{5}$$

1.2 计算正确森林的分数

在监督学习时, 需要为一幅图像 x 构造一个正确

的森林. 我们利用区域的标签信息, 只合并相同类别的区域对, 直到所有相同类别的区域合并完毕, 此时一个正确的森林也构造完成. 该详细合并过程如下:

首先, 通过邻接矩阵 A, 找出所有的相同类别的相邻区域对并加入到 C 中:

$$C=\{[a_i;a_j]: A(i,j)=1, L(a_i)=L(a_j)\}$$

以图 2 为例, 有 C={[a_1, a_2], [a_3, a_4], [a_4, a_6], [a_5, a_6]}.

然后计算 C 中所有相邻区域的分数, 假设 a_i 和 a_j 合并, 用 $f_{(i,j)}$ 表示合并后超级区域的特征, 则得到分数:

$$score_{(i,j)}=\max_{W^{ci}} W^c\cdot f_{(i,j)} \tag{6}$$

其中, W^c 为区域的类别所对应的参数权值集.

从这些分数中选择得分最高的区域对进行合并, 更新集合 Fea、C 和矩阵 A, 注意在更新集合 C 时需保证添加的相邻区域对为同类别. 合并完毕后, 将该分数所对应的权值 W^{ci} 加入集合 G 中.

重复以上过程, 直到 C 为空, 即相同类别的相邻区域全部合并完毕.

1.3 最大间隔参数学习

现在我们讨论如何学习该模型参数. 在最大间隔模型中, 对于训练集中的所有样本 $(x_i,y_i), i=1,...,N$ 都要满足正确森林的分数大于错误森林的分数. 其正规化风险函数如下:

$$\min_{W,\xi_i\geq 0}\frac{1}{2}W^TW+\alpha\cdot\sum_{n=1}^{N}\xi_i \tag{7}$$

$$st.\forall i, y\neq y_i,$$

$$\max_{y_i\in Y(x_i,l_i)} score(x_i,y_i\,|\,W)-\max_{y\in K(x_i),y\notin Y(x_i,l_i)} score(x_i,y\,|\,W)>\xi_i$$

其中 α 是正规化参数.

在训练时, 根据参数 W, 为每幅图像 x_i 构造 2 个森林, 一个是森林 y, 另一个是利用标签信息构造的正确森林 $Y(x_i,l_i)$. 这两个过程, 分别对应 2.1 节和 2.2 节. 除了森林, 我们还得到两个参数集合 I 和 G, 里面分别保存了训练森林和正确森林在合并过程中选择的那些参数权值. 算法 1 是由式(7)推导出的学习过程.

算法 1 最大间隔参数学习算法.

for each W^{ci} in W:

if W^{ci} in I but not in G

$$\Delta W^{ci}=\lambda\cdot(-W^{ci}+\alpha\cdot\sum_{r\in R} f_r)$$

elseif W^{ci} in G but not in I

$$\Delta W^{ci}=\lambda\cdot(-W^{ci}-\alpha\cdot\sum_{r\in R} f_r)$$

else

$$\Delta W^{ci}=\lambda\cdot(-W^{ci})$$

End

$$W=W+\Delta W$$

上述算法我们可以通俗地解释为: 对于参数 W^{ci}, 如果正确森林选择它, 而训练森林没选择, 此时应该增大该权值; 如果训练森林选择它, 而正确森林没有, 则应该减小该权值.

2 参数初始化

我们希望参数 $W=\{W^c\}$ 有这样的特点: 每个类别对应的权值参数与该类别的区域特征内积较大, 而与其他类别的区域特征内积较小. 这样我们就可以利用内积分数来预测区域的类别, 哪个类别的参数内积越大, 区域就越可能属于哪个类别. 利用聚类, 把区域特征的聚类中心作为该类别的权值参数. 这样相同类别的参数与特征具有相似性, 而不同类别的参数与特征则存在差异, 正好满足要求.

详细的聚类过程如下: 对于类别 c, 将训练样本中类别为 c 的所有区域特征组成原始数据集, 然后进行如下计算:

① 对该数据集用 k 均值进行聚类, 取类中心作为该类的参数权值 W^{ci};

② 将区域按照到类中心的距离进行升序排列, 然后将这些区域依次与其相邻的同类别区域合并;

③把合并后新区域的特征加入数据集中, 并从中删除合并之前的旧区域对;

重复以上过程, 直到所有训练图像中所有类别为 c 的相邻区域合并到一块为止.

通过聚类, 取其中心可以获取类别区域的一般性特征; 通过合并区域可以获取该类别大块区域的特征, 因此不断地迭代合并聚类, 我们学习到的参数能够有效抓取类别不同区域大小时的一般性特征.

在 1 步使用 k 均值算法时, 需要明确给出聚簇个数, 而由于图像数据的复杂性, 我们无法预知每个类别每次应该聚多少簇, 因此我们采用了让聚簇数不断增大, 直到所有点到类中心距离之和比上次迭代没有明显减少为止, 认为此时的聚簇数是比较合适的.

3 特征

我们用文献[8]中 3.1 节描述的方法进行区域特征的计算, 区域的特征是其包含所有像素特征的均值和协方差. 像素特征包括 2 部分: 1、像素基本特征, 如颜色、纹理等, 共 17 维; 2、像素分类器分数. 首先利用带标签像素的基本特征为每个类别用 gentleboost 方法训练分类器, 然后用这些分类器所有像素计算类别分数, 把这些分数作为像素的特征, 因为数据集有 8 个类别, 故有 8 个像素分类器分数.

在实验中我们计算了三套特征: (a)完整的像素特征, 计算后的区域特征维数为 350; (b)只包含像素的基本特征, 为 17 维, 则计算后的区域特征维数为 170; (c)不同于前 2 种计算的协方差, 我们把基本特征的均值、方差和像素特征的直方图作为区域特征. 若把每个像素特征分割 20 个刻度做直方图, 则计算后的区域特征为 374 维.

当两个区域合并时, 需计算合并后的超级区域的特征. 此时只需根据旧片段特征, 利用相应的均值和协方差公式计算即可.

均值公式:

$$f_{(i,j)} = \frac{(f_i \cdot p_i + f_j \cdot p_j)}{p_i + p_j} \tag{9}$$

其中, p_i 和 p_j 分别是 2 个旧区域所包含的像素数目. 协方差公式也可简单导出, 由于篇幅有限, 在此不再详述.

4 实验

我们使用 Standford background 数据集, 包括 572 张训练样本和 143 张测试样本, 共有包括建筑物、水、天空等在内的 8 个类别.

4.1 权值参数的数量

对于参数初始化部分, 需要调节的参数就是阈值, 阈值的大小决定了学习到的权值数目, 例如我们设定不同的阈值最多为每个类别学习到的权值数分别为{53,71,53,54,37,80,31,76}共 455 个, 最少的权值数分别为{11,33,17,11,6,22,5,38}共 143 个; 在实验中, 我们尝试了几组参数来训练模型, 其中得到最好效果的参数权值数目为{20,53,32,25,11,40,10,63}共 345 个.

4.2 实验结果

在模型学习中, 可调参数为 λ 和 α 分别是步长和正规化参数. 我们用第 4 节的三套特征分别进行模型学习, 得到的正确率如表 1 所示, 此处正确率是以像素为计算单元, 且是用数据集中的所有训练样本进行训练, 对所有测试样本测试的结果.

表 1 不同特征正确率的比较

特征	正确率
(a)完整的像素特征	42%
(b)基本的像素特征	39%
(c)像素特征直方图	40%

对于这三套特征, 特征(a)耗费的时间最长, 因其在特征提取阶段需为像素训练 8 个分类器, 然后为所有的样本中像素计算分数. 其次是特征(c), 其维数最多, 但由于是直方图, 在特征中有一部分元素为 0. 最后是特征(b), 其特征维数最少.

当我们从测试样本中挑选 30 张简单图像(图像中最多有 3 个对象) 用学习好的模型进行测试时, 得到的正确率为 64%. 如图 3, 可以看出我们的模型对结构简单的对象识别率较好, 如水域、草原等, 而对于复杂对象的识别有待提高.

图 3 简单图像的实验效果图

6 结语

本文提出了一种利用图像中的层次结构来解析自然场景的方法. 通过不断合并相邻区域, 由简单场景不断组合成复杂场景直到构成整个对象为止. 在合并过程中, 我们为每次合并后的新区域预测类别, 并利用这些类别来制约该区域以后的合并决策. 最终的合并结果为一个森林, 通常森林中的每棵树为一个对象. 我们的方法不仅实现了场景理解, 还深层地解释小区域是如何合并构成对象的. 在模型中, 我们采用了最大间隔学习和贪心推理. 对于参数的初始化, 我们用聚类方法从区域特征学习, 使得参数具有预测区域类别的特性.

未来我们会在该算法的基础上, 对特征选择、参数初始化等方面做改进工作.

参考文献

1 Cordelia S. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. The Conference on Computer Vision and Pattern Recognition, New York, 2006.

2 Socher R, Fei-Fei L. Connecting modalities: Semisupervised segmentation and annotation of images using unaligned text corpora. The Conference on Computer Vision and Pattern Recognition, San Francisco, 2010.

3 Rabinovich A, Vedaldi A, Galleguillos C, eds. Objects in context. Conference on Computer Vision. Rio de Janeiro. 2007.

4 Shotton J, Winn J, Rother C, eds. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. The European Conference on Computer Vision, Graz, 2006.

5 Socher R, Lin CC, Ng AY, eds. Parsing natural scenes and natural language with recurisve neural networks. The International Conference on Machine Learning, Bellevue, 2011

6 Liu B, Fan HQ. Semantic labeling of indoor scenes from RGB-D images with discriminative learning. The international Conference on Machine Vision, London, 2013.

7 Gupta A, Davis L. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. The European Conference on Computer Vision, Marseille, 2008.

8 Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. The International Conference on Computer Vision, Kyoto, 2009.