分割位置提示的可变形部件模型快速目标检测

杨 扬 [1] 李善平 [1]

摘 要 针对滑动窗口目标检测方法需要穷举搜索目标、检测速度较慢的问题, 提出一种可变形部件模型候选点检测算法. 图像先经过两种不同原理的分割方法预处理, 尽量使至少一个分割接近目标真实位置, 分割的左上角附近称为候选点. 然后, 将可变形部件模型作为底层检测器, 模型的训练和测试都只在候选点上进行, 这大大提高了检测速度. 在 PASCAL 2007 数据集上的实验结果表明, 候选点检测在一半类别上的正确率超过了穷举搜索方法.

关键词 目标检测, 可变形部件模型, 图像分割, 隐支持向量机, 滑动窗口方法

引用格式 杨扬, 李善平. 分割位置提示的可变形部件模型快速目标检测. 自动化学报, 2012, **38**(4): 540−548

DOI 10.3724/SP.J.1004.2012.00540

Fast Object Detection with Deformable Part Models and Segment Locations′ Hint

YANG Yang[1] LI Shan-Ping[1]

Abstract Sliding window detectors need to compute overall scores on all the positions and scales in the image pyramid, which causes the detection speed to be relatively slow. In order to accelerate the detection speed, we propose a candidate points′ detection algorithm for deformable part models. Multiple segmentation algorithms are used for each image to generate image segments. The segment′s top-left corner is treated as a candidate detection point. We adapt mixture deformable part models as our underlying detectors. The detection operations are only carried on these candidate detection points to accelerate detection speed dramatically. We evaluate the detection performance of our approach on PASCAL 2007 challenge dataset and find that the candidate points′ detection is even better than exhaustive search.

Key words Object detection, deformable part models, image segmentation, latent support vector machine (latent SVM), sliding windows

Citation Yang Yang, Li Shan-Ping. Fast object detection with deformable part models and segment locations′ hint. *Acta Automatica Sinica*, 2012, **38**(4): 540−548

目标检测是在静态图像中得到目标对象的类别和位置, 这是机器视觉领域的一个基础问题. 图像的不同拍摄角度、光线情况、尺寸大小, 包括混杂的背景, 加上同类对象的内部差异使得此问题仍然是一个巨大的挑战. 当前研究的一个趋势是采用滑动窗口方法 (Sliding window methods)[1−6]. 这类方法将目标检测建模为分类问题, 通常可以表示为 “矩形窗口覆盖下的图像区域是否为目标类别的一个对象?”. 这样可以很直接自然地使用各种复杂的机器学习分类模型, 从训练图像中学习最优的检测器. 为了检测多种尺寸大小的目标, 图像往往需要重复缩放多次. 检测窗口将置于不同尺度的图像中的每个位置, 然后, 判断窗口覆盖的区域能否通过检测. 这种穷举搜索的方式虽然能保证不漏掉任何一个可能的位置, 但是有两个明显的缺点: 1) 检测效率低, 计算量太大, 计算复杂度是像素级别; 2) 检测过程中容易出现多个相互重合的窗口, 常常需要采用非极大值抑制 (Non-maximum suppression, NMS)[1] 等后处理合并压缩窗口区域得到目标的范围. 但是这些后处理大都是基于启发式的规则, 与训练阶段的监督脱节.

另一类方法将图像预先分割为若干区域, 然后对每个区域提取颜色、纹理、形状等特征, 问题转化为区域级别上的分类问题[7−10]. 最为普遍的方法是条件随机场 (Conditional random field, CRF)[7−9], 用于建模区域之间的空间关系, 输出联合类别. 属于目标类别的区域可以合并得到整个目标. 这样可以避免穷举搜索以及 NMS 等后处理. 这类方法不仅能得到目标的范围方框, 还能自然地得到目标更为细致的轮廓形状. 但是分割方法本身就会出现错误, 同时分割的粒度难以控制, 往往不易得到完整的目标. 以图 1 为例, 图 1 (a) 的分割结果完全失败了, 图 1 (b) 的目标 “鸟” 被分割为 “头部”、“身体”、“翅膀” 等. 这些分割的类别信息都会被标记为 “鸟”, 但实际上它们都不是 “鸟”, 只是 “鸟” 的一个部分. 这些被标记为同一类的分割从视觉上区别很大, 分成多个类反而更合理. 再比如, “小汽车” 的 “轮子” 分

收稿日期 2011-05-04 录用日期 2011-10-28
Manuscript received May 4, 2011; accepted October 28, 2011
本文责任编委 戴琼海
Recommended by Associate Editor DAI Qiong-Hai
1. 浙江大学计算机科学与技术学院 杭州 310027
1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027

割和 “货车” 的 “轮子” 分割从视觉上看很像, 但是它们被标为两个不同的类. 类别信息不准确相当于又增加了同类之间的差异, 给分类带来更大难度, 而且被错误分类的分割仍然会参与到最终目标范围的合并中. 在复杂的国际基准数据集上, 基于图像分割的方法能否与滑动窗口方法竞争仍然有待证明.

(a) 分割方法 1

(a) Segmentation method 1

(b) 分割方法 2

(b) Segmentation method 2

图 1　两个分割方法得到的结果

Fig. 1　The results of two segmentation methods

本文的目的是将两类方法的优点结合起来: 既使用高精确度可理解的窗口模型, 又能避免穷举搜索, 提高检测效率. 我们从 PASCAL VOC 2007[11] 数据集里的每一类都选择出多张图像用于观察分割出错的比例. 每张图像都采用两种不同原理的分割方法处理, 在经过较多分割实验之后, 我们观察到以下现象: 几乎每一张图都存在分割错误; 但是绝大多数图像中至少存在一个分割的左上角与目标真实范围方框的左上角非常接近. 比如图 1 中 “鸟” 的例子, 图 1 (a) 分割无法捕获目标, 图 1 (b) 可以. 从这些采样观察中可以得到两个推论:

1) 目标的范围应该由模型决定, 而不应该由分割组合决定;

2) 穷举搜索可以通过只关注分割的左上角附近区域而避免.

在理想情况下, 假设测试图像中目标对象左上角的真实位置已经给定, 那么只需要将检测窗口置于此位置, 即使不做穷举搜索也能获得基本相当的检测精确度. 所以, 首先要做的是采用多种不同原理的分割方法, 尽量保证至少有一个分割位置很接近目标真实位置. 这样, 检测只需要在这些候选点上进行.

接下来要做的是选择合适的检测模型. 在复杂的数据集上, 简单方法, 比如 “规则模板”[2], 往往比复杂的模型效果更好. 这是因为简单的方法更容易通过诸如支持向量机之类的判别式方法来训练. 复杂模型训练起来更困难, 特别是由于它们往往利用了隐信息 (Latent info). 但是可变形部件模型 (Deformable part model)[1] 在最近几年的 PASCAL 比赛中表现优异, 连续获得了一流的检测结果. 文献 [1] 采用了混合可变形模型, 使得模型能够表示更为丰富的类别内部的变化, 进一步提高了检测的结果. 所以, 本文采用混合可变形模型作为底层的检测器.

将上面两部分结合起来, 先对图像采用预分割, 然后模型的训练和测试都只在候选点上进行. 训练和测试的效率都有大幅度的提高, 在 PASCAL 2007 数据集上的实验结果表明, 候选点检测在一半类别上比穷举搜索精确度更高.

本文后续内容组织如下: 第 1 节简要介绍可变形部件模型; 第 2 节是候选点检测算法; 第 3 节是模型的训练过程; 第 4 节是实验结果和分析; 第 5 节讨论相关工作; 最后, 第 6 节得到结论并讨论进一步的工作.

1　可变形部件模型

可变形部件模型是滑动窗口方法的一个代表. 这种模型由两层滤波器组成: 一个用于覆盖目标整体的根滤波器以及若干用于覆盖目标某个主要部件的局部滤波器. 根滤波器的作用是捕获目标的整体轮廓特征, 而局部滤波器能够捕获目标某个具有明显判别作用的局部特征. 比如用两层滤波器对人脸建模, 根滤波器可以训练成人脸的粗糙线条和轮廓, 而局部滤波器可以用来表示人脸的局部细节特征, 如眼睛、鼻子和嘴等.

图像被表示为一个特征位图 (Feature map) G, 它是一个二维矩阵, 其中每个元素是一个 d 维向量用于表示图像对应的局部区域. 模型的每个滤波器 F 是一个矩形模板, 其中每个元素是一个 d 维权重向量. 将模板的左上角置于特征位图的坐标 (x, y)

处, 两个矩形内部对应位置上做 "点积", 所得分数定义为

$$\sum_{x',y'} F[x',y'] \cdot G[x+x', y+y'] \tag{1}$$

分数越高, 表明滤波器覆盖区域成为对象的可能性越大. 由于滤波器的尺寸一般是固定的, 为了能检测多种尺度大小的对象, 常常需要将图像通过下采样缩小去构造特征金字塔. 图像先通过反复的平滑操作 (Smoothing) 和下采样 (Subsampling) 得到 N 层, 然后在每一层计算 M 种不同尺度 (Scale) 的特征位图, 相邻尺度大小为 2 倍关系. 这 M 个特征位图组成一个图像音阶 (Image octave), 越向下尺度越大, 图像特征越细致. 检测的时候, 先确定起始层, 在那层的图像音阶中, 局部滤波器将置于根滤波器相邻下方 2 倍尺度更细致的特征位图中. 全体 $M \times N$ 个特征位图统称为特征金字塔 (Feature pyramid) H.

模型被表示为一个 $n+2$ 元组 $(F_0, P_1, \cdots, P_n, b)$, F_0 是根滤波器, P_i 是第 i 个局部滤波器的外覆, b 是一个实数. P_i 被表示为 $(F_i, \boldsymbol{v}_i, \boldsymbol{d}_i)$, F_i 是局部滤波器; $\boldsymbol{v}_i$ 是一个二维向量, 指定 F_i 到 F_0 的 "锚" 位置; $\boldsymbol{d}_i$ 是一个四维向量, 代表 F_i 到 F_0 的变形开销系数.

单次检测会输出 $n+1$ 个位置, $\boldsymbol{z} = (\boldsymbol{p}_0, \cdots, \boldsymbol{p}_n)$, 其中, $\boldsymbol{p}_i = (x_i, y_i, l_i)$ 是第 i 个滤波器置于特征金字塔的层数和坐标. 单次检测的总分数定义为

$$s(p_0, \cdots, p_n) = \sum_{i=0}^{n} F_i' \cdot \phi(G, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(\mathrm{d}x_i, \mathrm{d}y_i) + b \tag{2}$$

其中 $(\mathrm{d}x_i, \mathrm{d}y_i) = (x_i, y_i) - (2(x_0, y_0) + v_i)$ 代表了第 i 个部分到根滤波器锚位置的偏移, $\phi_d(\mathrm{d}x, \mathrm{d}y) = (\mathrm{d}x, \mathrm{d}y, \mathrm{d}x^2, \mathrm{d}y^2)$ 是变形特征.

如果令 $\boldsymbol{d}_i = (0, 0, 1, 1)$, 则变形开销即为第 i 个部分到其锚位置的距离的平方. 模型的总分数是各个模板分数之和减去变形开销.

如果将根滤波器的位置固定在 $\boldsymbol{p}_0$ 处, 下面就是要为每个局部滤波器寻找最佳位置, 使得总分数最大:

$$s(\boldsymbol{p}_0) = \max_{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n} s(\boldsymbol{p}_0, \cdots, \boldsymbol{p}_n)$$

滑动窗口方法会在整个特征金字塔上的所有位置计算总分数. 分数超过模型阀值的根位置代表了一次通过的检测, 而所用的部分模型覆盖的区域和根滤波器覆盖的区域合起来构成了对象的范围. 为了加快检测速度, 动态规划和泛化距离迁移[12] 等编程技巧常被引入优化检测过程. 但总的来说, 由于没有可靠的对象位置提示, 滑动窗口方法不得不检测所有可能的位置, 检测的速度还是比较慢.

2 候选点检测算法

一旦给定了候选检测点, 在图像每个位置都计算分数就没有必要了. 问题简化为对于一个固定的根滤波器位置, 如何安置局部滤波器使得在此位置的总分数最大. 实际上, 根据式 (2) 可以独立放置每一个局部滤波器, 因为它们彼此之间相互独立. 单个局部滤波器 F_i 对分数的贡献为

$$D_i = F_i' \cdot \phi(H, p_i) - d_i \cdot \phi_d(\mathrm{d}x_i, \mathrm{d}y_i) \tag{3}$$

最直接的方法就是将此局部滤波器置于第 l_i 层的所有位置, 然后计算每个位置的分数, 取最大值. 但是这样是没有必要的, 试想一下, 如果局部滤波器离根滤波器的锚位置很远, 即使得分较高, 也不能将其作为一个合理位置. 这样对象岂不是过度变形夸张了? 所以只需要计算锚位置附近的 D_i 的值. 令此滤波器的高度和宽度分别记作 h 和 w, 根滤波器的检测位置记作 $\boldsymbol{p}_0 = (x_0, y_0, l)$, 那么局部滤波器的理想位置应该是 $(x = 2x_0 + v_{i1}, y = 2y_0 + v_{i2}, l_i)$, 问题定义为

$$\begin{aligned} \max_{x_i, y_i} \quad & F_i' \cdot \phi(H, (x_i, y_i, l_i)) - d_i \cdot \phi_d(\mathrm{d}x_i, \mathrm{d}y_i) \\ \text{s.t.} \quad & x - w < x_i < x + w \\ & y - h < y_i < y + h \end{aligned} \tag{4}$$

具体计算的时候, 由于局部滤波器在某一范围的得分在相邻的根位置之间可以共享, 所以可以考虑只计算一次缓存起来, 从而进一步加快速度. 计算复杂度从原先的像素级别降低到了区域级别. 令图像的平均区域数为 N, 每个区域左上角附近 8×8 区域作为候选点, 特征金字塔为 L 层, 局部滤波器 M 个, 平均大小为 $H \times W$, 计算一次 D_i 值的时间为 t, 则一张图的平均检测时间大约为: $L \times 64N \times M \times 2H \times 2W \times t$. 典型的测试图像大小大约为 480 像素 $\times$ 320 像素, 区域数量大约 30 个, 候选点计算次数与穷举检测之比大约为 $(30 \times 64)/(480 \times 320) = 1/80$.

3 模型训练

3.1 隐支持向量机

单个可变形模型的检测结果 $\boldsymbol{z} = (\boldsymbol{p}_0, \cdots, \boldsymbol{p}_n)$ 的总得分可以表示为两个向量 "点积" 的形式: $\boldsymbol{\beta} \cdot \boldsymbol{\psi}(H, z)$, 其中:

$$\boldsymbol{\beta} = (F_0', \cdots, F_n', d_1, \cdots, d_n, b) \tag{5}$$

是模型的参数, 而

$$\begin{aligned}\boldsymbol{\psi}(H,z)=(&\phi(H,p_0),\cdots,\phi(H,p_n),\\&-\phi_d(\mathrm{d}x_1,\mathrm{d}y_1),\cdots,\\&-\phi_d(\mathrm{d}x_n,\mathrm{d}y_n),1)\end{aligned}\tag{6}$$

是全体滤波器所覆盖位置的图像特征向量与变形开销向量的连接.

按照同样的方式, 混合模型的参数向量 $\boldsymbol{\beta}$ 可以表示为每个可变形模型参数的连接:

$$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_m)\tag{7}$$

而其所作用的 $\boldsymbol{\psi}(H,z)$ 为一个稀疏向量, 其定义为

$$\boldsymbol{\psi}(H,z)=(0,\cdots,0,\psi(H,z'),0,\cdots,0)\tag{8}$$

两个向量 "点积" 的形式为模型和线性分类器建立了联系. 向量 $\boldsymbol{\beta}$ 可以看作分类器的参数, 向量 $\boldsymbol{\psi}(H,z)$ 可以看作单个实例的特征. 这样一来, 模型参数可以通过隐支持向量机 (Latent support vector machine, Latent SVM) 学习而得到. 令训练集合为 $D=(\langle x_1,y_1\rangle,\cdots,\langle x_n,y_n\rangle)$, 其中, $y\in\{-1,1\}$, Latent SVM 的目标函数为

$$L_D(\beta)=\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^{n}\max(0,1-y_if_\beta(x_i))\tag{9}$$

其中

$$f_\beta(x)=\max_{z\in Z(x)}\beta\cdot\Phi(x,z)\tag{10}$$

如果每个 $\boldsymbol{x}_i$ 只有一个可能的隐变量, 即 $|Z(\boldsymbol{x}_i)|=1$, 那么 $\boldsymbol{f}_\beta$ 与 $\boldsymbol{\beta}$ 是线性关系. 此时, Latent SVM 就成了普通的线性 SVM. 在目标检测这个具体任务中, x 代表单张图像的特征金字塔, $z\in Z(x)$ 代表被模型覆盖的一个检测窗口, $\boldsymbol{\beta}\cdot\Phi(x,z)$ 代表模型在此窗口上的分数, 而 $f_\beta(x)=\max_{z\in Z(x)}\beta\cdot\Phi(x,z)$ 代表此图像上的最高分 (对应窗口就是最佳检测). 以普通 SVM 的术语来说, $(f_\beta(x_i),y_i)$ 才是一个训练实例, 如果最佳检测的分数超过阀值, 那么对应窗口范围之内就被认作目标对象. 一旦隐变量确定, 也就是说每张图像的检测窗口确定, Latent SVM 就可以转化为普通的 SVM 优化问题.

3.2 迭代训练

实际情况中, 由于训练图像给定了对象的范围方框, 所以范围方框之内的特征可以作为一个正类数据, 此图像剩下的区域以及其他不包括目标对象的图像的任何区域都可以作为负类数据. 即使只考虑候选点, 单张图像也可以产生 10^3 级别的负类数据. 将全部数据统一起来训练很困难, 一般考虑构造训练集只包括正类数据和 "困难" 负类数据. 对于全部候选点 D 上的当前模型参数 $\boldsymbol{\beta}$ 来说, 容易集合与困难集合的定义为

$$H(\boldsymbol{\beta},D)=\{(x,y)\in D|yf_{\boldsymbol{\beta}}(x)<1\}\tag{11}$$

$$E(\boldsymbol{\beta},D)=\{(x,y)\in D|yf_{\boldsymbol{\beta}}(x)>1\}\tag{12}$$

困难集合 $H(\boldsymbol{\beta},D)$ 是被当前模型错误分类或者是在间隔 (Margin) 之内的数据, $E(\boldsymbol{\beta},D)$ 是被当前模型正确分类且在间隔之外的数据.

模型训练是一个迭代的过程. 令 $C_1\subset D$ 为初始训练数据缓存, 模型训练与缓存更新交替进行的算法如下:

1) 令 $\boldsymbol{\beta}_t=\boldsymbol{\beta}^*(C_t)$ (用 C_t 训练模型);

2) 如果 $H(\boldsymbol{\beta}_t,D)\subset C_t$, 停止并返回 $\boldsymbol{\beta}_t$;

3) 对任意 $X\in E(\boldsymbol{\beta}_t,C_t)$, 令 $C_t'=C_t\backslash X$;

4) 对任意 $X\in(H(\boldsymbol{\beta}_t,D)\backslash C_t)$, 令 $C_{t+1}=C_t'\cup X$, 回到第 1) 步用 C_{t+1} 继续训练.

第 3) 步是删除当前缓存中的容易数据, 第 4) 步是加入新的未被当前模型很好分类的困难数据. 如果新困难数据已经不存在了, 那么在第 2) 步将退出训练, 得到对应的模型.

令 $\boldsymbol{\beta}^*(D)=\arg\min_{\boldsymbol{\beta}}L_D(\boldsymbol{\beta})$ 是在全部候选点 D 上的全局最优解, 由于 L_D 是严格的凸函数, 所以这个全局解唯一. 我们需要证明经过迭代之后, 在部分数据 $C\subset D$ 上得到的模型 $\boldsymbol{\beta}^*(C)=\boldsymbol{\beta}^*(D)$, 并且算法可以退出迭代过程.

定理 1. 令 $C\subset D$, $\boldsymbol{\beta}=\boldsymbol{\beta}^*(C)$. 如果 $H(\boldsymbol{\beta},D)\subset C$, 那么 $\boldsymbol{\beta}=\boldsymbol{\beta}^*(D)$.

证明. 由上述算法的第 1) 步和第 2) 步得知, 对于 C 上训练得到的模型, 在全局集合 D 上的困难数据也完全存在于 C 中. 这意味着目标函数 L 在集合 $D\backslash C$ 上的错误为 0, 即 $L_C(\boldsymbol{\beta})=L_D(\boldsymbol{\beta})$. 又因为存在唯一最优解, 所以 $\boldsymbol{\beta}^*(C)=\boldsymbol{\beta}=\boldsymbol{\beta}^*(D)$. □

定理 2. 上述算法能够退出.

证明. 算法第 3) 步的集合 C_t' 包含了 C_t 中全部错误不为 0 的数据, 即 $L_{C_t}=L_{C_t'}$. 同样由于存在唯一最优解, 即 $L_{C_t}(\boldsymbol{\beta}^*(C_t))=L_{C_t'}(\boldsymbol{\beta}^*(C_t'))$. 对第 4) 步 $X\in(H(\boldsymbol{\beta}_t,D)\backslash C_t)$ 分类讨论:

1) $X=\varnothing$, 此情况不存在, 否则, 算法在第 2) 步就已经退出.

2) X 中至少包含一个困难数据使得 $\boldsymbol{\beta}_t$ 错误非 0, 故 $C_t'\subseteq C_{t+1}$, 且 $L_{C_{t+1}}(\boldsymbol{\beta})\geq L_{C_t'}(\boldsymbol{\beta})$ 对任意 $\boldsymbol{\beta}$ 都成立. 再分两种情况:

a) 如果 $\boldsymbol{\beta}^*(C_{t+1})=\boldsymbol{\beta}^*(C_t')$, 由于存在新困难数据 X, 则 $L_{C_{t+1}}(\boldsymbol{\beta}^*(C_{t+1}))\geq L_{C_t'}(\boldsymbol{\beta}^*(C_t'))$;

b) 如果 $\boldsymbol{\beta}^*(C_{t+1})\neq\boldsymbol{\beta}^*(C_t')$, 由于 $L_{C_t'}$ 存在唯一最小值, 故 $L_{C_{t+1}}(\boldsymbol{\beta}^*(C_{t+1}))\geq L_{C_t'}(\boldsymbol{\beta}^*(C_t'))$ 仍然

成立.

综上, $L_{C_{t+1}}(\boldsymbol{\beta}^*(C_{t+1})) \geq L_{C'_t}(\boldsymbol{\beta}^*(C'_t))$. 由于数据为有限集, 则算法会在有限迭代次数内退出. □

由于我们模型的训练与检测都只在候选点上进行, 上述两个定理保证了在部分候选点上训练的模型能获得在全体候选点上一致的效果. 而多种不同原理的分割方法往往能捕获到图中的线条边界, 这使得在候选点上的困难数据很接近于图像全部位置的困难数据. 这两点结合起来保证了在部分候选点上迭代训练的模型很接近于在图像全体位置上训练的结果.

4 实验

PASCAL VOC 2007[11] 数据集包括 9 963 张图, 共 20 个目标类, 训练图像与测试图像大致各占一半, 有些图中包含不止一类目标. 检测得到的范围框与人工标记范围框之间的面积交集如果能超过它们并集的 50 %, 则算作一次成功的检测.

每张图采用两种分割工具 EGB (Efficient graph-based segmentation)[13] 和 EDISON (Edge detection and image segmentation on command prompt system)[14] 得到两组分割. EGB 将图像表示成一个无向图, 每个像素作为一个节点, 相邻像素之间的颜色值差异作为边的权重, 从而将图像分割转化为无向图的划分问题. 这种分割方法主要利用了局部区域的颜色特征. EDISON 利用了一种普遍的非参数特征空间分析方法: Mean shift, 并将其应用到了图像分割中去发现任意形状的聚类. 这种分割方法在度量相邻像素相似度时, 同时考虑了空间域 (Spatial domain) 与范围域 (Range domain), 能够利用图像中的边界形状特征.

EGB 有 3 个参数: 平滑因子 σ、相似度阀值 k 以及最小分割尺寸 m. EDISON 有 4 个参数: 空间宽度 sw、范围宽度 rw、最小区域面积 m 以及加速系数 s. 将每个分割的范围方框左上角作为中心, 边长为 8 的正方形之内的区域都看作候选检测点.

实验环境采用文献 [1] 的公开系统 Voc-release 3.1, 图像特征金字塔的构造、滤波器的数量以及训练参数使用系统默认值. 剩下的参数列在表 1 中, 分割通过坐标映射, 得到其在金字塔中每层左上角的位置, 然后, 仍然将边长为 8 之内的区域看成候选检测点. 图像缩得越小, 相当于检测点的范围越大, 这样进一步使得候选点能接近目标的真实位置.

表 1 实验参数

Table 1 The parameter settings in the experiment

步骤	参数
EGB	$\sigma = 1.25$, $k = 700$, $m = 1\,200$
EDISON	$sw = 5$, $rw = 19$, $m = 500$, $s =$ Medium
检测范围边长	8

实验所用的机器配置为: 处理器 2.4 GHz Intel Core 2 Duo, 内存 2 GB 1067 MHz DDR3. 运行时间的比较列在表 2 中.

表 2 每个步骤的运行时间

Table 2 The running time of each step

步骤	处理单张图像的平均时间 (s)
EGB	0.8
EDISON	1.9
候选点检测	0.3
文献 [1]	2.2

在图像完成预分割的基础上, 候选点检测算法处理单张图像的时间仅为 0.3 秒. 如果将分割时间考虑进去, 本文方法平均处理时间为 3.0 秒, 丧失了速度优势. 但是图像分割与后续的检测是相对独立的任务, 一旦完成了分割, 其结果可以反复重用. 无论后续检测方法如何调整, 都无需再一次切割. 当需要迭代训练或反复调整测试时, 基准方法速度不变; 而本文方法由于不需要再次切割, 就会逐渐体现出速度优势.

检测结果列在表 3 中. 参赛者的最佳结果列在第 2 列, 这个结果由当年多个优秀的方法所组成. 第 3 列是文献 [1] 的结果, 这个方法在大部分类别上超过了 VOC2007 的参赛者, 这一列的结果作为比较的基准. 第 4 列 Our_Real 是可变形模型在目标真实范围方框左上角检测的结果. 第 5 列 Our_Seg 是候选点检测方法. 下划线表示除 Our_Real 之外的最佳检测结果. 图 2 为检测结果示例.

Our_Seg 的训练和测试都只在候选检测点上进行. 从直觉上看, 只在某些点上检测的方法最多也就是接近穷举搜索的方法, 但是 Our_Seg 却在一半类别上超过了穷举搜索的基准方法. 主要原因有三个.

1) 训练数据是由正类图像中的目标范围框与迭代挖掘出的困难负类数据所组成. 无论是 Our_Seg, 还是基准方法, 所使用的正类数据是一样的, 负类数据经过多次迭代也逐渐收敛趋于一致. 这是因为所用的两种分割方法原理不同, 图像中的目标边界往往能被至少其中一种方法所捕获. 在整张图像上挖掘困难数据和在这些目标边界附近挖掘, 在多次迭代之后, 效果不会有太大的差别. 由于在这两种不同的设置下, 训练数据总体利用得差不多, 这就使得 Our_Seg 的模型并不比基准模型粗糙.

2) 在测试时, 基准模型往往会产生多个相互层叠重合的窗口. 这并不难想象, 检测窗口如果仅仅向附近移动一点, 其所覆盖区域与刚才变化不大, 此处很可能也会获得较高分数. 尤其是对某些大型物体, 比如 "飞机", 在有些图像上, 甚至会产生超过 100 个相互层叠重合的窗口. 这就需要像 NMS 这样的后处理来压缩窗口的数量和覆盖区域. 而 NMS 一般是基于某个固定的重合标准或是使用更复杂的启发式标准, 并没有任何完全可信的监督, 这就给检测结果带来了不确定性. 而 Our_Seg 只在分割左上角的附近检测, 只取分数最大的一个窗口, 就让模型自己的一次检测覆盖面积决定整个目标. 虽然有时并不足以覆盖整个目标, 但是评判的标准是检测窗口与实际目标之间的交集能超过两者并集的 50 %, 所以只要分割左上角离实际目标很近, 模型往往能够覆盖 50 % 以上, 从而不需要多窗口 NMS.

3) 还有一个容易被忽视但很重要的原因, 就是 VOC 比赛的评价标准 Average precision (AP) 同时兼顾了识别率和精确度. 对于单个类别来说, 令 n 记作此类在测试集中的目标数量, True positive (tp) 代表成功检测, False positive (fp) 代表不成功检测. 识别率 (Recall) 定义为: tp/n, 而精确度 (Precision) 定义为: $tp/(fp+tp)$. 识别率的范围被划分为 10 等份, 从而得到 11 个边界点. 取 11 个精确度, 使得每个精确度相应的识别率等于相应的边界点. 最终取这 11 个精确度的平均值, 即可得到 AP. 具体细节的计算方法参见官方 VOCdevkit Matlab 文件 VOCevaldet.m. 基准方法由于检测了全部位置, 相比之下, 它的识别率大都高于 Our_Seg; 但是测试图像中每个目标类别大约只占 1/20, 在剩下大量不包括目标类别的图像中, 基准方法会产生更多的 False positive, 这使得 Our_Seg 会获得更高的精确度. 这也是使用候选点检测的一大优点.

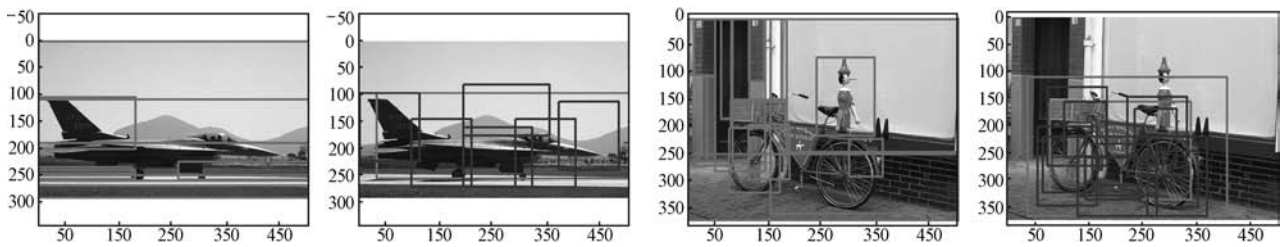
Our_Seg 的检测结果好坏与分割的结果息息相关. 室内图像往往背景环境更为繁杂, 物体更多, 分割就更加困难. 所以, 对于 "Bottle", "Chair", "TV" 等常出现在室内的物体, 此方法不如基准方法好. 室外图像一般来说背景较为简单, 比如 "天空", "草地" 等, 分割较为容易. 所以, 对于 "Bus", "Horse", "Train" 等常出现在室外的物体, 此方法优于基准方法.

总的来说, Our_Real 在大部分类别上都比 Our_Seg 要好. 这并不奇怪, 因为前者毕竟是在人为给定的目标真实范围方框的左上角附近检测. 但值得注意的是, 即使给定了真实位置, Our_Real 仍然在几个类别上不如基准方法. 图 3 生动地解释了这个现象. 这张图来自 PASCAL 2010 的讨论, 范围方框评价标准并不是适用于所用的情况和类别. 就像图中 "人" 这个物体, 用范围方框作为标准其实划入了很多错误的区域, 而真正有意义的检测反而很可能会被认为是错误的. 这同时也说明, 把模型置于左上角, 有时候是完全没有用的. 就像图中, 如果将 "人" 的模型置于左上角, 其所覆盖的区域其实什么都没有, 那自然就检测不出来了. 但是穷举搜索就可以做到. 这就解释了为什么给定了真实范围, 反而不如穷举搜索的原因.

表 3　PASCAL VOC 2007 上的检测结果

Table 3　The detection results on PASCAL VOC 2007

Class	Best	Baseline[1]	Our_Real	Our_Seg
Plane	26.2	29.0	29.3	26.6
Bicycle	40.9	54.6	43.8	42.6
Bird	9.8	0.6	11.8	6.7
Boat	9.4	13.4	18.5	13.3
Bottle	21.4	26.2	24.9	15.6
Bus	39.3	39.4	50.1	45.5
Car	43.2	46.4	48.9	43.8
Cat	24.0	16.1	16.3	15.4
Chair	12.8	16.3	18.0	12.0
Cow	14.0	16.5	22.8	16.0
Table	9.8	24.5	29.5	28.2
Dog	16.2	5.0	15.2	12.9
Horse	33.5	43.6	49.8	45.8
Motbike	37.5	37.8	44.9	40.8
Person	22.1	35.0	33.5	23.5
Plant	12.0	8.8	17.6	16.7
Sheep	17.5	17.3	27.6	22.3
Sofa	14.7	21.6	27.2	23.5
Train	33.4	34.0	42.0	38.0
TV	28.9	39.0	42.8	36.6



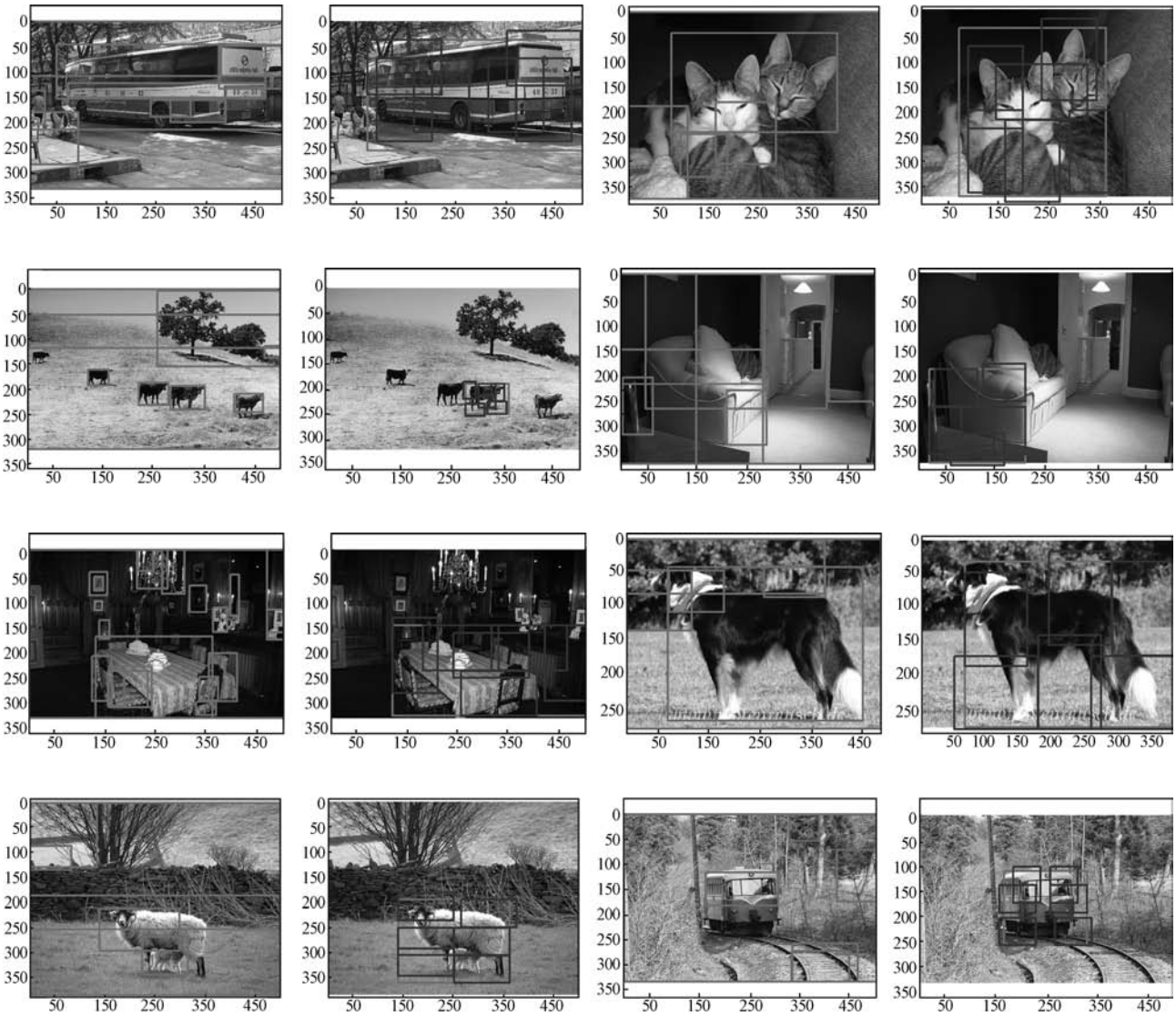


图 2 检测结果示例 (每一对图左边是分割预结果, 右边是候选点检测结果)

Fig. 2 The examples of detection results (The left picture of each pair is the pre-segmentation and the right one is the detected object on the candidate points.)

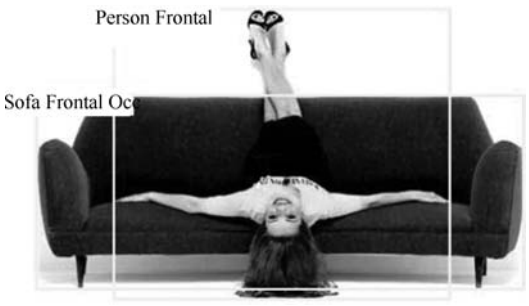


图 3 此例中范围方框标准适合 “沙发” 而不适合 “人”

Fig. 3 The bounding box criterion is suitable for “Sofa” but not for “Person” in this picture

5 相关工作

组成对象的最小单元是像素, 如果能给每个像素加一个类别, 那么属于目标类别的像素就可以决定目标的范围. 基于这个思想, 目标检测可以看作分割级别上的分类问题. 分割本身就可以分成: 单一分割[7]、层次分割[8−9]、多种分割[10] 等不同的方式. 单一分割最为普遍, 就是用单一的分割方法的单一参数, 得到单一粒度级别的分割. 层次分割采用单一分割方法, 但用多组参数得到多个粒度级别的分割, 最粗糙的分割大致为目标大小, 次之可以看作目标的某个部位, 再细致可以是某个细节. 多种分割采用

多个原理可互补的分割方法, 对单张图像得到多包分割. 最流行的方法是采用条件随机场[7−9] 建模分割之间的空间关系, 输入联合类别. 这类方法能得到比范围方框更细致的目标轮廓形状, 也不存在穷举搜索和启发式的后处理. 但是这类方法存在根本的缺点, 分割过程是无监督的, 分割自身往往就存在错误, 这样最终无法合并成有意义的目标.

另一个大类是滑动窗口方法. 这类方法从单窗口检测[2] 扩展到了各种基于部件的复杂结构[1, 3, 12, 15]. 除了为每个类建立一个单独的检测器之外, 还有的方法建模成结构化分类问题, 一次输出图像中多个不同的类别. 文献 [4] 考虑了目标之间的关系, 比如 “摩托车” 上有 “人”, 或者 “瓶子” 放在 “桌子” 上, 旁边有 “椅子” 等. 与之相近的工作还有文献 [5], 采用了结构化的回归去预测单个检测的范围框. 这两种方法模型的优化问题都表示为结构化的支持向量机[16] (Structural SVM). 在国际基准数据集上获得优秀实验结果的大多是滑动窗口方法.

但是检测速度慢是这类方法的最大问题. 假设图像大小为 $n \times n$, 要想从中获得最优窗口, 需要同时考虑窗口的位置和大小, 穷举计算的复杂度是 $\mathrm{O}(n^4)$[6]. 为了提高效率, 常用的启发式简化搜索包括粗糙化搜索区域和固定窗口大小, 或者用局部最优化标准, 在缩小后的可能区域中使用特定的梯度下降[17−18], 也有基于全局最优的 Branch-and-bound 优化策略[6]. 文献 [3] 将可变形模型修改为 Cascade 方式对局部滤波器的窗口位置采用一系列的阀值剪枝, 也大幅提高了检测速度.

与上面的工作相比, 本文的研究有两个最大的特点: 1) 引入了图像分割, 用分割位置提示了目标可能出现的区域, 直接为根滤波器减少了大量的计算, 而不是像文献 [3] 侧重局部滤波器的加速; 2) 缩小检测范围之后, 还有一半类别上结果有提高, 这是上面很多局部最优剪枝方法所不具备的.

6 结论

候选点检测的可变形部件模型打破了基于图像分割的方法与滑动窗口方法的对立. 检测位置由分割决定, 而目标范围由模型决定. 这一结合解决了分割错误带来的固有问题, 也避免了滑动窗口的穷举检测. 目前由于滤波器从左上角计算卷积比较自然方便, 故只将分割左上角附近作为候选检测点. 进一步工作可以扩大候选检测点的范围, 也可以考虑将分割和其他的滑动窗口模型相结合.

References

1 Felzenszwalb P, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

2 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893

3 Felzenszwalb P, Girshick R, McAllester D. Cascade object detection with deformable part models. In: Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2241−2248

4 Desai C, Ramanan D, Fowlkes C. Discriminative models for multi-class object layout. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 229−236

5 Blaschko M, Lampert C. Learning to localize objects with structured output regression. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 2−15

6 Lampert C, Blaschko M, Hofmann T. Beyond sliding windows: object localization by efficient subwindow search. In: Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE, 2010. 1−8

7 Awasthi P, Gagrani A, Ravindran B. Image modeling using tree structured conditional random fields. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: Morgan Kaufmann Publishers, 2007. 2060−2065

8 Reynolds J, Murphy K. Figure-ground segmentation using a hierarchical conditional random field. In: Proceedings of the 4th Canadian Conference on Computer and Robot Vision. Montreal, Canada: IEEE, 2007. 175−182

9 Plath N, Toussaint M, Nakajima S. Multi-class image segmentation using conditional random fields and global classification. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 817−824

10 Russell B, Freeman W, Efros A, Sivic J, Zisserman A. Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of the IEEE Computer

Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2006. 1605−1614

11 Everingham M, Van Gool L, Williams C, Winn J, Zisserman A. The PASCAL visual object classes challenge 2007 (VOC2007) results [Online], available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, January 6, 2012

12 Felzenszwalb P, Huttenlocher D. Distance Transforms of Sampled Functions, Technical Report TR-2004-1963, Department of Computer Science, Cornell University, USA, 2004

13 Comaniciu D, Meer P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(5): 603−619

14 Felzenszwalb P, Huttenlocher D. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005, **61**(1): 55−79

15 Felzenszwalb P, Huttenlocher D. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 2004, **59**(2): 167−181

16 Tsochantaridis I, Hofmann T, Joachims T, Altun Y. Support vector machine learning for interdependent and structured output spaces. In: Proceedings of the 21st International Conference on Machine Learning. Alberta, Canada: ACM, 2004. 1−8

17 Rowley H, Baluja S, Kanade T. Human face detection in visual scenes. In: Proceedings of the Neural Information Processing Systems. Denver, USA: the MIT Press, 1995. 875−881

18 Ferrari V, Fevrier L, Jurie F, Schmid C. Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(1): 36−51

杨 扬 浙江大学计算机学院博士研究生. 主要研究方向为机器学习与机器视觉. 本文通信作者.
E-mail: youthyang@zju.edu.cn
(**YANG Yang** Ph. D. candidate at the College of Computer Science and Technology, Zhejiang University. His research interest covers machine learning and machine vision. Corresponding author of this paper.)

李善平 浙江大学计算机学院教授. 主要研究方向为超大规模信息系统, 金融信息学, Linux 平台及应用.
E-mail: shan@zju.edu.cn
(**LI Shan-Ping** Professor at the College of Computer Science and Technology, Zhejiang University. His research interest covers very large information system, financial informatics, and Linux OS.)