基于优化可形变区域全卷积神经网络的人头检测方法

吉训生，王昊*
江南大学物联网工程学院，江苏 无锡 214122

摘要 人头检测技术是人数统计领域一项重要的研究内容，基于检测的人数统计方法常用于视频监控领域。人头检测常常受到遮挡、背景干扰、光照等因素影响。为解决上述问题，提出一种基于区域全卷积神经网络进行头部检测的方法。特征学习阶段通过残差网络和区域候选网络获得特征及感兴趣区域，并在残差网络中添加可形变卷积层。再将感兴趣区域输入池化层，进行可形变位置敏感均值池化。最后进行分类与目标位置精修，并提出将位置敏感感兴趣区域对齐并进行池化操作。为了改善网络在多尺度头部的检测效果，更新区域候选网络中锚点生成规则。利用在线难例挖掘算法提高复杂任务下头部目标的检测能力，通过软非极大值抑制减少检测边界框间的相互干扰。研究结果表明，在 HollywoodHeads 数据集上平均识别精度最高可达 83.24%，优于目前相关文献的方法。

关键词 图像处理；区域全卷积神经网络；人头检测；可形变卷积
中图分类号 TP391.41 **文献标识码** A **doi**：10.3788/LOP56.141009

Head Detection Method Based on Optimized Deformable Regional Fully Convolutional Neutral Networks

Ji Xunsheng, Wang Hao*
School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract Human head detection is an important research subject for counting people and is often considered to be a useful approach for video monitoring. The challenges associated with human head detection include instance occlusion, background interference, and uneven illumination; this study aims to address these challenges through a method based on the regional fully convolutional neural network. Initially, in the feature learning stage, features are acquired using a residual network (ResNet), and the region of interest is obtained through regional proposal networks. Subsequently, a deformable convolution layer is added into ResNet, and the region of interest is provided as input into the pooling layer for deformable position-sensitive mean pooling. Finally, the target location is classified and refined along with the alignment of the proposed position-sensitive region of interest to complete the pooling operation. Further, the anchor generation rules in regional proposal networks are updated to improve the detection effect of the network based on multi-scale head. The detection ability of head targets under complex tasks is improved using an online hard sample mining algorithm; subsequently, the mutual interference between the bounding boxes is reduced by the soft non-maximum suppression. After applying the proposed method to the HollywoodHeads dataset, the average recognition accuracy is confirmed to become 83.24%, which is better than those of other methods in the current literature.

Key words image processing; regional fully-convolutional neural network; head detection; deformable convolution
OCIS codes 100.3008; 100.4996; 100.4999

1 引 言

人数统计是视频监控系统的关键技术之一，在商场、银行、学校、车站等公共设施内进行人数统计，可以实现人员流量分析与预测，提高安全性与公共资源利用率。通过检测人体目标实现实时、准确的人数统计，是处理该任务的一个主要思路。但人体目标易受光照条件、摄像头位置、背景变化、视角、行

收稿日期：2019-01-07；**修回日期**：2019-01-30；**录用日期**：2019-02-26
基金项目：国家自然科学基金(61771223)、江苏省重点研发计划(SBE2018334)
* **E-mail**：2928412867@qq.com

人相互遮挡等因素的影响，其检测精度往往不高。而人头目标相对于人体目标，受遮挡与视角因素的影响将会大大减小，故人头目标更利于检测。

目前，人头检测的方法大致分为两类。1）基于深度信息确定头部区域，如 Aziz 等[1]通过骨骼图轮廓确定头部，进而标记更新学习；张姗姗等[2]利用双目摄像头获取图像深度信息以及图像强纹理点，确定头部区域。实际应用中，该类方法需要的深度彩色图（RGB-Depth、RGB-D）较难获取。2）利用 RGB 图像学习特征构建检测模型。传统方法多数基于人工设计特征如方向梯度直方图（HOG）、积分通道等作为重点描述特征，利用线性支持向量机（SVM）、自适应提升（AdaBoost）算法等进行分类。近年来，深度卷积网络广泛用于计算机视觉领域，在引入深度学习后，人脸检测效果得到了大幅提高，同时卷积神经网络也已广泛用于目标检测中。其中，文献[3]中提出了一种基于深度图像的人头检测方法，对运动区域进行立体匹配，根据深度分布提取人头，对光线和阴影的抗干扰性良好，能很好地适应环境。文献[4]中采用加速区域卷积神经网络框架实现夜红外图像中的行人检测，无需手动选取目标特征，达到了实际应用中的实时性要求。文献[5]中采用在线高斯模型的行人检测快速候选框生成方法，在静态环境下的精度和实时性都有所提高。Girshick 等[6]提出的区域卷积神经网络（R-CNN）在目标检测中取得巨大成功。文献[7]中使用无监督的学习方法获取行人边缘的中层特征，提升了检测精度，但对硬件的配置要求高。文献[8-9]中提出快速区域卷积神经网络（Faster R-CNN）、基于区域实例分割的卷积神经网络（Mask R-CNN）等方法。这些方法把卷积神经网络在目标检测中的应用推向了一个新的高度，也给目标检测的研究提供了新的思路。传统方法多是采用人工设计特征，卷积神经网络方法在输入图像后直接获得卷积特征，这些相比于人工设计特征有更好的检测效果。人头检测属于目标检测的一类，Vu 等[10]首次基于区域卷积神经网络，对选择性搜索产生的区域内局部上下文信息完成初步学习，通过全局模型预测图像所有像素，最后利用成对模型确立头部的尺度与位置关系。不过，其基于区域卷积神经网络的方法在特征学习阶段耗费的计算资源、硬盘空间较多。

本文提出一种优化的可形变区域全卷积网络，实现人头检测，通过在线难例挖掘[11]（OHEM）算法与软非极大值抑制[12]（S-NMS）算法提升了检测效果。

2 优化区域全卷积网络的人头检测模型

2.1 优化区域全卷积网络

优化区域全卷积网络的人头检测模型结构如图 1 所示。通过基础分类网络的残差网络[13]（ResNet）生成特征映射，再将其输入区域提议网络（RPN），进行前后景目标搜索并确定目标，获取目标的感兴趣区域（ROI）。接着将 RPN 与 ResNet 最后一层输出到 ROI 池化层，对其进行可形变位置敏感 ROI 均值池化，池化操作后降维输出至 softmax 层输出分类概率，其中 ResNet 为了和感兴趣区域池化层连接，剔除了全连接层，并且在 ResNet 的 5a、5b、5c 层添加可形变卷积层[7]，增强模型对目标特征的学习能力。

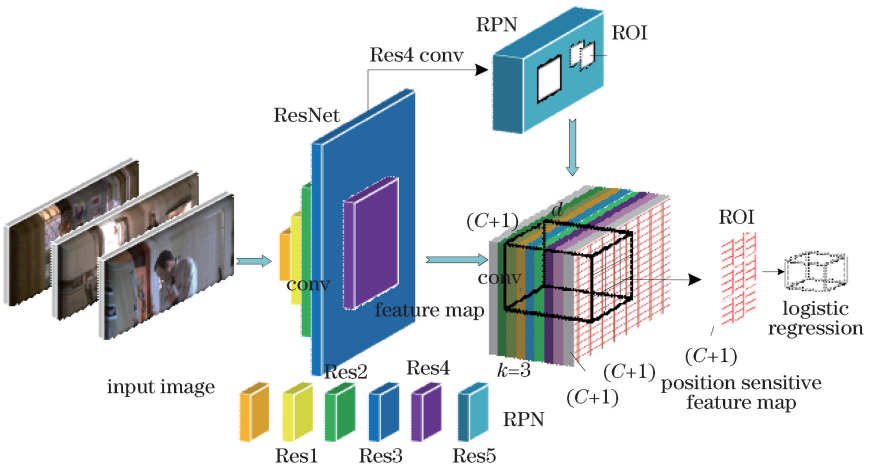

图 1 优化区域全卷积神经网络模型示意图

Fig. 1 Schematic of optimized regional fully convolutional neural network model

图 1 中用 $k^2(C+1)$ 个 1024×1×1 的卷积核去卷积即可得到 $k^2(C+1)$ 个特征图。其中，$k=3$ 表示将一个 ROI 划分成 3×3 对应的 9 个位置；C 表示一共有 C 类目标类别；1 表示背景类别。

2.2 深度残差网络

深度残差网络是由何凯明博士提出的。在其出现之前，当网络加深遇到梯度爆炸与消失的问题时，常通过添加初始归一化与中心归一化层来解决。而当寻求更深的网络时，精度由开始饱和到陡降，训练误差持续增加。网络加深不能简单地复制浅层的体系结构。残差网络在 34 层深度使用 2 个 3×3 卷积层进行卷积操作，在继续加深网络至 50 层后，使用 3 个 1×1、3×3 和 1×1 的卷积层替代 34 层深网络中的 2 个 3×3 卷积层。其中，第 1 个 1×1 用来降维，第 2 个 1×1 用来升维，如图 2 所示，其中 d 为通道的深度。另外，在残差网络中，每个卷积层后都跟上批归一化(BN)层用于归一化，避免梯度爆炸或弥散。这样设计更深的残差网络时复杂度小于 VGG-19 网络的复杂度。

本方法使用的 ResNet 的 5a、5b、5c 中的 3×3 卷积层替代可形变卷积层，并且将 ResNet 的全连接层剔除，以连接后续的池化操作。

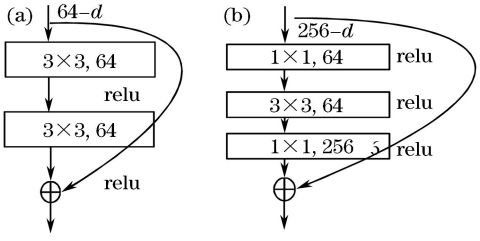

图 2 不同卷积层连接方式。(a)传统 VGG 卷积连接方式；(b) 50 层残差网络卷积连接方式

Fig. 2 Connection modes of different convolution layers. (a) Traditional VGG convolution connection; (b) convolutional connection of 50-layer residual network

2.2.1 可形变卷积层

可形变卷积设计的目的是处理目标可能出现的形变。虽然在数据增强时对数据进行镜像、90°、180°、翻转等选择变化可以完成有限已知的形变，但无法处理未知形变。可形变卷积是在每次卷积的采样点后加一个偏移值，该操作在另外添加的卷积层中进行。可形变卷积示意图如图 3 所示，卷积核大小为 3×3。图 3(b)为添加偏移后，产生的不规则排列；图 3(c)为理想排列并可实现尺度增加；当出现图 3(d)时，形变卷积可实现旋转变换。

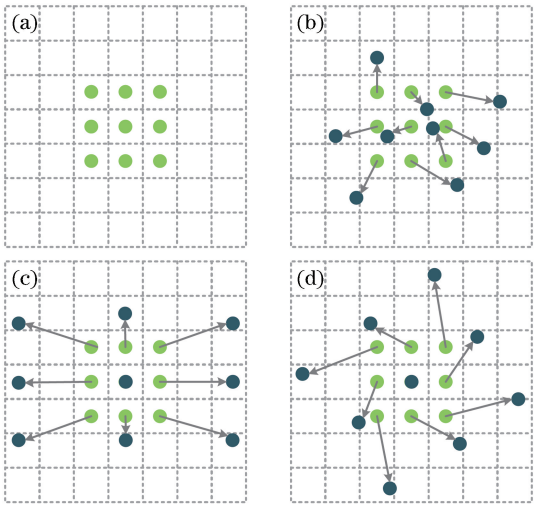

图 3 可形变卷积不同偏移方式示意图。(a)一般卷积；(b)可形变卷积；(c)卷积理想排列；(d)卷积旋转变换

Fig. 3 Schematics of deformable convolution with different migration modes. (a) General convolution; (b) deformable convolution; (c) convolutional ideal arrangement; (d) convolutional rotation transformation

常规卷积使用一个规则的网格 R 在输入的特征图上采样，再对采样获得的值加权求和。如一个 3×3 的核，扩张量为 1。R 定义了感受野大小和扩张量，即

$$R=\{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}。\tag{1}$$

在输出特征图 y 上，令 p_n 代表网格 R 中所有可能的位置，则 $w(p_n)$对 p_n 进行加权求和可表示为

$$y(p_0)=\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n)。\tag{2}$$

在可变形卷积中，用偏移$\{\Delta p_n\mid n=1,\cdots,N\}$对规则的网格 R 进行扩充，其中 N 为单元格中的像素数目。则输出特征图 $y(p_0)$调整为

$$y(p_0)=\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n+\Delta p_n)。\tag{3}$$

采样在不规则有偏移的位置 $p_n+\Delta p_n$ 处进行。偏移量 Δp_n 很有可能不是整数像素，而是一个精度较高的小数。因此，不能直接获取像素坐标。本研究通过双线性插值取整，避免用简单的直接取整而产生误差，也无法采用梯度下降求解。

$$x(p)=\sum_q\max(0,1-|q_x-p_x|)\cdot \max(0,1-|q_y-p_y|)\cdot x(q),\tag{4}$$

式中：$p=p_0+p_n+\Delta p_n$ 为任意一个位置；q 为特征图 x 上所有整数空间位置；max(·)函数将 q 的集

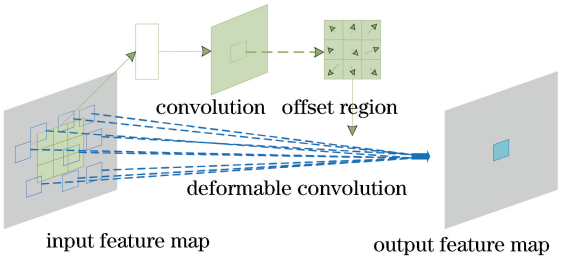

图 4 可形变卷积流程图

Fig. 4 Flow chart of deformable convolution

合范围约束为距离位置 p 最近的 4 个单元格。

图 4 所示为可形变卷积流程，将一般卷积流程分成两条线路，共享输入的特征图。上方偏移用额外的一个核为 3×3 的卷积层学习偏移 Δp_n，得到 $H\times W\times 2N$ 的偏移输出，其中 H 为输入特征图的长度，W 为输入特征图的长度，$2N$ 表示 x 与 y 方向的偏移。添加偏移后的每一个卷积窗口不再是规则的3×3 滑动窗，已变为经过偏移后的窗口。输入特征图和偏移一同作为可形变卷积层的输入，可形变卷积层中采样点发生偏移后再进行卷积。计算过程与卷积一致。

2.2.2 ROI 获取和 RPN 网络细节

基础分类网络选取 ResNet 对输入图像进行卷积。ResNet 训练以迁移学习的方式，使用在 ImageNet 上迭代数十万次的预训练模型。为提升网络速度，将 ResNet 的 Res4f 与 RPN 连接，而不是将 Res5c 输出直接输入至 RPN，此连接方式主要是为了后续边界框进行类不可知方式回归，其在边界框回归输出较少，网络更快，占用内存更少。Res4f 输出为 1×1024×63×38 的张量。RPN 网络结构如图 5 所示。RPN 通过滑动窗对 Res4f 输出的特征映射进行候选区域搜索，其滑动窗步长为 1，单次搜索 512，紧接着用 3×3 的卷积核完成卷积操作。将其获得的张量分别用于前景背景分类部分与候选位置预测部分。前景背景分类部分使用 1×1×512×2×9 卷积完成张量到下层网络的映射，同理候选位置预测部分使用 1×1×512×4×9 卷积。前者中的 2 个参数为前景(头部)与背景概率，后者中的 4 个参数为候选位置的中心坐标及宽和高。

为适应头部目标检测，优化了 RPN 中滑动窗搜索方式。考虑到研究目标为头部目标，头部目标的边界框一般长与宽相近，并且场景中常出现不同尺度的目标，因此设定 4 种尺度为 162、322、642、1282，长宽比为 1∶2、1∶1、3∶2、2∶1 的 16 个锚点

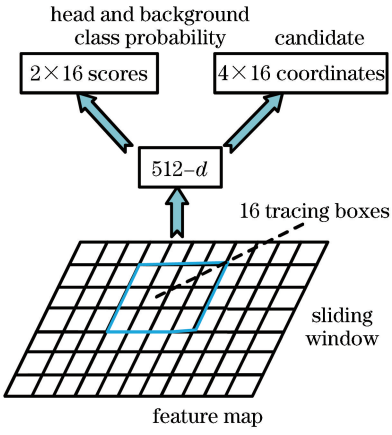

图 5 模型中 RPN 网络结构示意图

Fig. 5 Schematic of RPN network structure in model

框，在默认区域提议网络的锚点框的设置上增加 162 的尺度和长宽比 3∶2，共增加 7 个锚点框。其中，锚点框与真实值边界框的交叉联合(IOU)重叠率 $\eta>0.7$ 时，标记为正样本(头部)；若 $0.3\leqslant\eta\leqslant 0.7$ 时，取最大的 IOU 的锚点框为正样本；若 $\eta<0.3$，则标为负样本(背景)。实验中会出现部分超出图片边界的锚点框，该部分不参与后续操作，而当出现多个锚点框重叠于同一目标时，通过非极大值抑制选择锚点框，减少不必要的计算量。

2.2.3 位置敏感感兴趣区域池化

通常，网络构造越深，其平移旋转不变性越强，这个性质对于保证分类模型的稳健性有着积极的意义。但是在检测问题中，设计的模型需要对目标的位置信息有一定的感知能力，过度的平移旋转不变性会削弱这一性能。较深的全卷积神经网络如 Inception、ResNet 以及 Faster-RCNN 等检测模型都存在着一个明显的缺陷，即目标检测器对目标的位置信息的感知敏感度下降，检测准确度降低。之前，最直观的解决方法是将 RPN 与 ResNet 的连接位置向浅层移动，正如本文方法将 RPN 嵌入到 ResNet 的 Res4 位置，不过这样做会明显增加计算量，使得检测速度变慢。

位置敏感感兴趣区域池化层的主要思想是在特征聚集时人工引入位置信息，从而在一定程度上改善较深的神经网络对目标位置信息的敏感程度，并且基于区域全卷积网络(R-FCN)直接对整张输入图像进行大部分操作，很大程度上优化了网络的运行速度。位置敏感感兴趣区域池化层具体操作方法如图 1 右侧池化操作部分所示。

对位置敏感感兴趣区域池化操作的具体实现是将 ResNet 输出经过卷积后与 RPN 输出一同输

入至 ROI 池化。将 ROI 分为 $k\times k$ 个位置敏感区域，每个位置敏感区域输出 $C+1$ 个通道，其中 C 在本模型中为头部类别，1 为背景类别。对每个 ROI 进行位置敏感池化，得到 $k^2(C+1)$ 个通道位置敏感的得分图，本模型 $k=3$，共得 18 个通道位置敏感得分图。一个大小为 $w\times h$ 的 ROI 分为 9 个大小为 $\frac{w}{3}\times\frac{h}{3}$ 的子区域。用 (i,j) 表示子区域的位置，$i\in[0,2]$，$j\in[0,2]$。位置敏感池化公式可表示为

$$r_c(i,j\mid\theta)=\frac{1}{n}\sum_{(x,y)\in\mathrm{bin}(i,j)}z_{i,j,c}(x+x_0,y+y_0\mid\theta),\quad(5)$$

式中：$r_c(i,j)$ 为子区域 (i,j) 对头部类的池化映射；$z_{i,j,c}$ 为该子区域对应的位置敏感得分图；(x_0,y_0) 为 ROI 的左上角坐标；(x,y) 为 ROI 每个元素的坐标；n 为该子区域块的像素值；θ 为通过网络学习的参数。

2.2.4 位置敏感感兴趣区域对齐

感兴趣区域对齐[8]是一种区域特征聚集方式，最初用于语义分割。初始设计是为了解决感兴趣区域池化操作中两次量化会产生的区域不匹配问题。

主流的两步目标检测框架（如 Fast-RCNN[9]、Faster-RCNN[10]）中，感兴趣区域池化操作通常在特征图中根据预选框的位置坐标将对应的区域映射为统一不变的尺寸特征图，以便进行后续的分类和边界框回归操作。候选边界框的位置多由网络回归获得，通常为浮点数，但要求固定池化操作后的特征图尺寸。因此，ROI 池化这一操作存在两次量化的过程。将量化后的边界区域均匀分割成 $k\times k$ 个单元，对每个单元的边界进行量化，并将候选边界框量化为整数点坐标值。

不匹配问题是指在实际操作过程中，两次量化操作后，候选边界框和刚开始回归得到的位置会出现一定的偏差，这会影响最终分割准确度或检测精度。图 6 给出 Faster-RCNN 检测框架。输入一张 750 pixel × 400 pixel 的图片，图片上有一个 280 pixel×28 pixel 的包围边界框。图片经过主干网络提取特征后，图缩放操作的步长为 32。因此，缩放后的图像和包围边界框的长宽都是输入时的 1/32。750 除以 32 变为 23.4375，400 除以 32 以后得到 12.5，280 除以 32 得 8.75，都带有小数，于是 ROI 池化直接将其分别量化成 23、12 和 9。接下来，需要把边界框内目标的特征池化为 7×7 的大小，因此将包围边界框平均分割成 7×7 个矩形区域。每个矩形区域的边长为 1.25，又含有小数，于是 ROI 池化再次将其量化到 1。经过上述两次量化，候选区域出现了较明显的偏差（如图 6 中绿色部分所示）。更重要的是，该层特征图上 0.1 pixel 的偏差，缩放到原图就是 3.2 pixel，即 0.25 的偏差，在原图上接近 8 pixel 的差别。若输入图片较大且除以 32 和 7 后的余数较大，这些偏差对候选区域特征学习的影响会很大。

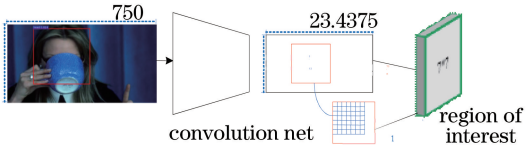

图 6 感兴趣区域池化操作

Fig. 6 Pooling operation of ROI

为了解决 ROI 池化映射时存在的不足，ROI Align 取消了映射时的量化操作，使用双线性内插方法获得坐标为浮点数的像素点上的图像数值，从而将整个特征聚集过程转化为一个连续的操作。ROI 对齐并不是简单地补充候选区域边界上的坐标点后对区域边界坐标点进行池化，其步骤如下：1）ROI对齐遍历每一个候选区域，保持浮点数边界不作量化；2）将候选区域分割成 $k\times k$ 个单元，每个单元的边界也不作量化；3）在每个单元中计算固定的 4 个坐标位置，用双线性内插的方法计算出这 4 个位置的值，然后进行最大值池化操作。其中步骤 3）中固定的 4 个坐标位置是指在每一个矩形单元中按照固定规则确定的位置，即若采样点数目为 1，采样点就是这个单元的中心点，若采样点数是 4，采样点就是把这个单元平均分割成 4 个小方块之后每个方块各自的中心点。这些采样点的坐标通常是浮点数，需要使用双线性插值法得到其像素值。另外，采样点数设定为 4 时会获得最佳性能，而若直接设为 1 则对性能无太大影响。ROI 对齐遍历采样点的数目并没有 ROI 池化那么多，却可以获得更好的性能，这主要是因为解决了量化过程中的不匹配问题。不过不同大小的目标受不匹配问题的影响程度不同。同样是 1 pixel 的偏差，对于较大的目标来说，影响微不足道，但对于小目标，误差的影响会高很多。因为头部目标会出现部分小尺度，故可引入 ROI 对齐。

ROI 池化的反向传播公式表示为

$$\frac{\partial L}{\partial x_i}=\sum_r\sum_j[i=i\times(r,j)]\frac{\partial L}{\partial y_{rj}},\quad(6)$$

式中：L 为池化前的特征图；x_i 为池化前特征图上的像素点；y_{rj} 为池化后第 r 个候选区域的第 j 个点；$i\times(r,j)$ 为点 y_{rj} 最大值池化操作时选出的最大像素值所在点的坐标。

由(6)式可以看出，只有当池化后某个点的像素值在池化过程中采用了当前点 x_i 的像素值，即满足 $i=i\times(r,j)$ 时，才在 x_i 处回传梯度。

类比于 ROI 池化和 ROI 对齐的反向传播需要作出修改。在 ROI 对齐中，$x_i\times(r,j)$ 是一个浮点数的坐标位置(前向传播时计算出来的采样点)，在池化前的特征图中，每一个与 $x_i\times(r,j)$ 横纵坐标距离均小于 1 的点都应该接受与此对应的点 y_{rj} 回传的梯度，故 ROI 对齐的反向传播公式为

$$\frac{\partial L}{\partial x_i}=\sum_r\sum_j[d(i,i\times(r,j))<1]\cdot(1-\Delta h)(1-\Delta w)\frac{\partial L}{\partial y_{rj}},\tag{7}$$

式中：$d(\cdot)$ 表示两点之间的距离；Δh 和 Δw 分别为 x_i、$x_i\times(r,j)$ 横纵坐标的差值，这里作为双线性内插的系数。

将 ROI 对齐移植到位置敏感 ROI 池化中，实现位置敏感感兴趣区域对齐。主要改进就是取消两次量化：ROI 的边界坐标值和每个 ROI 中所有矩形单元的边界值保持浮点数形式，在每个矩形单元中计算出固定位置、固定数量的采样点的像素值作平均池化。前向传播的具体步骤如下：

1) 遍历池化后特征图上的每一个像素点，在池化前特征图上寻找对应通道上的对应区域；

2) 将每个候选区域平均划分成 $n\times n$ 个单元；

3) 在每个单元内，按照设置的采样点数目计算出采样点的坐标值；

4) 使用双线性内插的方法计算出特征图上每个采样点处所对应的值；

5) 依照平均池化的方式计算出步骤 1)中当前点的值，并记录所有采样点的位置坐标。

反向传播步骤如下：

1)遍历池化后特征图上的每一个像素点，在池化前特征图上寻找对应通道上的对应区域；

2)在步骤 1)的当前区域中遍历每一个点，分别与前向传播中记录下来的所有采样点坐标比较，如果横纵坐标都小于 1，则回传平均后的梯度值。

在本方法中，直接使用位置敏感感兴趣区域对齐替代位置敏感感兴趣区域池化。

2.2.5 可形变位置敏感感兴趣区域池化

感兴趣区域池化是将 ROI 分成 $k\times k$ 个单元，对每个单元中的多个像素作均值池化，最后输出 $k\times k$ 个特征图，可表示为

$$y(i,j)=\sum_{p\in\mathrm{bin}(i,j)}x(p_0+p_n)/n_{ij},\tag{8}$$

式中：p 为每个单元中的任一像素点；$\mathrm{bin}(i,j)$ 为每个单元中像素点的横纵坐标集合；n_{ij} 为一个单元中像素个数。可形变感兴趣区域池化与可形变卷积思路一致，在空间位置单元增加偏移，将(8)式变为

$$y(i,j)=\sum_{p\in\mathrm{bin}(i,j)}x(p_0+p_n+\Delta p_{ij})/n_{ij},\tag{9}$$

式中：Δp_{ij} 为偏移量，$0\leqslant i,j\leqslant k$。可形变位置敏感感兴趣区域池化主要是将可形变感兴趣区域池化中整体的特征图用位置敏感分值图替换，即 x_{ij} 替代 x，

$$y(i,j)=\sum_{p\in\mathrm{bin}(i,j)}x_{ij}(p_0+p_n+\Delta p_{ij})/n_{ij}。\tag{10}$$

可形变位置敏感 ROI 池化与其他可形变操作的偏移量学习的方式不同。如图 7 所示，k 表示把一个 ROI 划分成 $k\times k$ 对应的 k^2 个位置，C 表示一共有 C 类目标类别，1 表示的是背景类别。在上方的分支中，输入特征图经过一个核为 1×1 的卷积层后产生完整的空间像素偏移域。对于每个 ROI 和每一类(背景类和头部目标类)，位置敏感 ROI 池化会被应用在这些空间像素偏移域上获取归一化的偏移量，然后通过上述可形变 ROI 池化相似的方法变成真正的偏移量。在本文方法的具体实现如图 8 所示。

2.2.6 分类检测阶段

如图 8 所示，可形变位置敏感池化中对抽取部分进行均值操作得到一个 3×3 的矩阵，再对该矩阵求和得到 1×2 的向量，最后对其进行多项逻辑斯特回归(Softmax)分类。

在 Softmax 与边界框回归阶段，增加在线难例挖掘，以加速收敛并提高分类与检测精度。在线难例挖掘首先读入所有 ROI 的损失值，然后根据损失值排序选出最高损失的 ROI，最后将其放入反向传播中进行梯度回传。为减少损失值重复计算，损失值排序使用 NMS，批尺寸选为 128，因为单张图片产生的 ROI 已较多。其思路类似于自助法。

本模型的损失函数定义为交叉熵损失与边界框回归损失之和，表示为

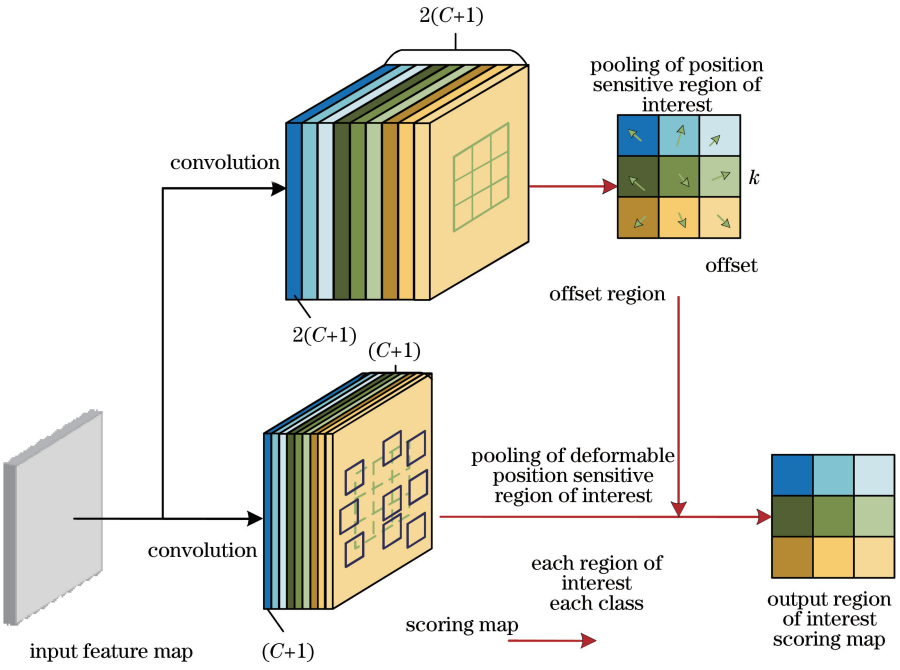

图 7 可形变位置敏感感兴趣区域池化

Fig. 7 Pooling of deformable position-sensitive ROI

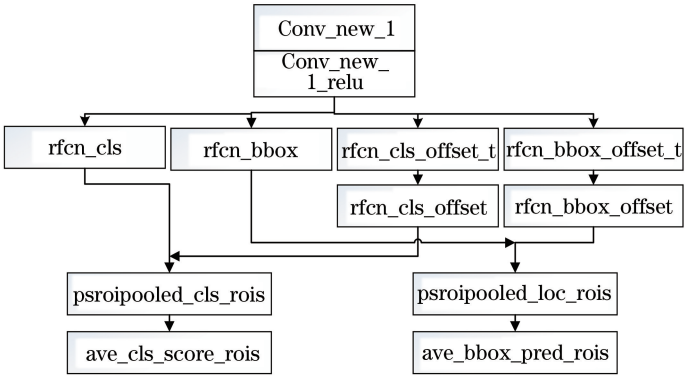

图 8 本方法中可形变位置敏感 ROI 池化

Fig. 8 Pooling of deformable position-sensitive ROI in the proposed method

$$\begin{cases} L(s,t)=L_{\mathrm{cls}}(s_{c*})+\lambda[c*>0]L_{\mathrm{reg}}(t,t*) \\ L_{\mathrm{cls}}(s_{c*})=-\ln s_{c*} \\ L_{\mathrm{reg}}(t,t*)=R(t-t*) \end{cases}, \tag{11}$$

式中：s 为网络得到的类别参数；s_{c*} 为类别预测值与真实类别的差值；$c*$ 为 ROI 的真实类；t 为网络得到的回归框参数；$t*$ 为真实回归框参数；L_{cls} 为交叉熵损失；L_{reg} 为边界框回归损失。

测试时使用 S-NMS，常规 NMS 是直接抑制多个得分较高的边界框，即将高得分附近的边界框得分直接设置为 0，这样可以抑制重复边界框导致的冗余计算，但也会抑制目标重叠导致的部分高得分边界框，从而使模型产生漏检，表达式为

$$s_i=\begin{cases} s_i, \mathrm{IOU}(M,b_i)<N_{\mathrm{t}} \\ 0, \mathrm{IOU}(M,b_i)\geqslant N_{\mathrm{t}} \end{cases}, \tag{12}$$

式中：M 为得分最高的边界框；b_i 是其他边界框；N_{t} 是设定的 IOU 阈值；IOU(·)为交叉联合函数。

S-NMS 使 M 附近的边界框随着其 IOU 阈值升高而得分降低。这样在减小重复边界框影响的同时，也不会遗漏部分利于检测的边界框，表达式为

$$s_i=\begin{cases} s_i, \mathrm{IOU}(M,b_i)<N_{\mathrm{t}} \\ s_i\,[1-\mathrm{IOU}(M,b_i)], \mathrm{IOU}(M,b_i)\geqslant N_{\mathrm{t}} \end{cases}。 \tag{13}$$

在网络训练过程中，为了减小每次梯度计算过

程中噪声的影响，考虑到网络中有较多的连续，且梯度较小，使用动量为0.9、学习率为0.01、权重衰减为0.005的带动量的随机梯度下降算法(SGD)最小化独立对数损失的和值，以优化网络参数。

3 实验与结果分析

使用的数据集是文献[10]中制作的HollywoodHeads头部检测数据集。其来自于21部好莱坞电影中224、740 frame的369、846个头部注释。因该数据集不是标准VOC格式，对其进行格式转换，在转换过程中发现部分目标注释出现超出图像边界和图片内无目标注释的问题。训练集制作时将剔除这类图片，训练集最终使用HollywoodHeads数据集中15部电影216719 frame中的216694张图像，验证集为3部电影6719 frame中的6676张图片。测试集不变，仍为另外3部电影的1302 frame图像。为了评估检测性能，使用平均精度均值(mAP)度量，测试速度使用每秒传输帧数度量。认为与真实值有高重叠率(N_t大于0.5)的检测是正确的。

本文方法对比了基于R-CNN[12]的目标检测器、基于可形变零件模型(DPM)[13]的面检测器(DPM Face)、文献[10]中提出的结合上下文信息的Local-RCNN、基于Zeiler-Fergus网络(ZF)与VGG16的Faster-RCNN。R-CNN使用HollywoodHeads数据集的训练子集训练人头上的R-CNN目标检测器。受内存限制，R-CNN训练的SVM阶段是在一组训练图像上完成。对基于DPM的面部检测器，使用香草DPM模型。Faster-RCNN的预训练模型使用ImageNet的VGG16和ZF的Caffe模型。在Faster-RCNN上同时对比了增加锚点框与软非极大值抑制操作后对检测结果的影响，R-FCN使用本文优化后的模型，同样采取了迁移学习的方法，使用Caffe Zoo提供的ResNet-50与ResNet-101。主要对比本文方法中增强多尺度目标学习的锚点框(Anchor)、降低边界框相互间影响的S-NMS、增强目标表征的可形变卷积、可形变位置敏感感兴趣区域池化以及位置敏感感兴趣区域对齐在HollywoodHeads数据集的表现，不同模型的mAP与测试速度如表1所示。

表1 不同模型在HollywoodHeads上的mAP及测试速度

Table 1 mAP and test speed of different models on HollywoodHeads

Method	Anchor	S-NMS	Deformable conv	Iterations	Test speed /(frame·s^{-1})	mAP /%
DPM Face	—	—	—	—	—	37.4
R-CNN						67.1
Local-RCNN	—	—	—	—	—	72.7
Faster RCNN(ZF)	—	—	—	50000	17.24	73.48
Faster RCNN(ZF)	√	—	—	50000	17.24	74.50
Faster RCNN(VGG16)	—	—	—	50000	6.36	79.17
Faster RCNN(VGG16)	√	—	—	50000	6.36	80.02
R-FCN(ResNet-50)	—	—	—	30000	7.29	81.00
R-FCN(ResNet-50)	—	√	—	30000	7.29	81.12
R-FCN(ResNet-50)	√	—	—	30000	7.29	81.96
R-FCN(ResNet-50)	√	—	—	40000	7.29	82.00
R-FCN(ResNet-50)	√	√	—	40000	7.29	82.41
R-FCN(ResNet-50)	√	—	√	30000	7.29	82.76
R-FCN(ResNet-50)	√	—	√+	30000	6.95	83.24
R-FCN(ResNet-50)*	√	—	√	30000	7.04	82.83
R-FCN(ResNet-101)	—	—	—	30000	8.50	84.49
R-FCN(ResNet-152)	—	—	—	30000	8.32	84.32

Note: "*": position sensitive ROI align; "+": deformable position sensitive ROI; "√": corresponding network added.

使用部分HollywoodHeads数据集的测试结果如图9～16所示。

HollywoodHeads数据集训练的可形变区域全卷积网络模型在部分网络资源图片上的测试表现如图17和图18所示。

Faster-RCNN与R-FCN均使用了在线难例挖掘算法。其中，S-NMS只在测试时使用，检测效果有一定的提升，实验显示训练时使用S-NMS反而会使检测效果下降。分析其原因在于：训练数据中的目标并不都存在遮挡问题，使用S-NMS会降低

图 9 暗光下小尺度测试

Fig. 9 Small scale measurement in dark light

图 10 暗光与背景干扰下测试

Fig. 10 Test under dark-light and background interference

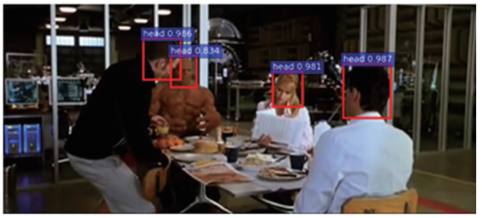

图 11 多对象边界框重叠测试

Fig. 11 Multi-object boundary box overlapping test

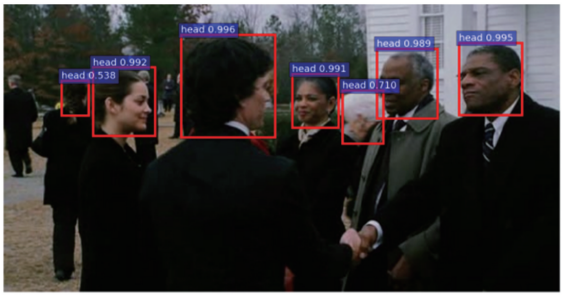

图 12 多对象多尺度测试

Fig. 12 Multi-object and multi-scale test

图 13 小尺度遮挡测试

Fig. 13 Small-scale occlusion test

图 14 多对象有遮挡测试

Fig. 14 Multi-object occlusion test

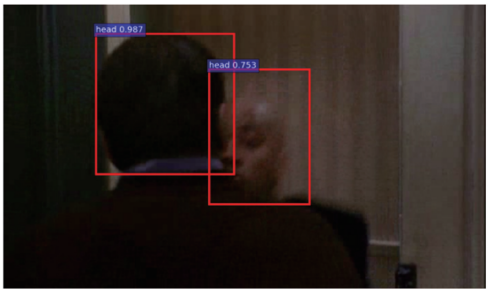

图 15 双对象遮挡测试

Fig. 15 Two-object occlusion test

图 16 小尺度模糊对象测试

Fig. 16 Small-scale fuzzy object test

图 17 多对象正面头部测试

Fig. 17 Multi-object frontal head test

目标周围部分区域信息的学习效果。R-FCN 因使用了迁移预训练模型，迭代 30000～40000 次已达到收敛状态。

从实验结果可以发现，基础网络加深确实可以提高对目标特征的学习效果，检测精度明显提高，但并不是越深越好。基于 ResNet-152 的 R-FCN 在任务下表现低于 ResNet-101。增强锚点框对多尺度

图 18 多对象有遮挡侧面头部测试

Fig. 18 Multi-object occlusion side head test

目标的学习效果更好，故使 Faster-RCNN 与 R-FCN 上的检测精度提升明显，但在 R-FCN 中迭代约 40000 次才达到收敛，另外 16 个锚点框所需的训练时间比 9 个锚点框时长。测试时使用 S-NMS 会使检测精度提高，16 个锚点框 & S-NMS 的 R-FCN 相比不做修改的 R-FCN，检测精度可提高 2.41%。添加可形变卷积与 16 个锚点框的 R-FCN 相比，16 个锚点框 & S-NMS 的 R-FCN 检测精度可提高 0.35%。

在 16 个锚点框和可形变卷积条件下，提出的位置敏感感兴趣区域对齐与可形变位置敏感感兴趣区域池化相比，原始位置敏感感兴趣区域池化均有所提高，分别为 0.07%和 0.48%。在检测速度上两者都有轻微下降，优化后基于 ResNet-50 的可形变区域全卷积神经网络在头部检测任务上检测精度可达到 83.24%，比文献[10]中的方法提高了 10.54%，而修改后的锚点框对多尺度目标有更好的学习效果，很好地解决了文献[10]中方法在小目标检测上出现的漏检率偏高的问题。提出的可形变区域全卷积神经网络的人头检测模型为 128.9 MB，结合位置敏感感兴趣区域对齐的区域全卷积神经网络的人头检测模型为 127.7 MB。

基础分类网络为 ResNet-101 优化后的 R-FCN 检测精度为 84.49%。因为 ResNet-101 对显存需求较高，使用 NVIDIA GPU 1070(8G)训练测试，而 ResNet-50 使用 NVIDIA GPU 1060(6G)进行训练测试，在帧率表现上两者不作相关对比。

本文 R-FCN 与 Faster-RCNN 实验基于 Ubuntu 的 Caffe[14] 框架进行，其他对比实验在 Matconvnet[15]框架上进行。其他配置为内存 8G，CPU Intel i5。

4 结 论

主要对区域全卷积网络进行了进一步优化，并将其应用于人头检测过程中。在残差网络中加入可形变卷积层，区域提议网络中针对头部目标的特点优化特征提取锚点选择，感兴趣区域池化阶段使用可形变位置敏感感兴趣区域池化，并提出位置敏感感兴趣区域对齐。池化操作后加入在线难例挖掘算法增强较难目标的学习，测试时添加软非极大值抑制降低边界框间的相互干扰。研究结果表明，优化后的可形变区域全卷积网络在实现端到端检测的同时，较大程度地提高了人员头部检测效果。而在大型 HollywoodHeads 数据集上的仿真结果表明，优化后 ResNet-50 的可形变区域全卷积神经网络检测精度可达到 83.24%。

参 考 文 献

[1] Aziz K, Merad D, Iguernaissi R, *et al*. Head detection based on skeleton graph method for counting people in crowded environments[J]. Journal of Electronic Imaging, 2016, 25(1): 013012.

[2] Zhang S S, Jing W B, Liu X, *et al*. A head detection method based on depth information[J]. Journal of Changchun University of Science and Technology (Natural Science Edition), 2016, 39(2): 107-111, 115.
张姗姗, 景文博, 刘学, 等. 一种基于深度信息的人头检测方法[J]. 长春理工大学学报(自然科学版), 2016, 39(2): 107-111, 115.

[3] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[4] Ye G L, Sun S Y, Gao K J, *et al*. Nighttime pedestrian detection based on faster region convolution neural network [J]. Laser & Optoelectronics Progress, 2017, 54(8): 081003.
叶国林, 孙韶媛, 高凯珺, 等. 基于加速区域卷积神经网络的夜间行人检测研究[J]. 激光与光电子学进展, 2017, 54(8): 081003.

[5] Qin J, Wang M H. Fast pedestrian proposal generation algorithm using online Gaussian model [J]. Acta Optica Sinica, 2016, 36(11): 1115001.
覃剑, 王美华. 采用在线高斯模型的行人检测候选框快速生成方法[J]. 光学学报, 2016, 36(11): 1115001.

[6] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C] // 2014 IEEE Conference

on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 580-587.

[7] Lim J J, Zitnick C L, Dollar P. Sketch tokens: a learned mid-level representation for contour and object detection [C] // 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 3158-3165.

[8] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[9] He K M, Gkioxari G, Dollar P, *et al*. Mask R-CNN [J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence [2018-12-25]. https://ieeexplore.ieee.org/document/8372616.

[10] Vu T H, Osokin A, Laptev I. Context-aware CNNs for person head detection [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 2893-2901.

[11] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 761-769.

[12] Bodla N, Singh B, Chellappa R, *et al*. Soft-NMS: improving object detection with one line of code[C] // 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice. New York: IEEE, 2017: 5562-5570.

[13] Liu H, Peng L, Wen J W. Multi-scale aware pedestrian detection algorithm based on improved full convolutional network[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091504.
刘辉, 彭力, 闻继伟. 基于改进全卷积网络的多尺度感知行人检测算法[J]. 激光与光电子学进展, 2018, 55(9): 091504.

[14] Yang M, Zhang B, Song Y L. Application of support vector machine based on optimized kernel function in people detection [J]. Laser & Optoelectronics Progress, 2018, 55(4): 041001.
杨萌, 张葆, 宋玉龙. 基于优化核函数支持向量机在行人检测中的应用[J]. 激光与光电子学进展, 2018, 55(4): 041001.

[15] Vedaldi A, Lenc K. MatConvNet: convolutional neural networks for MATLAB[C] // Proceedings of the 23rd ACM International Conference on Multimedia, October 26-30, 2015, Brisbane, Australia. New York: ACM, 2015: 689-692.