doi:10.6041/j.issn.1000-1298.2019.S0.051

# 基于竞争融合区域建议网络的在线行人跟踪算法研究

王兵兵[1,2]　王　莹[1,2]　陈治昌[1,2]　杨邦杰[1,2]　高万林[1,2]　王敏娟[1,2]

(1. 中国农业大学信息与电气工程学院, 北京 100083;
2. 中国农业大学农业农村部农业信息化标准化重点实验室, 北京 100083)

**摘要**: 针对复杂背景下行人跟踪任务的深度学习网络模型和在线行人跟踪算法问题,在多层竞争融合模型目标检测预训练区域建议网络基础上,结合长期和短期并存的在线学习更新策略,实现行人跟踪任务。预训练网络的特征提取过程以 VGG16 为主干网络,将提取的特征投入多层竞争融合区域建议网络中,进而生成定位更准确的候选目标。在线跟踪算法使用预训练过的区域建议网络初始化参数值,选取第 1 帧 500 个正样本和 5 000 个负样本对区域建议网络进行微调,建立长期和短期更新的帧索引集,通过正负样本对区域建议网络进行更新,最终实现在线行人跟踪算法。在公开数据集 Caltech、ETH、PETS 2009 和 Venice 上对本文模型进行实验验证,结果表明,竞争融合区域建议网络在行人跟踪任务中性能优越,在几个环境背景较复杂的行人数据集中改进的方法均取得了很好的效果。

**关键词**: 行人跟踪; 区域建议网络; 在线学习算法; 竞争融合模型

**中图分类号**: TP39　　**文献标识码**: A　　**文章编号**: 1000-1298(2019)S0-0331-08

# Competitive Fusion Region Proposal Network Based Online Pedestrian Tracking Algorithm

WANG Bingbing[1,2]　WANG Ying[1,2]　CHEN Zhichang[1,2]　YANG Bangjie[1,2]　GAO Wanlin[1,2]　WANG Minjuan[1,2]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *Key Laboratory of Agricultural Information Standardization*, *Ministry of Agriculture and Rural Affairs*, *China Agricultural University*, *Beijing* 100083, *China*)

**Abstract**: Target tracking is an important part of computer vision, specially the pedestrian detection and tracking is a crucial and difficult field. Many researchers have been devoted to the improvement of target detection and tracking methods. With the wide application of deep convolution network, the result of pedestrian detection and tracking has been improved. However, some complex scenarios are difficult to identify and track by present methods. Therefore, it's necessary to propose an optimal algorithm to improve the performance of pedestrian detection and tracking. The region proposal network, which included multi-layer competitive fusion model was used as pre-training network, and long-term and short-term update strategy in pedestrian tracking task. The pre-training network applied VGG16 to extract feature maps, and then they were put into the multi-layer competitive fusion region proposal network to generate more accurate candidate targets. The online pedestrian tracking algorithm was initialized by the pre-training region proposal network, and the region proposal network was fine-tuned through 500 positive samples and 5 000 false positive examples from the first frame, and then created frame index datasets for long-term and short-term update. Finally, the pedestrian tracking algorithm with continuous updating of region proposal network was accomplished. The model was verified by experiment and in the public datasets named by Caltech, ETH, PETS 2009 and Venice. The test result showed that the region proposal network which included multi-layer competitive fusion model had the perfect performance in pedestrian detection and tracking task, and showed good effects in complex background environment.

**Key words**: pedestrian tracking; region proposal network; online learning algorithm; competitive fusion model

收稿日期: 2019-04-23　修回日期: 2019-05-26

**基金项目**: 教育部科学事业费重大项目(2017PT19)

**作者简介**: 王兵兵(1988—),女,博士生,主要从事计算机视觉和植物表型组学研究,E-mail: snowice310@163.com

**通信作者**: 高万林(1965—),男,教授,博士生导师,主要从事智慧农业、网络信息安全和农业信息化与标准化研究,E-mail: gaowlin@cau.edu.cn

## 0 引言

目标跟踪作为计算机视觉领域的一种常见任务[1],在机器人导航、视频监控、自动驾驶等任务中有着重要应用。传统的视频处理方法[2]都是将视频信息转换为图像信息进行处理,由于图像算法在处理目标随时间产生的形状变化时效果并不理想,使得视频信息的处理费时、费力,因此,需要使用目标跟踪技术处理视频信息。

目标跟踪是一种不断估计目标对象状态的过程,目标状态是跟踪算法显著的特性之一,它决定了跟踪过程中估计的参数。传统的目标跟踪算法大多集中于基于生成式或判别式的统计模型。文献[3]为了捕捉跟踪目标的时空特性,提出一种高斯空间颜色混合的外观模型,该模型能够同时编码空间布局和颜色信息;文献[4]基于核的生成式方法,建立一个以颜色直方图为基础的视觉表示正则化的空间平滑各向同性核,利用 Bhattacharyya 系数作为相似性度量,进而实现目标定位;文献[5]基于子空间的生成式方法,应用偏最小二乘分析学习低维特征子空间进行目标跟踪;文献[6]基于稀疏表示的生成式方法,利用压缩感知方法生成低维的压缩特征描述子,提升实时跟踪算法的性能。近年来,判别式跟踪方法因其强大的判别学习特性,在目标跟踪任务中被广泛应用,其中基于在线 Boosting[7]、支持向量机、随机学习的目标跟踪方法得到广泛研究和改进。

目前,深度学习方法[8]在图像分类和目标跟踪领域应用较为广泛。卷积神经网络拥有自动特征提取特性,在目标跟踪领域也得到应用。然而,现在大多数卷积网络中使用的主干网络[9]如:VGG[10]、ResNet[11]、DenseNet[12]等都是针对图像分类任务在 Imagenet 图像数据集上进行训练的。由于不能直接用于目标跟踪任务,研究者通常使用集成方法[13]和跨层特征选择[14]两种方式对已有网络进行处理,使之适应目标跟踪任务。因此,影响了目标跟踪任务的效率和增大了复杂度。

本文以基于多层竞争融合模型的目标检测预训练区域建议网络为基础,进行行人跟踪。采用一种长期和短期相结合的方式实现在线训练算法,不断学习跟踪目标的形变特征。

## 1 目标检测原理

### 1.1 区域建议网络

区域建议网络(Region proposal networks, RPN)是 2016 年提出的 Faster RCNN[15-16]目标检测框架中的一个子结构,用于在第 1 阶段的检测中生成多个候选区域,替代了选择搜索(Selective search)和边缘框图(Edge boxes)这些传统的候选区域生成方法,效果显著。区域建议网络使用任意大小的图像作为输入,最终输出多个预测得到的边框坐标及其对应的置信度。整体结构如图 1 所示。

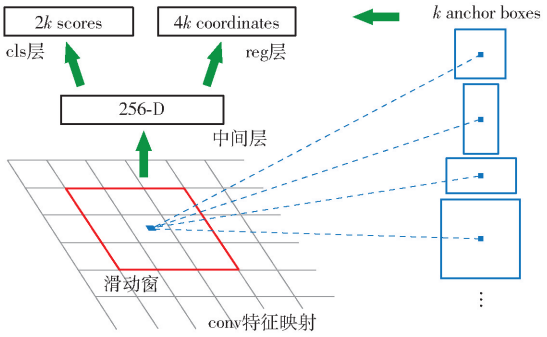



图 1 区域建议网络整体结构

Fig. 1 Overall structure of RPN

(1)anchor

在每一个滑动窗口的位置同时预测多个候选区域,每个位置有 $k$ 个大小不同的可能候选区域。因此边框回归层(reg layer)对应生成 $4k$ 个值,置信度预测层生成 $2k$ 个结果。这 $k$ 个候选区域被参数化地对应到 $k$ 个相关的参考区域上,这些参考区域就被称为 anchor。

(2)损失函数

在训练区域建议网络过程中,对每个 anchor 使用一种二元分类标记。把两种类型的 anchor 标记为正样本,一种是和真实目标重叠面积(IOU)最大的 anchor,另一种是和真实目标重叠面积(IOU)超过 0.7 的 anchor。因此一个真实目标可能对应多个 anchor,会生成多个对应的正样本。选取和真实目标重叠面积 IOU 值小于 0.3 的 anchor 作为负样本,在目标的边框回归层中这些负样本不进行计算。

区域建议网络使用的损失函数为

$$L(\boldsymbol{p}_i,\boldsymbol{t}_i)=\frac{1}{N_{\mathrm{cls}}}\sum_i L_{\mathrm{cls}}(\boldsymbol{p}_i,\boldsymbol{p}_i^*)+\lambda\frac{1}{N_{\mathrm{reg}}}\sum_i \boldsymbol{p}_i^* L_{\mathrm{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*) \tag{1}$$

式中 $i$——每次训练中不同 anchor 的序号

$\boldsymbol{p}_i$——anchor 置信度的预测值

$\boldsymbol{p}_i^*$——anchor 作为正样本的真实值

$\boldsymbol{t}_i$——回归层生成的候选区域的坐标向量

$\boldsymbol{t}_i^*$——正样本对应的真实边框的坐标向量

$N_{\mathrm{cls}}$——分类数 $L_{\mathrm{cls}}$——分类损失

$N_{\mathrm{reg}}$——回归数 $L_{\mathrm{reg}}$——回归损失

置信度预测的损失函数和回归层预测的损失函数的值都进行了归一化,并在回归损失函数前乘以

$\lambda$ 来平衡两项的差距。对于边框回归函数，计算后的 4 种参数化坐标为

$$\begin{cases} t_x = \dfrac{x - x_a}{w_a} \\ t_y = \dfrac{y - y_a}{h_a} \\ t_w = \lg(w/w_a) \\ t_h = \lg(h/h_a) \end{cases} \tag{2}$$

$$\begin{cases} t_x^* = \dfrac{x^* - x_a}{w_a} \\ t_y^* = \dfrac{y^* - y_a}{h_a} \\ t_w^* = \lg(w^*/w_a) \\ t_h^* = \lg(h^*/h_a) \end{cases} \tag{3}$$

式中　$x$、$y$——边框的中心坐标

$w$、$h$——边框的宽度、高度

$x^*$、$y^*$——真实边框坐标

$x_a$、$y_a$——anchor 边框的坐标

$w_a$、$h_a$——anchor 边框的宽度、高度

$t_x$、$t_y$——预测边框中心坐标

$t_w$、$t_h$——预测边框的宽度、高度

$w^*$、$h^*$——真实边框的宽度、高度

$t_x^*$、$t_y^*$——预测 anchor 边框中心坐标

$t_w^*$、$t_h^*$——预测 anchor 边框的宽度、高度

在坐标回归函数中，所有特征从最后一层特征图上以相同的大小卷积滤波器计算得到。为了计算不同大小的目标，共学习得到 $k$ 个回归器，每一个回归器都对应一种尺度和大小。因此，区域建议网络可以通过固定的特征图预测得到不同大小的目标。

### 1.2　MDNet

MDNet[17] 使用深度卷积神经网络进行目标跟踪并到达性能飞跃式提升，为了充分发挥卷积神经网络在视觉跟踪中的表现力，需要利用视觉跟踪专用的大规模数据对它们进行训练，这些数据包括了目标和背景组合中的广泛变化。然而，由于不同视频序列中包含完全不同的特性，因此卷积神经网络很难学习出统一的表达。

为了解决上述问题，提出了一种多领域网络[18]（Multi-domain network）来学习目标跟踪中多个不同标记的视频序列的共性，并将每一个视频看作一个分离的领域。所提出的网络在其末端为每个域生成一个特定的单独分支，并且共享之前所有序列中学习得到的通用表达特性。MDNet 中的每个域分别被训练和迭代，而在每次迭代中更新共享层[19]。通过这种策略，将每一个领域中的无关信息和特定信息分离，并学习用于视觉跟踪的通用特征表示。

具体的框架设计如图 2 所示，使用 107 像素 × 107 像素的 RGB 图像作为输入，构建了一个共 6 层的网络结构，其中前 3 层为卷积神经网络，中间 2 层为全连接神经网络，其中前面 5 层用于学习目标的共有表达特性。最后 1 层神经网络共有 $k$ 个分支，每个分支代表一个特定的目标域。卷积神经网络部分使用与 VGG－M 对应部分相同的结构和参数，后面 2 层全连接网络使用 512 个神经元，并包含相应的 Relu 和 Dropout 操作。最后 $k$ 个分支都使用 softmax 进行二元分类，将该视频中的目标和背景区分开来。

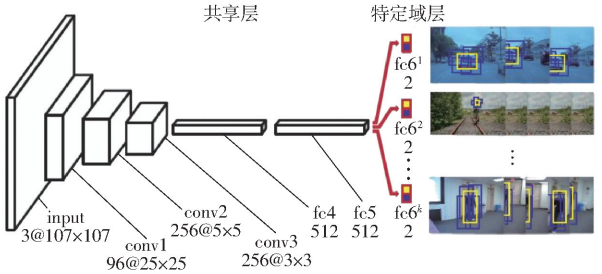



图 2　MDNet 整体结构图

Fig. 2　Overall architecture of MDNet

训练得到模型后，在测试过程中为新的视频序列构建一个新的特定域。然后固定之前的共享表达层，利用第 1 帧的图像数据对新产生的第 6 层全连接网络进行训练，学习该视频目标的特有表现。然后进行跟踪，在跟踪过程中使用在线跟踪方法，不断对最后的特定网络层进行微调，来适应目标的形变和遮挡等问题。

## 2　行人跟踪方法与实验

### 2.1　行人跟踪方法的细节设定

#### 2.1.1　预训练模型选择

目标跟踪算法使用文献[20]中的基于多层竞争融合模型的区域建议网络作为预训练模型。在目标跟踪任务中，使用在行人检测数据集上训练过的多层竞争融合模型有以下几个优点：

（1）多层级结构可以得到目标的全方面信息，包含细节信息和全局信息，在卷积神经网络的低层存在大量的细节特征，如颜色、纹理等特征，在其高层则包含更多的语义信息，采用多层级竞争融合结构可以使网络自适应地选择所需要的有用信息。

（2）可以很好地解决局部形变和部分遮挡问题，由于行人检测数据集中包含大量行人信息，通过数十万轮的训练之后，模型对于不同形态和部分遮挡的行人都有很好的识别效果，这表明模型提取了行人多个方面的特征，这些特征在处理行人跟踪问

题上也非常有效。

(3)对于环境因素影响有很好的鲁棒性,如光照不同、对比度不同等问题。一方面得益于预训练模型是在大规模数据集下进行训练而得到的,很多行人出现在阴影区域,这使得模型必须排除光照因素的干扰。另一方面是因为多层级结构能够获得更多的细节信息,分类器可以在光照变化时提取具有共性的特征对目标进行区分。

基于上述优点,故采用文献[20]提到的预训练模型进行目标跟踪。网络模型架构如图 3 所示。

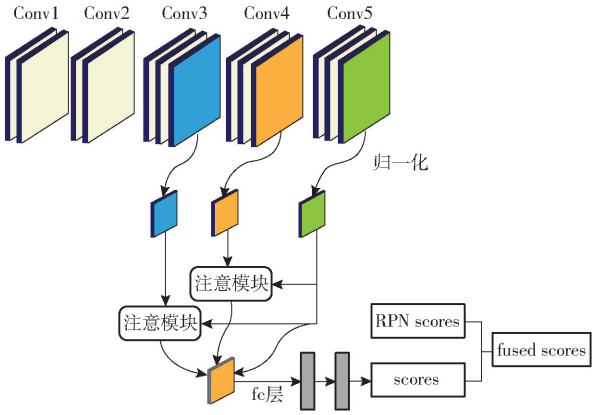



图 3 预训练网络 RCNN 阶段架构

Fig. 3 Overall structure in RCNN stage

### 2.1.2 输入图像尺寸选择

在 MDNet 中输入图像的尺寸为 107 像素 × 107 像素,这个尺寸的选择是全部依靠经验来完成的,在行人跟踪方法中采用基于感受野的原则来确定最终的输入大小。在卷积神经网络中,特定神经元的感受野是指输入图像中相关像素的数目。由于在跟踪任务中输入的都是较大的包含整个目标的图像,因此感受野的选择非常重要。如果神经元的感受野小于输入图像中的目标,生成的特征就会太局部化,从而不能得到目标的全部表现;如果神经元的感受野大于输入图像中的目标,那么得到的特征对于最终的分类器来说就会变得过于冗余。

在本文中,使用以 VGG16 网络为基础的区域建议网络[21-22],在网络的 Conv5 层特征图上生成候选区域,这一层的感受野为 163 像素 × 163 像素。在 MDNet 中,本文提出的观点是除目标区域外让输入图像覆盖一小部分背景内容是有益的,因此使用 224 像素 × 224 像素作为输入图像的尺寸,两个相邻的 anchor 之间的距离为 16 像素。

### 2.1.3 跟踪算法 anchor 设计

区域检测网络[21]中一个非常重要的方面就是 anchor 的设计,一个 anchor 本质上代表了特定特征图上 1 像素 × 1 像素位置的特征。在区域建议网络中,目标置信度及其对应的边框回归值都是以这个特征为基础进行计算的。通过采用 anchor 的方式为计算真实边框与 anchor 之间的重叠面积(IOU)提供基础。进而可以判定 anchor 对应的区域是正样本还是负样本。通常使用 0.5 或者 0.7 作为 IOU 的阈值,这样每一个特定位置上的 anchor 就可以被分成目标或者背景。这个设计保证了置信度预测和边框回归预测的有效性。

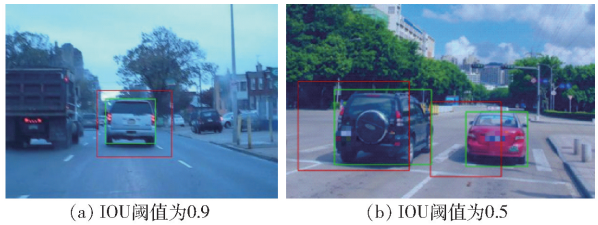
anchor 在设计上的灵活性使得可以通过改变原始的 anchor 设计来更好地适应行人跟踪任务。由于在跟踪任务中,只能使用第 1 帧中的目标来训练数据,不能负担大规模参数的微调(fine-tune),故跟踪任务中 anchor 的设计首先要简单,使整体网络在有限的数据量下仍然能够训练起来。其次,新设计的 anchor 还要和预训练网络中的 anchor 兼容,使离线网络不需要额外训练就可以直接使用。

因此,在行人跟踪任务中,只使用一种宽高比和一种尺度的 anchor 来减少需要训练的参数数量,考虑到与预训练模型的兼容性和网络输入的图像分辨率大小,使用高度为 161 像素和宽高比 0.41 作为 anchor 的参数。

### 2.1.4 改进的损失函数

原始区域建议网络的损失函数[23]如式(1)所述,共包含两项,一项用于计算类别预测的误差,另一项计算边框回归值的误差。如果在跟踪任务上直接使用原始损失函数就会造成 anchor 边框和其对应目标边框 IOU 阈值难以选择的问题,当使用较大阈值时,随着跟踪时间推移,预测得到的边框则会越来越大,如图 4a 所示。使用较小阈值可以提高边框回归值的准确性,但是在跟踪变化速度很快的目标时会产生漂移现象从而导致跟踪失败,如图 4b 所示。

这些问题出现的主要原因在于原始的边框回归损失项在训练时没有得到高品质的真实标记数据,在跟踪任务时只在第 1 帧含有真实标记数据,后续在线更新用到的都是根据当前帧预测得到的数据。这使得边框回归损失项的噪声变得非常大,最终这些误差的累积导致跟踪结果和真实目标偏差很大。如果去掉边框回归损失项并使用一个较小的 IOU 阈值,基于区域建议网络的追踪方法就变得和



(a) IOU阈值为0.9 (b) IOU阈值为0.5

图 4 不同 IOU 阈值效果

Fig. 4 Prediction result with different thresholds of IOU

MDNet 非常相似。

为了保证跟踪和边框回归同时得到一个比较好的结果，则通过改变区域建议网络的损失函数来达到这一目标。去掉边框回归损失项，使用两个类别损失项来优化类别判断并保证回归框的品质。新的损失函数为

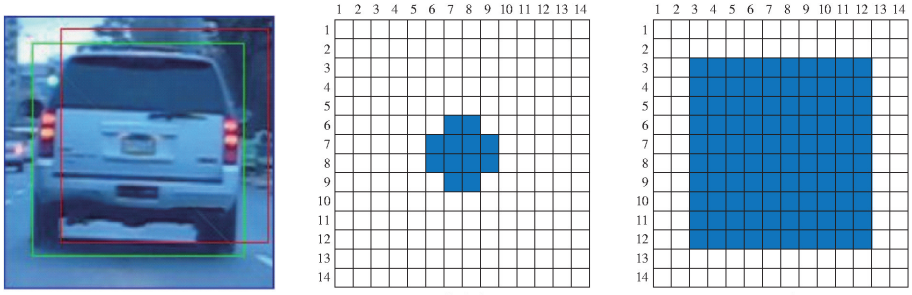
$$L=\alpha\sum_i L_{\text{cls}}(a_i,a_i^*)+\beta\sum_i L_{\text{cls}}(b_i,b_i^*) \quad (4)$$

式中　$a_i$、$b_i$——区域建议网络预测得到的两个不同的类别置信度值

$a_i^*$、$b_i^*$——其对应的真实样本类别值

$\alpha$、$\beta$——两项间的平衡权重

图 5a 为原图，其中绿色为真实边框，红色是能与其匹配的 anchor 边框。$a_i^*$ 使用 IOU 阈值为 0.7 来生成正负样本，如图 5b 所示，强调样本类别分类。$b_i^*$ 不使用 anchor 和真实样本之间的 IOU 值作为选取正负样本的判断指标，而是直接使用原图在特征图上的映射区域作为正样本，如图 5c 所示，强调回归框的品质。



(a) 原图　(b) 分类标记　(c) 回归标记

图 5　不同标记方式得到的正负样本

Fig. 5　Positive and negative samples with different labels

## 2.2　在线跟踪方法

### 2.2.1　在线算法的更新策略

在线算法的更新策略[24]要满足鲁棒性和适应性两个要求，因此使用长期(long-term)和短期(short-term)并存的更新策略。长期更新[25]是在固定的时间间隔后定时更新，所用到的正样本采集于这个固定的时间间隔之内。短期更新[26]则是在跟踪失败时进行更新，如将所要跟踪的目标判别为背景时，使用的正样本采集于一个较短的时段内。在两种更新策略中，长期更新生成时间间隔较长的负样本通常会冗余，并且和当前帧的关联性较小，因此只使用短期时段内产生的负样本。

为了估计每一帧中目标的状态，从前一帧使用跟踪算法估计的目标周围选取 $n$ 个目标分别为 $x_1$，$x_2$，…，$x_n$，然后将这 $n$ 个目标区域作为输入放入区域建议网络中进行计算，分别预测得到每个区域正样本的概率 $f^+(x_i)$ 和负样本概率 $f^-(x_i)$，然后通过每个区域的正样本概率的最大值确定最终的跟踪结果，公式为

$$x^*=\arg\max_{x_i} f^+(x_i) \quad (5)$$

### 2.2.2　困难样本挖掘

在跟踪算法中大部分的负样本都是非常平凡或者冗余的，只有一小部分的负样本在训练过程中对分类器非常有效。由于缺乏有效训练样本，当使用普通的随机梯度下降法进行训练时，每一个训练样本对训练结果的贡献完全相等，这使得分类器容易受到漂移等问题的干扰。在目标检测领域中常用的解决方法是困难样本挖掘，该方法以一种训练和测试交替的方式来得到难以区分的负样本，如一些假正性的样本[27](False positive examples)，在跟踪算法中使用这种思想进行在线算法学习。

在每一帧选择得到的样本中，一共包含了 $M^+$ 个正样本和 $M^-$ 个负样本，其中负样本中存在一些和正样本非常相似而难以区分的负样本 $M_h^-$，通常情况下 $M^-$ 的个数远大于 $M_h^-$ 的个数。在困难样本挖掘中，首先对所有样本进行测试，选择出正样本概率 $f^+(x_i)$ 最高的多个负样本作为困难负样本，然后将这些困难负样本放入网络中进行训练。

### 2.2.3　在线跟踪算法流程

整个在线跟踪算法流程为：

(1)准备预训练的 CNN 网络。

(2)初始帧中目标 $x_1$ 的状态。

(3)从初始帧中收集正负样本 $S_1^+$ 和 $S_1^-$。

(4)使用 $S_1^+$ 和 $S_1^-$ 更新区域建议网络。

(5)初始化 $T_s\leftarrow\{1\}$，$T_l\leftarrow\{1\}$。

(6)重复步骤(1)~(5)。

(7)选出候选样本 $x_t^i$，通过式(5)得到最优的目标状态 $x_t^*$。

if $f^+(x^i)>0.5$ then

选出对应的正负样本 $S_i^+$ 和 $S_i^-$

更新 $T_s\leftarrow T_s\cup\{t\}$，$T_l\leftarrow T_l\cup\{t\}$

if $|T_s|>t_s$ then $T_s\leftarrow T_s\backslash\{\min_{v\in t_s}v\}$

if $|T_l|>t_l$ then $T_l \leftarrow T_l \setminus \{\min_{v\in t_l} v\}$

(8)使用边框回归调整 $x_t^*$。

if $f^+(x^i)\leqslant 0.5$ then

使用 $S^+_{v\in t_s}$ 和 $S^-_{v\in t_s}$ 更新区域检测网络

else if $t$ mod1 0 = 0 then

使用 $S^+_{v\in t_s}$ 和 $S^-_{v\in t_s}$ 更新区域检测网络

until 视频结束。

在整个在线跟踪算法中,首先使用预训练过的区域建议网络来初始化参数值,然后在第 1 帧时选取 500 个正样本和 5 000 个负样本对区域建议网络进行微调。区域建议网络的前 2 个卷积层 Conv1 和 Conv2 的参数固定不变。$T_s$ 和 $T_l$ 分别为长期更新和短期更新的帧的索引集,长期更新的索引集长度 $t_l=100$,短期更新的索引集长度 $t_s=20$。在第 2 帧开始进入循环,通过式(5)得到最优的目标状态 $x_t^*$,如果当前最优状态 $x_t^*$ 的 $f^+(x^i)>0.5$ 则收集对应的正负样本,并更新当前的帧索引集,把当前帧的索引添加进去,如果长期更新的索引集 $|T_s|>t_s$,则把索引集中最小的索引去掉,同理处理短期更新的索引集,然后使用边框回归对得到的最优状态进行调整。如果当前最优状态 $x_t^*$ 的 $f^+(x^i)\leqslant 0.5$,则使用收集到的短期更新的正负样本通过 $S^+_{v\in t_s}$ 和 $S^-_{v\in t_s}$ 来更新区域检测网络,如果当前帧数是 10 的倍数,使用长期更新索引集中的正样本和短期更新索引集中的负样本对区域检测网络进行更新。

### 2.3 行人跟踪算法实验

为了验证本文提出的目标检测和跟踪算法的有效性,采用带有复杂背景的公开数据集对基于多层竞争融合模型的区域建议网络进行模型验证,通过真实场景下的行人检测与跟踪实验,验证本文提出的行人跟踪算法的性能,为后续研究农业领域目标跟踪机器人提供可行性依据。

实验采用处理器为因特尔 i7 - 6700hq,主频为 2.6 GHz,内存为 16 GB,显卡为 GTX960,操作系统为 Ubuntu 16.04 的计算机来模拟云端进行行人检测和跟踪。并且通过部署 caffe 深度学习框架和 Python 2.7 实现整体算法。实验采用的软硬件平台符合系统要求,可以为行人检测和跟踪算法实验提供可靠支持,实现后续对系统云端的设计目标。通过录制一段真实场景下行人视频,对其中的行人进行检测和跟踪,来检验算法的可行性和有效性。

## 3 结果分析

### 3.1 基于公开数据集的算法验证

(1)基于多层竞争融合模型的区域建议网络性能评估结果

采用 Caltech 行人检测数据集对基于多层竞争融合模型的区域建议网络进行测评,Caltech 行人检测数据集共包含约 35 000 个有标记的行人边缘框图。在训练阶段,使用 10 倍于标准训练数据的数量对网络进行训练。在测试阶段,使用标准测试数据集中的 4 024 幅图像并在 reasonable 设置下进行评价,最终用 FPPI - MR 评价标准来衡量算法的整体表现。

图 6a 显示了使用基于多层竞争融合模型的 CA 和其他多种方法在 Caltech 行人检测数据集上的对比。可以看到,基于多层竞争融合模型的区域建议网络框架的丢失率为 7.02%,达到了目前行人检测任务中的最优结果。图 6b 显示了单一的区域建议网络和 RPN + BF 网络的结果对比。结果表明,RPN + BF 方法中的区域建议网络检测的丢失率比单一区域建议网络的检测器丢失率低近 5%,达到目标检测任务评价要求。这证明了通过竞争关注模型进行的多层级和多尺度的特征融合非常适合区域建议网络,并使其达到了更好的效果。区域建议网络与全连接网络相比,是一种简单且不具有强分类作用的分类器,因此在使用更具有判别性的特征时会带来更大的提升。

(2)基于区域建议网络的行人跟踪算法验证

为了验证基于区域建议网络的行人跟踪方法的准确性,使用几个用于行人跟踪的公开数据集进行

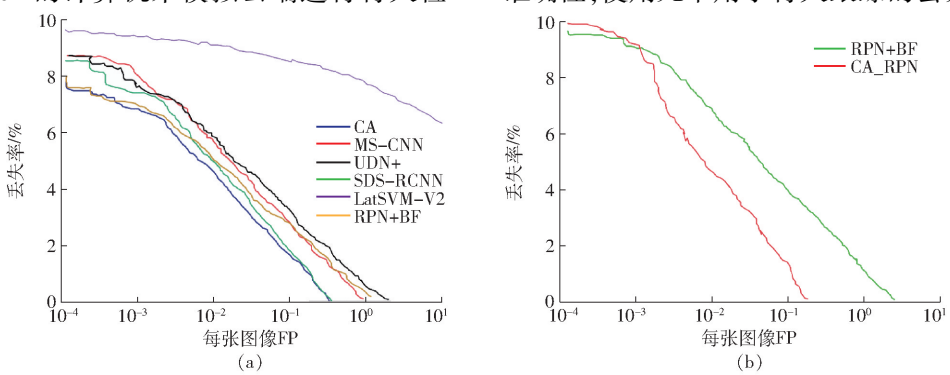



图 6 基于多层竞争融合模型的区域建议网络性能评估结果

Fig. 6 Performance evaluation results of multi-layer competitive fusion RPN

测试。其中包括 ETH、PETS 2009 和 Venice 3 个数据集，这 3 个数据集都收集了大量行人视频。

其中 ETH 数据集包含了多个不同的场景，并且拍摄角度变化非常大，视频的分辨率为 640 像素 × 480 像素，每秒 14 帧，行人的衣着和姿态变化范围比较大，并且存在着一定的行人间相互遮挡和行人被其他物体遮挡的情况。

PETS 2009 数据集和 Venice 数据集的环境场景相似，都是在空旷操场或者空旷街头利用固定摄像头拍摄得到的，PETS 2009 采用高空放置的摄像头进行拍摄，Venice 采用和人等高的摄像头拍摄，视频中的行人变化情况比较小，行人衣着上也比较相似。其中 PETS 2009 数据集的视频分辨率为 768 像素 × 576 像素，每秒 7 帧，存在着一定的行人间相互遮挡，行人被其他物体遮挡的情况比较少。Venice 数据集的视频分辨率为 1 920 像素 × 1 080 像素，每秒 30 帧，也存在着一定的行人间相互遮挡情况，同时也有行人被其他物体遮挡的情况。

图 7 ~ 9 分别展示了基于区域检测网络的行人跟踪方法在不同数据集上的表现结果。本文使用的基于区域检测网络的跟踪方法属于深度学习方法，参数量远大于传统算法，相对于传统方法计算效率处于劣势。但是可以看出，在图 7 ~ 9 几个环境比较复杂的行人数据集中，所使用的算法取得了不错的效果。

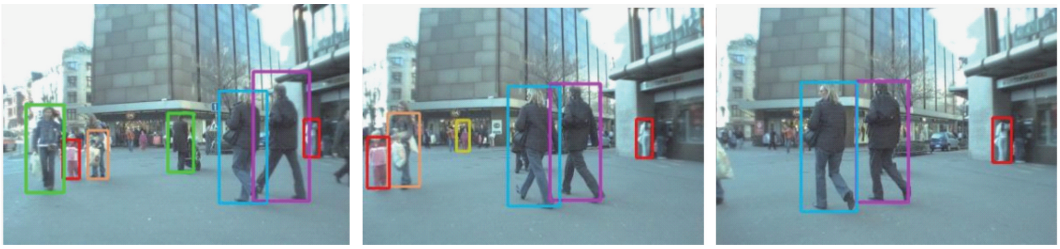


图 7　ETH 数据集跟踪效果

Fig. 7　Results of tracking effect on ETH dataset

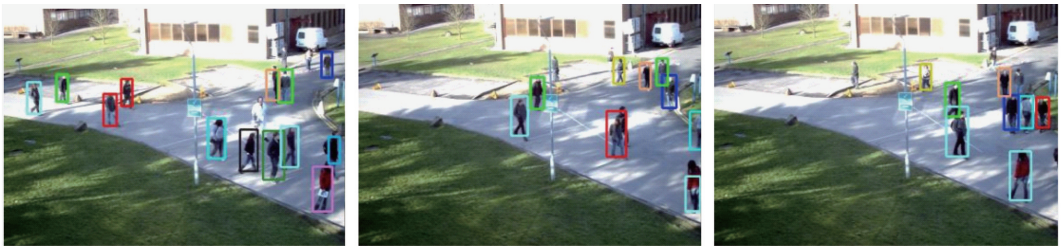


图 8　PETS 2009 数据集跟踪效果

Fig. 8　Results of tracking effect on PETS 2009 dataset



图 9　Venice 数据集跟踪效果

Fig. 9　Results of tracking effect on Venice dataset

### 3.2　基于行人跟踪算法实验结果分析

在实验中，先使用行人检测算法对视频每隔两帧进行检测，当检测到行人之后把检测结果传给行人跟踪算法，然后进行行人跟踪，直到视频结束。可以看出所设计的算法可以良好运行并得到行人检测和跟踪结果，由于使用的环境相同，未来在机器人和云端进行部署时也能达到较好的效果。模拟环境所用到的便携式计算机无法满足实时检测需求，但是在搭建云端服务器时使用多块性能优秀的显卡可以满足实时检测和跟踪需求。真实行人跟踪实验效果如图 10 所示。

## 4　结论

(1)阐述了区域建议网络的原理和 MDNet 算法整体框架，提出了改进的区域建议网络损失函数，这使跟踪和边框回归效果更好。

(2)提出了一种使用长期和短期并存的在线算法更新策略，设计了跟踪算法实验流程。通过实验和公开数据集的验证，基于多层竞争融合模型的区域建议网络的目标检测丢失率为 7.02%，达到了目前行人检测任务中的最优结果。RPN + BF 方法中的区域建议网络检测的丢失率比单一区域建议网络

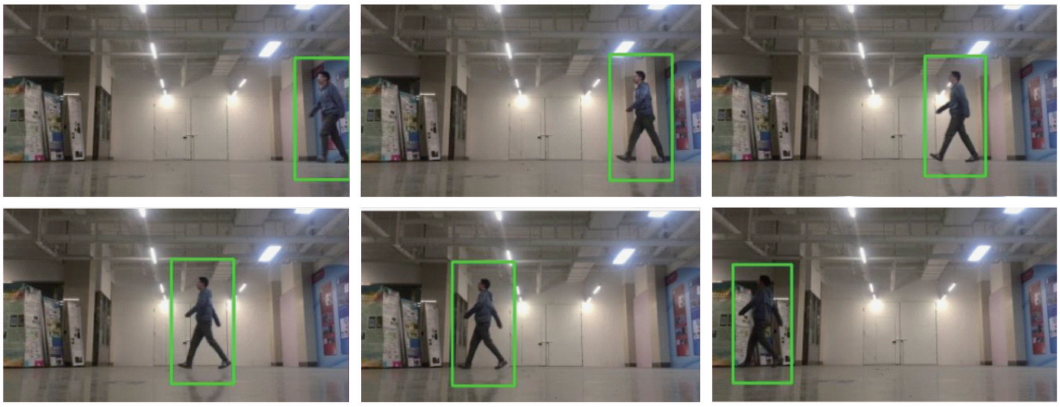


图 10 行人跟踪实验效果

Fig. 10 Results of pedestrian tracking experiment

的检测器丢失率降低近 5%。

(3)在线目标跟踪算法实验验证结果表明,本文提出的算法能很好地实现行人实时跟踪,对未来农业领域目标跟踪机器人和云端部署提供了理论依据。

## 参 考 文 献


[1] WANG H, SUTER D, SCHINDLER K, et al. Adaptive object tracking based on an effective appearance filter[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1661 – 1667.

[2] THOMPSON W B, PONGT C. Detecting moving objects[J]. International Journal of Computer Vision, 1990, 4(1): 39 – 57.

[3] WANG Q, CHEN F, XU W L, et al. Object tracking via partial least squares analysis[J]. IEEE Transactions on Image Processing, 2012, 21(10): 4454 – 4465.

[4] COMANICIU D, RAMESH V, MEER P. Kernel-based object tracking[J]. Pattern Analysis & Machine Intelligence, 2003, 25(5): 564 – 575.

[5] OUYANG W, WANG X. A discriminative deep model for pedestrian detection with occlusion handling[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3258 – 3265.

[6] ZHANG K, ZHANG L, YANG M H. Real-time compressive tracking[C]// European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 864 – 877.

[7] LI X, HU W, SHEN C, et al. A survey of appearance models in visual object tracking[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(4): 58.

[8] 卢宏涛,张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理, 2016, 31(1): 1 – 17.
LU Hongtao, ZHANG Qinchuan. Applications of deep convolutional neural network in computer vision[J]. Journal of Data Acquisition and Processing, 2016, 31(1): 1 – 17. (in Chinese)

[9] OUYANG W, WANG X, ZENG X, et al. Deepid-net: deformable deep convolutional neural networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2403 – 2412.

[10] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4293 – 4302.

[11] KRISTAN M, MATAS J, LEONARDIS A. The visual object tracking vot2015 challenge results[C]// Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015: 1 – 23.

[12] QU J, LIU Z. Non-background HOG for pedestrian video detection[C]// 2012 8th International Conference on Natural Computation. IEEE, 2012: 535 – 539.

[13] 姚雪琴,李晓华,周激流. 基于边缘对称性和 HOG 的行人检测方法[J]. 计算机工程, 2012, 38(5): 179 – 182.
YAO Xueqin, LI Xiaohua, ZHOU Jiliu. Pedestrian detection method based on edge symmetry and HOG[J]. Computer Engineering, 2012, 38(5): 179 – 182. (in Chinese)

[14] 种衍文,匡湖林,李清泉. 一种基于多特征和机器学习的分级行人检测方法[J]. 自动化学报, 2012, 38(3): 375 – 381.
CHONG Yanwen, KUANG Hulin, LI Qingquan. Two-stage pedestrian detection based on multiple features and machine learning[J]. Acta Automatica Sinica, 2012, 38(3): 375 – 381. (in Chinese)

[15] SMEULDERS A W M, CHU D M, CUCCHIARA R, et al. Visual tracking: an experimental survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1442 – 1468.

[16] REN S, HE K, GIRSHICK R, et al. Faster R – CNN: towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems, 2015: 91 – 99.

[17] WEN L, CAI Z, LEI Z, et al. Online spatio-temporal structural context learning for visual tracking[C]// European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012: 716 – 729.


(下转第 379 页)

*lucidum*[J]. Journal of Henan Institute of Science and Technology (Natural Science Edition), 2018,41(1):54 - 58. (in Chinese)

[13] 李路,黄汉英,赵思明. 大米蛋白质、脂肪、总糖、水分近红外检测模型研究[J]. 中国粮油学报,2017,32(7):121 - 126.
LI Lu, HUANG Hanying, ZHAO Siming. NIR spectra detection model of protein, fat, total sugar and moisture in rice[J]. Journal of the Chinese Cereals and Oils Association, 2017,32(7): 121 - 126. (in Chinese)

[14] 石振兴,朱茵茵,杨修仕,等. 藜麦粉末中主要营养成分近红外预测模型的建立及验证[J]. 粮食与油脂, 2017, 30(12): 55 - 57.
SHI Zhenxing, ZHU Yinyin, YANG Xiushi, et al. Development of near-infrared reflectance spectroscopy (NIRS) prediction model for the quality components of flour in quinoa[J]. Cereals & Oils, 2017, 30(12):55 - 57. (in Chinese)

[15] 汪庆平,黎其万,董宝生,等. 近红外光谱法快速测定山核桃品质性状的研究[J]. 西南农业学报,2009,22(3):873 - 875.
WANG Qingping, LI Qiwan, DONG Baosheng, et al. Study on the prediction of oils of *Carya cathayensis* Sargent using near infrared reflectance spectroscopy[J]. Southwest China Journal of Agricultural Sciences, 2009, 22(3):873 - 875. (in Chinese)

[16] 徐艳飞,回瑞华,侯岩冬. 新疆种植核桃与野生核桃蛋白质含量的测定[C]//中国化学会第 26 届学术年会分析化学分会场论文集. 中国化学会,2008.

[17] 马文强,张漫,李忠新,等. 基于近红外光谱的核桃仁蛋白质含量检测分析[J]. 农业机械学报,2017,48(增刊):407 - 411.
MA Wenqiang, ZHANG Man, LI Zhongxin, et al. Detection and analysis of walnut protein content based on near infrared spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017,48(Supp.):407 - 411. (in Chinese)

[18] 王彬,王巧华,肖壮,等. 基于可见-近红外光谱及增强回归树算法的鸡蛋种类鉴别[J]. 华中农业大学学报,2018, 37(1):95 - 100.
WANG Bin, WANG Qiaohua, XIAO Zhuang, et al. Identitying egg varieties based on regression trees algorithm and visible-near-infrared spectrum[J]. Journal of Huazhong Agricultural University, 2018,37(1):95 - 100. (in Chinese)

[19] 王海龙,杨国国,张瑜,等. 竞争性自适应重加权算法和相关系数法提取特征波长检测番茄叶片真菌病害[J]. 光谱学与光谱分析,2017,37(7):2115 - 2119.
WANG Hailong, YANG Guoguo, ZHANG Yu, et al. Detection of fungal disease on the tomato leaves with competitive adaptive reweighting sampling and correlation analysis methods[J]. Spectroscopy and Spectral Analysis, 2017,37(7):2115 - 2119. (in Chinese)

[20] 姜微,房俊龙,王树文,等. CARS - SPA 算法结合高光谱检测马铃薯还原糖含量[J]. 东北农业大学学报,2016,47(2): 88 - 95.
JIANG Wei, FANG Junlong, WANG Shuwen, et al. Using CARS - SPA algorithm combined with hyperspectral to determine reducing sugars content in potatoes[J]. Journal of Northeast Agricultural University, 2016, 47(2): 88 - 95. (in Chinese)

(上接第 338 页)

[18] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM International Conference on Multimedia. ACM, 2014: 675 - 678.

[19] WU Y, LIM J, YANG M H. Online object tracking: a benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2411 - 2418.

[20] CHEN Zhichang, ZHANG Li, KHATTAK A M, et al. Deep feature fusion by competitive attention for pedestrian detection[J]. IEEE Access, 2019, 7: 21981 - 21989.

[21] ZHANG J, MA S, SCLAROFF S. MEEM: robust tracking via multiple experts using entropy minimization[C]//European Conference on Computer Vision. Springer, Cham, 2014: 188 - 203.

[22] ZHONG W, LU H, YANG M H. Robust object tracking via sparsity-based collaborative model[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1838 - 1845.

[23] KALAL Z, MATAS J, MIKOLAJCZYK K. Pn learning: bootstrapping binary classifiers by structural constraints[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 49 - 56.

[24] SERMANET P, KAVUKCUOGLU K, CHINTALA S, et al. Pedestrian detection with unsupervised multi-stage feature learning [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3626 - 3633.

[25] ROSS D A, LIM J, LIN R S, et al. Incremental learning for robust visual tracking[J]. International Journal of Computer Vision, 2008, 77(1 - 3): 125 - 141.

[26] BAO C, WU Y, LING H, et al. Real time robust l1 tracker using accelerated proximal gradient approach[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 1830 - 1837.

[27] PEREZ P, HUE C, VERMAAK J, et al. Color-based probabilistic tracking[C]//European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2002: 661 - 675.