·图形图像处理· 文章编号：1000-3428(2019)09-0216-06 文献标志码：A 中图分类号：TP751

一种用于单目标跟踪的锚框掩码孪生 RPN 模型

李明杰，冯有前，尹忠海，周 诚，董方昊

（空军工程大学 基础部，西安 710051）

摘 要： 针对孪生区域候选网络（RPN）易受干扰且目标丢失后无法跟踪的问题，引入锚框掩码网络机制，设计一种新型孪生 RPN 模型。设置多尺度模板图片，并将其与目标图片进行卷积操作，实现全图检测以避免目标丢失。通过对前三帧图片的 IOU 热度图进行学习，预测连续帧目标锚框掩码，简化计算并排除其他目标干扰。在 VOT2016 和 OTB100 数据集中的实验结果显示，该模型对 VOT2016 数据集检测帧率达到 24.6 frame/s，预期平均覆盖率为 0.344 5，对 OTB100 数据集的检测准确率和成功率分别为 0.862 和 0.642。基于摄像头采集数据的目标丢失及干扰测试表明，该模型具有良好的抗干扰性与实时性。

关键词： 孪生区域候选网络；锚框掩码；锚框掩码网络；多尺度变换；目标跟踪

开放科学（资源服务）标志码（OSID）：

中文引用格式： 李明杰，冯有前，尹忠海，等. 一种用于单目标跟踪的锚框掩码孪生 RPN 模型[J]. 计算机工程，2019，45(9)：216-221.

英文引用格式： LI Mingjie，FENG Youqian，YIN Zhonghai，et al. An anchor mask Siamese RPN model for single target tracking[J]. Computer Engineering，2019，45(9)：216-221.

An Anchor Mask Siamese RPN Model for Single Target Tracking

LI Mingjie，FENG Youqian，YIN Zhonghai，ZHOU Cheng，DONG Fanghao

（Department of Sciences，Air Force Engineering University，Xi'an 710051，China）

【Abstract】 To address the problem that the Siamese Region Proposal Network（RPN） is susceptible to interference and cannot be tracked after the target is lost，this paper introduces the anchor mask network mechanism to design a new Siamese RPN model. The model sets the multi-scale template images and convolves them with the target image to achieve full-image detection and avoid target loss. By learning the IOU hot maps of the first three frames，the target anchor mask is predicted in the continuous frames to simplify the calculation and exclude other target interference. The experimental results in the VOT2016 and OTB100 datasets show that the model has a detection rate of 24.6 frame/s and an expected average overlap of 0.344 5 for the VOT2016 dataset，and a precision of 0.862 and a success rate of 0.642 for the OTB100 dataset. The target loss and interference tests are carried out on the data collected by the camera. The results show that the model has good anti-interference and real-time performance.

【Key words】 Siamese Region Proposal Network（RPN）；anchor mask；anchor mask network；multi-scale transformation；target tracking

DOI： 10.19678/j.issn.1000-3428.0053937

0 概述

目标跟踪在计算机视觉领域是一项基础任务，对于目标行为分析及确定目标位置等其他任务具有重要的作用。近年来许多针对目标跟踪任务的算法被相继提出[1-3]，尤其深度学习的广泛应用使得跟踪效果越来越好。在目标跟踪的主流方法中有一类方法为结合深度特征条件的检测跟踪网络。其中：C-COT[4]作为一种基于判别式学习的特征点跟踪方法通过学习连续卷积算子实现目标跟踪；T-CNN[5]网络通过拓展静态图像检测框架并融合具有时空特性的 Tubelets 实现目标跟踪；EBT[6]网络通过类物

基金项目： 国家自然科学基金（61472443）。

作者简介： 李明杰（1995—），男，硕士研究生，主研方向为计算机视觉、目标检测；冯有前，教授、博士生导师；尹忠海，教授；周 诚，博士研究生；董方昊，硕士研究生。

收稿日期： 2019-02-19 **修回日期：** 2019-03-25 **E-mail：** 13072988703@163.com

体度量生成少量高质量的建议框,再以现有的跟踪检测方法进行评估实现目标跟踪;文献[7]提出一种图像重分块检测机制对 SAMF 算法进行改进,提高了遮挡目标的跟踪准确率。除此以外,基于相关滤波的跟踪网络也取得较好的效果。其中:ECO[8]通过使用相关滤波算法实现特征降维,并利用高斯混合模型生成新的训练集,提高了跟踪精度与速度;KCF[9]通过使用目标周围区域的循环矩阵采集正负样本,训练目标检测器提高检测速度,并融合多通道数据,高效地实现目标跟踪;基于非线性核相关滤波的全景视觉目标跟踪算法[10]将岭回归与循环样本矩阵和经典的相关滤波进行联系,设计了适用于全景成像的自适应机制和基于极坐标表示的目标搜索机制,提高了跟踪精度。特别地,孪生区域候选网络(Region Proposal Network,RPN)[11]在孪生全卷积网络(Fully Convolutional Network,FCN)[12]的基础上引入 RPN 原理,以极快的速度达到了单目标跟踪的高准确率。

在对图像中目标进行跟踪时,孪生 RPN 只提取单帧图像的特征,没有结合之前的检测结果,因此,在具有相似图像目标的条件下,可能会产生误识别的情况,并且其检测位置是在上一帧图片目标位置一定范围内进行截图操作,未对整张图片进行处理,若视频中的目标曾被遮挡或消失,会导致跟踪失败。针对上述问题,本文构建一种多尺度模板图片的孪生 RPN 模型,并结合视频分析中常用的 3D 卷积操作和 FCN[13]网络引入固定卷积操作范围的锚框掩码机制,使模型适用于视频图片。

1　相关知识

1.1　3D 卷积

文献[14]将 3D 卷积用于视频维度的分析,随后文献[15]提出适用于视频且帧率极快的 C3D(Convolutional 3D)网络,此后基于 C3D 网络的行为识别方法相继被提出。本文提出的锚框掩码网络借助 3D 卷积学习了连续 3 帧图片的 IOU 热度图信息,与后续的类 FCN 预测层进行相连,估计下一帧图片的锚框掩码图像。3D 卷积原理如图 1 所示。

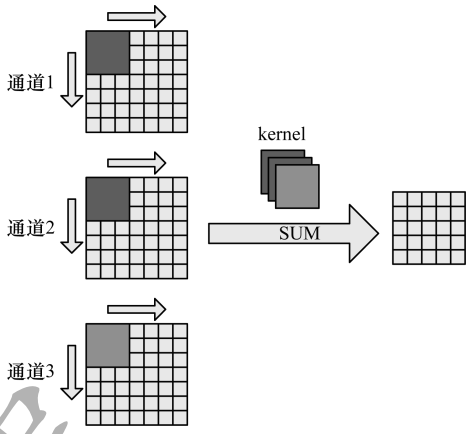


图 1　3D 卷积

1.2　全卷积网络

FCN 是一种用于图像分割的经典网络,在其基础上诞生了许多衍生网络,如 R-FCN[16]、Mask R-CNN[17]、全连接 CRF[18-19]等。FCN 通过对图像进行特征提取、全连接及分类,最后反卷积至原图大小,使得网络针对原图每个像素点都可进行分类,最终实现图像分割,其结构如图 2 所示。

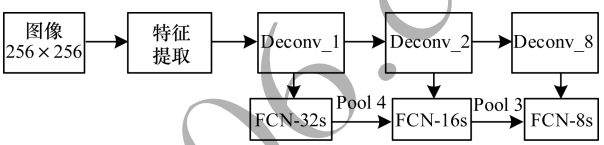


图 2　FCN 结构

本文借鉴 FCN 网络结构,在前段接入 3D 卷积学习 IOU 热度图时空信息的基础上,利用类 FCN 结构,全卷积操作预测一幅锚框掩码图片,用于 RPN 的定位回归。

1.3　孪生 RPN

孪生 RPN 网络借鉴了用于目标跟踪的全卷积孪生网络思想,将 RPN[20]与 FCN 结合,不仅实现了有效的目标跟踪,而且因 RPN 的回归能力大幅提高了目标位置的检测精度,其结构如图 3 所示。该方法检测帧率可达到 160 frame/s,在 VOT2015/2016/2017 及 OTB100 数据集中的各项指标都十分优异。但是该网络进行目标跟踪是基于上帧图片位置进行截图操作,不能对全图进行检测,无法有效解决目标丢失的问题。

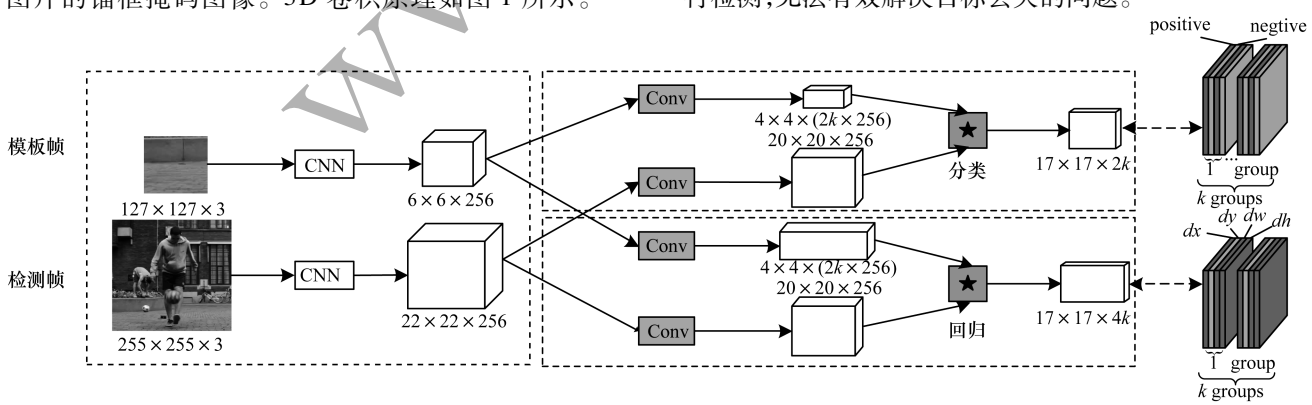


图 3　孪生 RPN 结构

在孪生 RPN 网络的基础上,本文重构网络,将整张图片作为输入,在全图范围内使用多尺度的模板图片生成不同大小的卷积核对跟踪图片进行卷积,最终根据生成结果,共同进行非极大抑制操作,寻找图片目标,并在此过程中引入锚框掩码机制,推测目标锚框位置,加快卷积操作及屏蔽不可能锚框位置,去除图片中相似物体干扰。

2 锚框掩码孪生 RPN

孪生 RPN 网络在检测当前帧图片目标时,是基于上一帧目标位置,截取一定大小的图片作为检测图片,即假设上下帧视频图像目标位置变化不大。这样的假设条件,对于高帧率视频及没有目标消失和遮挡的视频十分适用,但容错率较低。为此,本文提出一种锚框掩码孪生 RPN 网络。

2.1 锚框掩码

RPN 是 Faster-RCNN 中生成目标推荐框的核心机制,可以有效地生成推荐框,以及对前后景目标进行检测。根据其工作流程,本文提出的锚框掩码原理如图 4 所示。正常的 RPN 在特征图的每个锚点上会对应生成 k 个锚框,但是根据 IOU 热度图可知,并不是所有锚框均有效。因此,可根据先验知识,将生成的锚框掩码图与特征图点乘为一个稀疏矩阵,则其卷积对应的得分和回归框会出现大量的 0,从而屏蔽掉很多干扰锚框,加快卷积操作以及后期非极大抑制过程。本文提出的锚框掩码网络借助 3D 卷积学习连续3帧图片的IOU热度图信息,与后续的类 FCN 预测层进行相连,估计下一帧图片的锚框掩码图。

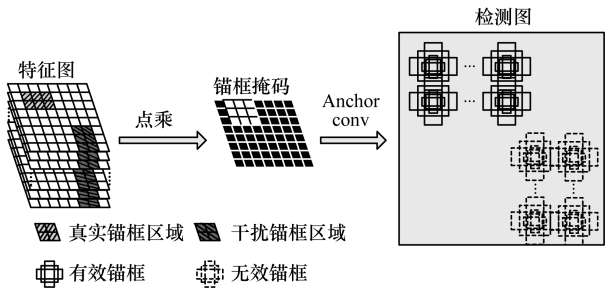


图 4 锚框掩码原理

2.2 锚框掩码网络

本文用一个考虑了目标时序特征的锚框掩码网络生成锚框掩码图,以屏蔽非追踪目标以及其他相似目标干扰。首先用锚框生成前 3 帧图像的 IOU 热度图(在训练时利用真实框,在使用时将前一帧检测得分最高的框作为真实框进行操作。当连续帧数小于 3 时,热度图设置为全 0 矩阵)。生成条件为每个锚点取 k 个锚框中的 IOU 最高值,如果 IOU 值小于有效阈值,则该位置设为 0,生成本图的 IOU 热度图。将前 3 帧图片的 IOU 热度图输入图 5 所示的锚框掩码网络,经 3D 卷积层综合时空特征后,进行全卷积操作,生成等大小锚框掩码预测图。将每个像素点分为 0、1 两类,表示是否屏蔽。生成结果综合前 3 帧图片的时空信息,对目标锚框可能的生成位置进行预测。

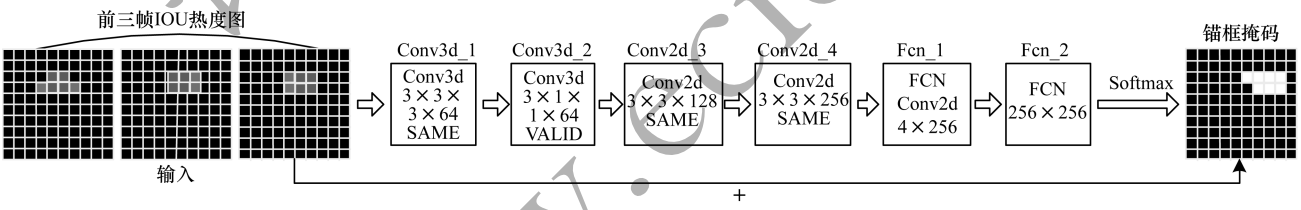


图 5 锚框掩码网络结构

为了减少因错误预测导致目标位置被屏蔽,本文将预测结果与前帧图片的 IOU 热度图转换结果进行叠加,提高连续图片的位置相关性。该网络在 VOT2016 数据集下的跟踪效果如图 6 所示。

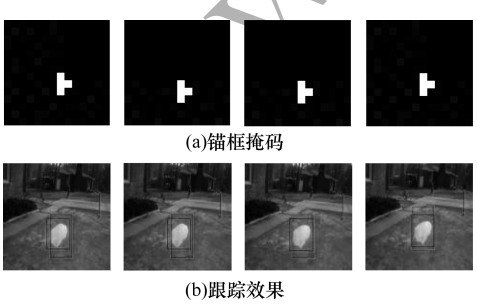


图 6 VOT2016 锚框掩码和跟踪效果

2.3 锚框掩码孪生 RPN

如图 7 所示,在本文锚框掩码孪生 RPN 结构中将全图调整为448 ×448,将模板区域按相应比例截取后调整为 224 ×224、128 ×128、64 ×64,输入网络进行特征提取。之后将相应特征图再进行卷积尺度扩充为 $2k$、$4k$ 两种大小,与经锚框掩码网络操作后的预测图片特征图进行全卷积操作,生成 3 种尺度锚框特征图,分别代表不同位置和大小的锚框,最后效仿 SSD 网络进行非极大抑制,得到得分最高的预测区域。通过设置不同尺度的模板图片与检测图片进行滤波操作,可以实现孪生 RPN 网络的变尺度检测。对全图片进行操作能避免因目标从屏幕中消失

或被遮挡而导致的跟踪失败。网络中的锚框掩码图在上帧图像跟踪到目标的条件下,接收锚框掩码网络生成结果,否则锚框掩码重置为全 1 矩阵,开启类全图检测防止目标跟踪丢失。

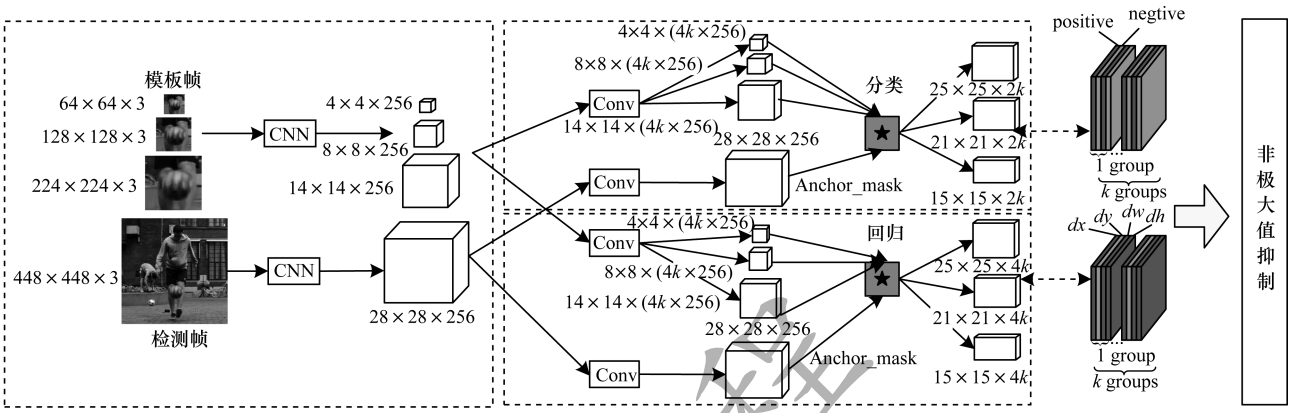


图 7　锚框掩码孪生 RPN 结构

2.4　损失函数

本文使用的锚框掩码网络和孪生 RPN 可以进行分离训练和线上组合,即根据相同的数据集分离训练锚框掩码和孪生 RPN。在训练孪生 RPN 时,将锚框掩码图置为全 1 矩阵。锚框掩码网络损失函数为:

$$loss_{\text{mask}} = \frac{-\sum_{i}^{M} label\_p_i \cdot \lg(score\_p_i)}{M} \tag{1}$$

其中,M 是卷积操作前特征图锚点数目,$label\_p_i$ 是真实框锚框掩码,$scores\_p_i$ 是锚框掩码图上每一个像素点的置信得分,$i = 1,2,\cdots,M$。

孪生网络损失函数为:

$$loss_{\text{sia}} = \sum_{i}^{3} loss_{sia\_cls_i} + \sum_{i}^{3} loss_{sia\_reg_i} \tag{2}$$

$$loss_{\text{sia\_cls}} = \frac{-\sum_{i}^{n} labels_i \cdot \lg(socres_i)}{n} \tag{3}$$

$$loss_{\text{sia\_reg}} = \frac{\sum_{i}^{n}\sum_{j}^{4} R(t_j - t_j^*)}{n} \tag{4}$$

其中,$loss_{sia}$ 表示孪生网络损失,$loss_{sia\_cls_i}$、$loss_{sia\_reg_i}$ 表示第 i 个模板图片处理后的网络分类损失和网络回归损失,$i = 1,2,3$,n 表示有效的锚框数目,$labels_i$ 表示第 i 个锚框的真实分类,t_j 是真实框位置和锚框位置转换参数(t_x,t_y,t_h,t_w)中的一项,t_j^* 是预测框位置和锚框位置转换参数(t_x^*,t_y^*,t_h^*,t_w^*)中的一项,$j = 1,2,3,4$。

3　实验与结果分析

3.1　实验环境及训练集

本文实验的硬件环境为锐龙 2600x 型 CPU、gtx1080Ti 型显卡;软件环境为 Linux 环境搭载 pycharm + tensorflow。训练集为 VOT2016 和 OTB100 数据集。其中,VOT2016 包含 60 个通用跟踪视频,OTB100 包含 100 个通用跟踪视频。

3.2　对比模型

本文实验与当前的 9 种主流模型进行对比,其中包括 SiamRPN[11]、C-COT[4]、T-CNN[5]、ECO-HC[8]、Staple[1]、EBT[21]、SiamRN[22]、MDNet-N[23] 和 SiamAN[22]。

3.3　数据集 VOT2016 实验结果

本文在数据集某一个视频片段中任取一帧图片真实位置作为模板图片,另一帧图片作为检测图片,生成一个训练样本,以 20 个训练样本作为一个批次进行训练,训练迭代 10 000 次。

本文通过准确率(跟踪成功的平均覆盖率)和鲁棒性(平均失败次数)对模型进行评估。具体指标为预期平均重叠率(Expected Average Overlap,EAO)、准确率(Accuracy)、平均批次失败个数(Failure)、等效过滤操作(Equivalent Filter Operations,EFO)。

本文实验与当前的 9 种主流模型进行了对比,结果如图 8 及表 1 所示。根据对比结果可以看出,本文模型相较于孪生 RPN(SiamRPN),在准确率略有下降的情况下,失误率减低了很多,所以在主流网络中其 EAO 表现优异。并在相同指标下跟踪速度 EFO 达到 15.4,即对大小为 448 ×448 图片,其跟踪帧率为 24.6 frame/s,可满足实时性要求。本文模型准确率下降的原因在于模板尺度变换有限,对于极小目标,模板难以匹配或者插值后特征不明显,但对于大部分跟踪目标本模型效果显著。

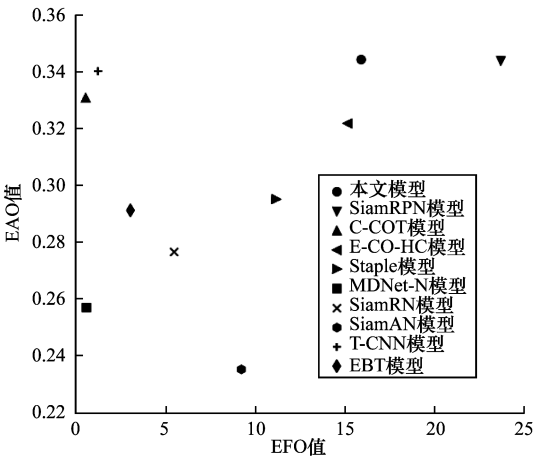


图 8 平均覆盖率及速度对比

表 1 10 种模型对比结果

模型	EAO	Accuracy	Failure	EFO
SiamRPN	0.344 1	0.56	1.08	23.300
C-COT	0.331 0	0.53	0.85	0.507
TCNN	0.327 5	0.55	0.89	1.200
ECO-HC	0.322 0	0.53	1.08	15.130
Staple	0.295 2	0.54	1.35	11.140
EBT	0.291 3	0.47	0.90	3.011
SiamRN	0.276 6	0.55	1.37	5.440
MDNet-N	0.257 0	0.54	1.20	0.534
SiamAN	0.235 2	0.53	1.65	9.210
本文模型	0.344 5	0.55	0.95	15.400

3.4 数据集 OTB100 实验结果

OTB100 数据集采用与 VOT2016 相同训练方式,以 20 个训练样本作为一个批次进行训练,训练迭代 10 000 次。

本文将预测框与真实框测距小于规定阈值的图片作为正确检测图片,正确图片在数据集中的占比作为准确率;预测框与真实框覆盖率高于设定阈值的图片比例作为成功率,对模型进行验证。实验结果如图 9 所示,其中,[ ]中的数值为该模型误差范围小于 0.5 的准确率值和覆盖率大于 0.5 的成功率值。从图 9 可以看出,本文模型在跟踪准确率和中心位置准确率上与孪生 RPN 结果相比皆有一定上升,准确率达到了0.862,成功率达到了0.642。并且本文模型成功率下降梯度小,准确率上升梯度大。这表明模型具有强鲁棒性,适用于单目标检测。

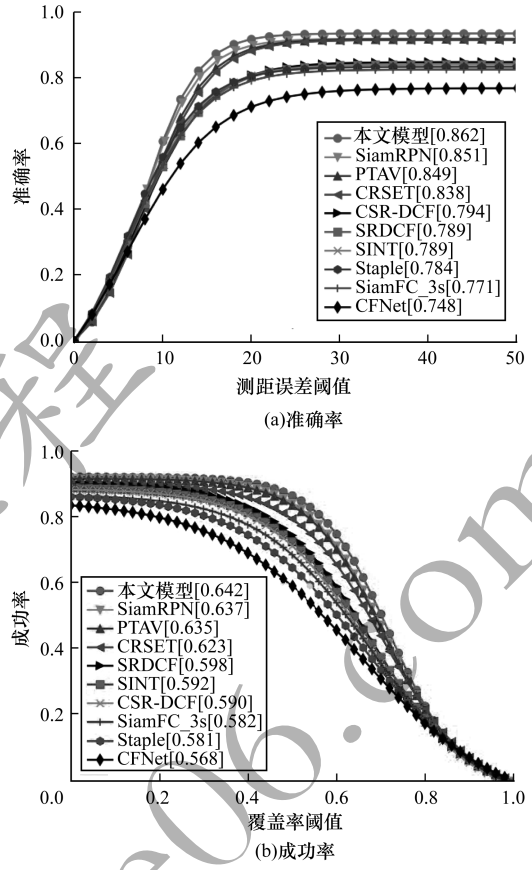
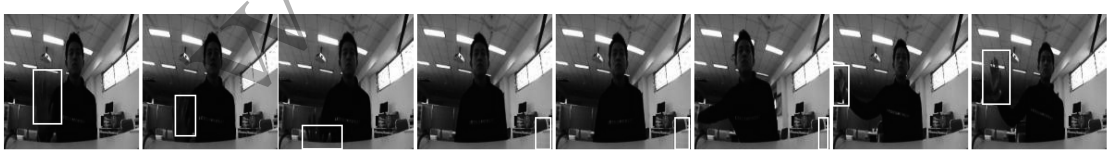


图 9 OTB 数据集下准确率及成功率对比

3.5 摄像头数据检验

本文利用摄像头采取手部数据训练,在实验室采集视频输入网络中进行检验,检验效果如图 10 所示。在图 10(a)中,目标在前 3 帧图像中出现,从第 4 帧开始目标在屏幕中消失,而在最后 2 帧图像中另一位置目标重新出现并被跟踪。这说明本文网络在目标丢失后可以重新跟踪目标,具有较强的鲁棒性。在图 10(b)中,具有相似特征的左右手轨迹具有交叉相近位置,目标跟踪没有受到干扰。这证明本文模型能对目标位置进行连续有效预测,屏蔽干扰位置,使得网络具有较强的抗干扰能力。上述实验证明,本文模型具有良好的跟踪特性可用于实时跟踪。



(a)目标消失视频

(b)目标干扰视频

图 10 实验室视频结果

4　结束语

本文构建一种用于单目标跟踪的锚框掩码孪生 RPN 模型,通过设置多尺度模板对全图进行相关滤波,以实现变尺度跟踪及防止目标丢失,同时采用锚框掩码网络预测目标锚框位置,提高网络的抗干扰能力。经过 VOT2016 和 OBT100 数据集实验检测,该模型具有高检测准确率及强鲁棒性。利用摄像头采集数据对模型抗干扰性与实时性的验证也取得了较好的效果。下一步尝试引入不同机制以提高目标的跟踪速度,并将本文方法应用到多目标跟踪领域。

参考文献

[1] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple:complementary learners for real-time tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016:1401-1409.

[2] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016:4293-4302.

[3] 江维创,张俊为,桂江生. 基于改进核相关滤波器的目标跟踪算法[J]. 计算机工程,2018,44(11):228-233.

[4] 李大湘,吴玲凤,李娜,等. 改进的 SAMF 目标跟踪算法[J]. 计算机工程,2019,45(2):258-264.

[5] DANELLJAN M, ROBINSON A, KHAN F S, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking [C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016:472-488.

[6] KANG Kai, LI Hongsheng, YAN Junjie, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10):2896-2907.

[7] ZHU Gao, PORIKLI F, LI Hongdong. Robust visual tracking with deep convolutional neural network based object proposals on PETS[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington D. C., USA: IEEE Press, 2016: 26-33.

[8] DANELLJAN M, BHAT G, SHAHBAZ K F, et al. ECO: efficient convolution operators for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017:6638-6646.

[9] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3):583-596.

[10] 朱齐丹,韩瑜,蔡成涛. 全景视觉非线性核相关滤波目标跟踪技术[J]. 哈尔滨工程大学学报,2018,39(7): 102-108.

[11] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with Siamese region proposal network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018:8971-8980.

[12] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]// Proceedings of European Conferenceon Computer Vision. Berlin, Germany: Springer, 2016:850-865.

[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015:3431-3440.

[14] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1):221-231.

[15] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the IEEE International Conferenceon Computer Vision. Washington D. C., USA: IEEE Press, 2015:4489-4497.

[16] DAI Jifeng, LI Yi, HE Kaiming, et al. R-FCN: object detection via region-based fully convolutional networks [C]// Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems Foundation, Inc., 2016:379-387.

[17] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of the IEEE International Conferenceon Computer Vision. Washington D. C., USA: IEEE Press, 2017:2961-2969.

[18] KRAHENBÜHL P, KOLTUN V. Efficient inference in fully connected CRFs with Gaussian edge potentials [C]// Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems Foundation, Inc., 2011:109-117.

[19] ZHENG Shuai, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional random fields as recurrent neural networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015:1529-1537.

[20] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems Foundation, Inc., 2015: 91-99.

[21] ZHU Gao, PORIKLI F, LI Hongdong. Beyond local search: Tracking objects everywhere with instance-specific proposals [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016:943-951.

[22] ZAGORUYKO S, KOMODAKIS N. Learning to compare image patches via convolutional neural networks [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015:4353-4361.

[23] FAN Heng, LING Haibin. Parallel tracking and verifying: a framework for real-time and high accuracy visual tracking[C]// Proceedings of the IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017:5486-5494.

编辑　刘盛龄