控制与决策
Control and Decision

基于负样本挖掘与特征融合的高速跟踪算法

李虹瑾, 彭力

引用本文:
李虹瑾,彭力. 基于负样本挖掘与特征融合的高速跟踪算法[J]. 控制与决策, 2023, 38(9): 2554−2562.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.2270

基于负样本挖掘与特征融合的高速跟踪算法

李虹瑾[1]，彭 力[1,2†]

(1. 江南大学 物联网技术应用教育部工程研究中心，江苏 无锡 214122；
2. 无锡太湖学院 物联网应用技术重点建设实验室，江苏 无锡 214064)

摘 要: 随着目标跟踪技术在多种视觉任务中的广泛应用,跟踪算法的实时性变得越来越重要. 全卷积孪生网络跟踪算法(SiamFC)虽然在跟踪速度方面较为理想,但在复杂的跟踪环境下很容易出现跟踪漂移. 为了能在提高算法精度的同时保证实时性,提出一种基于负样本挖掘与特征融合的高速跟踪算法. 首先,为了学到更深层次特征,又不过多增加额外参数运算,使用增加了剪裁层的轻量级网络ShuffleNetV2进行特征提取,提升跟踪速度;其次,在离线训练阶段引入不同种类的负样本对,加强对语义信息的学习,从而提升模型的特征判别能力;最后,为了得到更高质量的响应图,提出一种多尺度特征融合策略,充分利用浅层与深层特征,提高跟踪精度. 在OTB100和VOT2018两个数据集上与其他跟踪算法进行对比实验,结果表明:所提出算法较基准算法SiamFC在各项指标上有大幅度提升,在两个数据集下分别收获8.3 %和7.9 %的增益;同时在NIVIDA GTX l070下的速度可达114 FPS.

关键词: 目标跟踪；孪生网络；负样本挖掘；特征融合；轻量型网络

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.2270

引用格式: 李虹瑾,彭力. 基于负样本挖掘与特征融合的高速跟踪算法[J]. 控制与决策, 2023, 38(9): 2554-2562.

High-speed tracking algorithm based on negative example mining and feature fusion

LI Hong-jin[1]，*PENG Li*[1,2†]

(1. Engineering Research Center of Internet of Things Technology Applications of Ministry of Education，Jiangnan University，Wuxi 214122，China；2. Internet of Things Application Technology Key Construction Laboratory, Wuxi Taihu College，Wuxi 214064，China)

Abstract: With the wide application of object tracking technology in a variety of vision tasks, the real-time requirements for tracking algorithms are becoming increasingly important. Although the fully-convolutional siamese network algorithm for object tracking (SiamFC) is ideal in tracking speed, it is prone to tracking drift when dealing with complex tracking environment. In order to simultaneously improve the tracking accuracy and speed, a high-speed tracking algorithm based on negative example mining and feature fusion is proposed. Firstly, the improved lightweight network ShuffleNetV2 with the added crop layer is used for feature extraction, which can learn deeper features without adding extra parameters and calculations. Then, different types of negative example pairs are introduced in the offline training phase to strengthen the learning of semantic information, aiming at improving the feature discrimination ability of the model. Finally, a multi-scale feature fusion strategy is adopted, which makes full use of shallow and deep features to obtain a higher quality response map and greatly improve tracking accuracy. The experiment results of OTB100 and VOT2018 datasets show that the proposed algorithm significantly outperforms the benchmark algorithm SiamFC in various indicators, yielding 8.3 % gain in the OTB100 dataset and 7.9 % gain in the VOT2018 dataset. At the same time, the proposed tracker can perform 114FPS under NIVIDA GTX 1070.

Keywords: object tracking；siamese network；negative example mining；feature fusion；lightweight network

0 引 言

目标跟踪作为计算机视觉领域的一个基础研究内容,无论是在人机交互、无人驾驶还是现代军事等领域都有着重要应用. 近年来,随着深度学习的研究逐渐深入,诸多优秀的跟踪算法被相继提出,但这些算法普遍追求跟踪精度而忽略了速度. 然而在现实

收稿日期: 2021-12-29；**录用日期:** 2022-04-15.
基金项目: 国家自然科学基金项目(61873112)；台州市发改委基金项目(2106-331000-04-04-295510).
责任编委: 牛玉刚.
†通讯作者. E-mail: pengli@jiangnan.edu.cn.

场景应用中,跟踪算法的速度却是一项不可忽视的指标,许多优秀的跟踪算法落地困难也多源于此. 因此,本文研究设计一种目标跟踪算法旨在兼顾跟踪精度与速度,具有较强的现实意义.

目标跟踪技术经历了十几年的飞速发展,到现在主流算法基本分为两大类. 一类是相关滤波方法,其基本思想是设计一个滤波模板,利用该模板与目标候选区域做相关运算,最大输出响应的位置即为当前帧的目标位置. Bolme等[1]提出的MOSSE (minimum output sum of squared error)是相关滤波方法运用于目标跟踪领域的开山之作,在此基础上很多改进算法相继出现,其中以KCF (kernelized correlation filter)[2]为经典代表. KCF使用多通道HOG (histogram of oriented gradient)特征来代替单通道原始像素特征,大大提升了特征提取能力,同时利用循环矩阵可以被傅里叶矩阵对角化的性质,将矩阵的运算转化为元素的点乘,提高了运算速度. 这类跟踪方法虽然运行效率较高,但是应对复杂场景的效果不佳,跟踪精度受限. 另一类是基于深度学习孪生网络的方法,使用深度学习可以更好地提取目标特征,应对较大的目标变化和防止跟踪器漂移. Bertinetto等[3]提出的SiamFC (fully-convolutional siamese networks)奠定了孪生网络结构应用于跟踪领域的基础,该算法的孪生网络结构有两个输入,分别是作为基准的模板和待选择的搜索样本,孪生网络要在第1帧之后的每一帧中找出与第1帧目标对象最相似的候选区域,作为当前帧中的目标. 该文章发表之后,涌现出了相当多的跟进工作. Zhang等[4]提出了SiamDW (deeper and wider siamese networks for real-time visual tracking)算法,使用CIResNet替换SiamFC中的主干网络AlexNet[5],研究了深层网络应用在跟踪问题的退化现象,并给出了剪裁方案. Li等[6]提出了SiamRPN (high performance visual tracking with siamese region proposal network)算法,通过引入区域候选网络,使孪生网络提取出的目标特征在区域候选网络中进行分类和回归. DaSiamRPN (distractor-aware siamese networks for visual object tracking)算法[7]设计了抗干扰感知模块,提出在训练阶段改善样本的质量,进一步提升了模型的泛化能力. 文献[8-9]亦是基于孪生网络跟踪的相关跟进工作,这些跟进算法的跟踪精度指标都得到了显著提升. 但是,由于深度网络结构愈发复杂,模型的参数量和计算量激增,在一定程度上是以牺牲跟踪速度来换取精度的.

针对上述问题,本文以SiamFC为基准算法,提出一种基于负样本挖掘和多尺度特征融合的轻量级孪生网络高速跟踪算法. 为了能够兼顾跟踪精度与速度,做出以下3项改进: 1)将原有的骨干网络替换为轻量级网络ShuffleNetV2[10]进行特征提取,同时对网络结构进行调整,使其能够适用跟踪问题; 2)为进一步提升模型的特征判别能力,在训练阶段改善样本质量,通过生成不同的困难样本,解决样本分布不均的问题; 3)设计一种多尺度融合策略,高效融合结构特征与语义特征,并生成更高质量的响应图. 将所提出算法在OTB100和VOT2018数据集上进行测试,实验结果表明,相较于基准算法SiamFC,本文算法在跟踪精度与速度方面都有显著增益.

1　SiamFC跟踪框架

SiamFC算法使用孪生网络结构,具有两个权值共享的分支分别对应于输入模板图像和搜索图像,通过匹配计算特征相似度,确定目标的位置. 此外,使用端到端的离线训练,在帧率上满足了实时性的需求,故本文沿用该算法框架.

SiamFC算法中模板图像z和搜索图像x进入相同的卷积网络进行特征提取,然后使用互相关对两者进行相似度计算,即

$$f_\theta(z,x)=\varphi_\theta(z)*\varphi_\theta(x)+b\cdot 1. \tag{1}$$

其中: $f_\theta(z,x)$表示输入图像对的相似程度,φ_θ是通过参数为θ的神经网络后生成的特征图,$*$表示互相关运算,$b\cdot 1$表示相似度的偏置项. SiamFC算法在跟踪时以第1帧为模板图像送入神经网络提取特征,在之后的每一帧都以上一帧目标位置为中心进行填充得到搜索图像,进而以模板特征图作为卷积核在搜索特征图上进行卷积操作,计算出得分响应图,响应峰值处即为目标位置.

SiamFC算法通过端到端的离线训练进行相似度学习,将响应图上每一个点的损失定义为

$$l(y,v)=\log\left(1+\exp(-yv)\right). \tag{2}$$

其中: y为样本实际标签值,v为算法预测得到的相似度值. 于是在整个搜索区域形成了一张相似度响应图D. 将整体损失函数定义为所有响应图的均值,则有

$$L(y,v)=\frac{1}{D}\sum_{u\in D}\left(l(y(u),v(u))\right). \tag{3}$$

其中: u为相似度响应图上各个位置点; $v(u)$为算法预测该点的相似度值; $y(u)$为该点的真实标签值,定义如下:

$$y(u)=\begin{cases}+1, & \|u-c\|\leqslant R;\\ -1, & \text{otherwise.}\end{cases} \tag{4}$$

如果相似度与响应图的中心距离不超过R,则定义为正样本,用$+1$表示;剩余位置都认为是不相似,用-1表示.

2 本文跟踪算法

本文以SiamFC算法为基础,引入轻量级网络和多尺度特征融合模块,提升特征提取能力,同时,在训练阶段进行负样本挖掘,改善样本质量,整体框架如图1所示. 算法首先使用轻量型网络ShuffleNetV2作为主干网络提取特征,为了减轻多次填充(padding)操作对目标位置的影响,对卷积操作生成的特征图最外圈特征进行裁剪,从而减少网络层数,优化网络结构;其次,利用多尺度特征融合模块,对主干网络提取到的浅层与深层特征生成的响应图进行融合,从多个层级对目标进行表达;最后,在训练阶段构造负样本对,增强模型的判别能力.

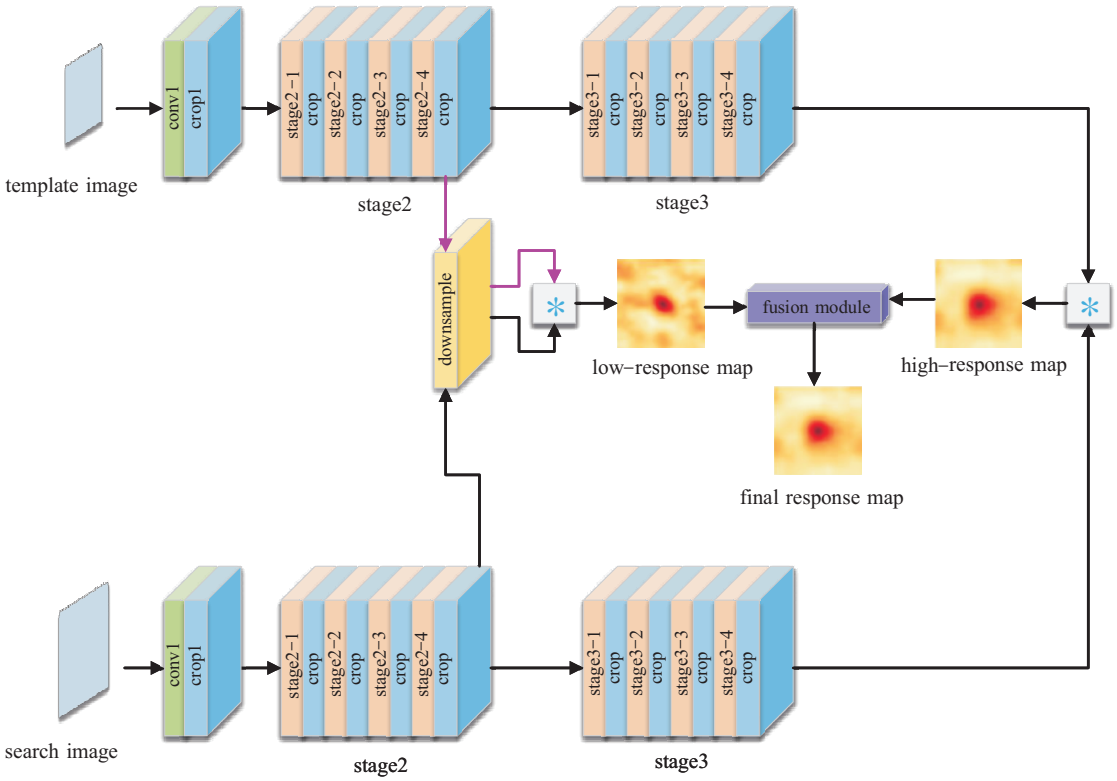


图 1 本文算法框架

2.1 改进的ShuffleNetV2

SiamFC算法的原始骨干网络AlexNet只有5个卷积层,虽然带来了可观的跟踪速度,但是对特征的提取能力有限,算法的跟踪精度并不理想. 因此,本文选择将骨干网络替换为比AlexNet参数量更小但卷积层数更多的轻量级网络ShuffleNetV2.

ShuffleNetV2是ShuffleNetV1[11]的升级版,它运用深度可分离卷积替换标准卷积来减少运算量. 深度可分离卷积可分解为逐通道卷积和逐点卷积两个过程. 逐通道卷积会对每个输入通道单独进行卷积操作,而逐点卷积就是用1×1的卷积核对逐通道卷积的输出结果进行卷积运算,整个过程的计算量和一次标准卷积的计算量分别为

$$\begin{aligned}\mathrm{MAC}_{\mathrm{dwconv}}=&D_K\times D_K\times M\times D_F\times D_F+\\ &M\times N\times D_F\times D_F,\end{aligned} \tag{5}$$

$$\mathrm{MAC}_{\mathrm{conv}}=D_K\times D_K\times M\times N\times D_F\times D_F. \tag{6}$$

其中:D_K和D_F分别是卷积核和输出特征的大小,M和N分别是输入输出的通道数. 将两个公式作比可知,卷积核越小,深度可分离卷积降低计算量的效果越显著.

本文通过实验研究发现,直接用ShuffleNetV2替换AlexNet作为骨干网络无法带来显著增益的原因有两个. 一是因为填充操作对目标位置产生偏差,如图2所示. 神经网络使用卷积提取特征时,往往会采用padding,即在输入特征图四周用零填充,使得输出特征图尺寸与输入一致. 与此同时带来的负面影响是模板特征也会包含这一填充信息,如图2(a)外围的灰色区域. 而在搜索图像中有些搜索区域只包含图像本身信息,如图2(b)左侧蓝色部分,边缘处的搜索区域却同时包含了原始信息和填充信息,如图2(b)

右侧橙色部分. 随着网络的加深,每一次卷积填充都会引入新的噪声,导致最终相似度计算结果不再精准. 因此,本文设计一种裁剪方法,在保证特征图尺寸合适的条件下,通过裁去卷积后的特征图的最外围区域,减少填充操作导致的偏差影响.

(a) 模板图像　　(b) 搜索图像

图 2　填充造成的位置偏差

限制增益的另一个关键因素是网络步长. 随着网络步长的增大,一方面最终的特征图尺寸变小,导致空间信息模糊; 另一方面网络感受野也会变大,这意味着提取到的特征难以精准反映目标位置信息. 而跟踪任务的关键就是需要获取更多的空间信息来对目标进行精细定位, ShuffleNetV2 原始总步长有 32,步长过大并不适用于目标跟踪. 考虑到步长过小也会出现计算负担变重、跟踪速度下降的问题,通过权衡跟踪的实时性与精确性,最终将网络总步长控制为 8. 具体修改方式是将原始网络中 stage 3 中瓶颈层的循环次数由 8 次减为 4 次, 去掉全部 stage 4 层以及第 1 次卷积后的最大池化层. 除此之外,在网络的最后增加了 Convlast 层,用于控制最终特征图输出通道数为 256, 在 stage 2 层之后增加 Convdown 层,进行下采样操作以便于之后进行特征融合,具体的网络结构见图 1.

2.2　负样本挖掘

一个模型的性能除了与网络结构本身有关,还非常依赖于具体的训练策略,比如数据增强. 本文研究发现,阻碍目标进一步表征学习的主要原因是训练数据中非语义背景和语义干扰的不平衡. 高质量的训练数据是端到端离线学习跟踪器的成功关键,网络表征的质量好坏在很大程度上取决于训练数据的分布. 然而, SiamFC 算法在训练阶段主要使用的仍是易于背景分类的样本,从而导致跟踪效率难以提升. 因此,本算法在训练阶段会生成不同语义的负样本对,以此提高模型对于特征的判别能力,抵消数据分布不平衡带来的影响.

SiamFC 算法的判别表征能力不佳源于两个层面的训练数据分布不均衡: 一方面是语义负样本稀缺造成的不平衡; 另一方面是同类干扰的不平衡,也就是目标跟踪中的困难样本. SiamFC 的训练数据主要是背景,这意味着大多数的负样本都是非语义的,即它们都不是真实的对象,所以很容易进行区分. 因此,网络仅仅能学习到前景与背景的区别,而语义对象之间的差异将被大量简单负样本所覆盖. 本文算法在训练阶段引入负样本对,该负样本对分别由来自相同类别和不同类别的目标组成,标签定义如下:

$$y(u)=\begin{cases}0.2, \ \text{samecategories};\\ 0, \ \text{differentcategories}.\end{cases} \tag{7}$$

图 3 展示了本算法构造的负样本对. 来自相同类别的负样本对能够帮助跟踪器着重细粒度表征,而不同类别的负样本对可以减少因目标遮挡或者出视野而引发跟踪器发生漂移的现象.

(a) 来自相同类别的负样本对　　(b) 来自不同类别的负样本对

图 3　负样本对

对比来看, SiamFC 的训练策略是所有训练样本对中的模板图像和搜索图像皆来自同一个图片序列,它的训练过程是重点强调模板与候选区域的相似性匹配程度,所以 SiamFC 算法在训练时学习到的是同类间相似度. 而本文所提出的负样本挖掘策略,从不同的图片序列中构建模板与候选图片对,在训练时不仅学习了同类间的相似度,还学习到不同类间的不相似度,使得模型在线跟踪时,可以明确地将目标与周围不同种类的显著干扰物体区分开来,从而提升了模型的判别能力.

2.3　多尺度特征融合策略

特征融合就是将来自不同层次或分支的特征进行组合,是现代网络架构中广泛应用的一部分. FPN (feature pyramid networks for object detection)[12] 通过构建特征金字塔网络,在目标检测任务上展现出了很好的效果提升. 文献 [13] 也提出了一种注意力融合方式,解决了融合语义与尺度不一致的特征时出现的问题. 低层特征的分辨率更高,包含着丰富的位置细节信息,有利于对目标精准定位,但是,经过的卷积较少就会残留更多的噪声; 而高层特征蕴含了更强的语义信息,能有效应对目标变化,但是由于分辨率

变低,对细节的敏感度也受到了影响. 因此,本算法提出一种多尺度特征融合策略,有效融合浅层和深层特征,进一步提升跟踪性能. 使用孪生网络框架进行目标跟踪的关键环节是通过计算相似度得到响应图,再根据响应图峰值处来判断目标位置. SiamFC原始骨干网络仅有5层卷积,蕴含的语义信息十分有限,并不能有效地完成特征融合;而所提出的跟踪器已将骨干网络替换为ShuffleNetV2,可以充分发挥深层网络的优势,采取多尺度特征融合策略,从多角度表征目标.

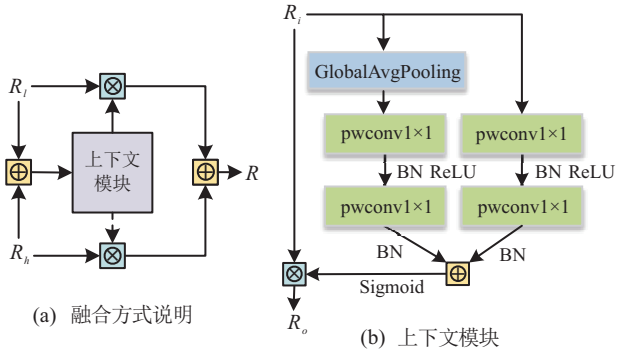
如图1所示,对stage 2层最后的输出特征经过的卷积进行简单的下采样操作,利用采样后的特征以及网络最终输出特征对模板图像与搜索图像进行相似度计算,生成两张大小为17 × 17的响应图,再对两者进行融合得到最终的响应图. 图4(a)展示了响应图融合方式,浅层与深层响应图进行相加后会通过一个上下文感知模块学习到不同响应图的权重,最后,根据融合权重线性加权得到最终的响应图,融合过程如下:

$$R = C(R_l \oplus R_h) \otimes R_l + (1 - C(R_l \oplus R_h)) \otimes R_h. \tag{8}$$

其中:$\oplus$和$\otimes$分别表示广播加法和按元素进行相乘;R_l和R_h分别表示浅层与深层响应图,在所提出算法中分别对应stage 2层与最终层得到的响应图;R表示融合后的响应图;C代表上下文感知模块,具体结构如图4(b)所示. 为了尽可能保持跟踪模型的轻量,仅使用两个分支来提取融合权重,左侧分支使用全局平均池化来提取全局特征,右侧分支使用逐点卷积提取局部特征,两者相加后通过激活函数获得权重,即

$$R_o = R_i \otimes \sigma(l(R_i) \oplus g(R_i)). \tag{9}$$

其中:R_i和R_o分别表示输入与输出的响应图,$l(R_i)$代表局部特征提取,$g(R_i)$代表全局特征提取.



(a) 融合方式说明 (b) 上下文模块

图4 响应图融合

相较于普通融合操作,如求和或连接,所提出算法采用了上下文感知模块,能够充分利用上下文信息求得融合权重. 多尺度融合策略有效平衡了浅层空间信息与深层语义信息,实现了对目标多层面表征从而得到高质量响应图,进一步提高跟踪器的鲁棒性.

3 实验结果与分析

3.1 实验参数与环境配置

本文所提出的算法是在Pytorch 0.4.1深度学习框架下实现的,实验环境配置:CPU是英特尔i7-8700,操作系统为Ubuntu16.04,16 G内存,显卡是NIVIDA GeForce GTX1070,8 G显存.

在训练阶段,使用GOT-10 K数据集以及构造的负样本对进行端到端的离线训练,GOT-10 K数据集约有10 000个图片对,构造负样本对约2 000组. 实验加载ShuffleNetV2的预训练模型完成卷积层参数初始化,使用随机梯度下降方法对网络进行优化,动量控制为0.9,学习率由10^{-2}到10^{-5}呈指数衰减.

在测试阶段,孪生网络两分支分别输入的是大小为$127 \times 127 \times 3$的模板图像和大小为$255 \times 255 \times 3$的搜索图像,经相关计算后得到响应图大小为17 × 17,利用3个尺度下的搜索图像块分别进行相似度衡量,确定最佳尺度,缩放因子定为1.053.

3.2 基于OTB100实验分析

当前,单目标跟踪领域应用最广泛的数据集之一是OTB100,它包含了100组视频序列,具备光照变化(IV)、尺度变化(SV)、遮挡(OCC)、形变(DEF)、运动模糊(MB)、快动作(FM)、平面内旋转(IPR)、平面外旋转(OPR)、离开视野(OV)、背景复杂(BC)以及分辨率低(LR)共11项困难属性,采用一次评估欧氏距离精度图和重叠成功率图来评估跟踪器的性能. 欧氏距离精度图是根据中心位置的误差绘制的,它衡量的是跟踪算法预测的目标框与真实的目标框中心的欧氏距离,即

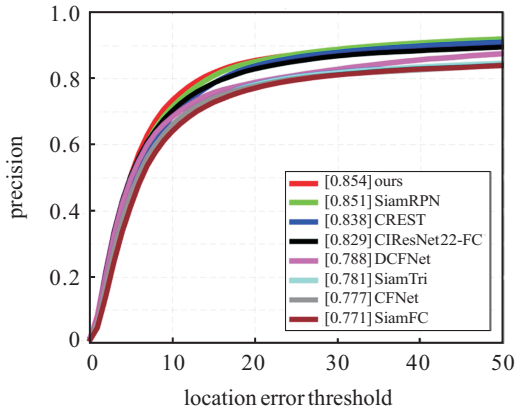
$$V_{\mathrm{CLE}} = \sqrt{(x_A - x_G)^2 + (y_A - y_G)^2}. \tag{10}$$

其中:(x_A, y_A)为算法预测的目标中心坐标,(x_G, y_G)为真实的目标中心坐标,V_{CLE}是两者的欧氏距离. 重叠成功率是指跟踪算法预测的目标框与真实的目标框之间的重叠率,也称为交并比,即

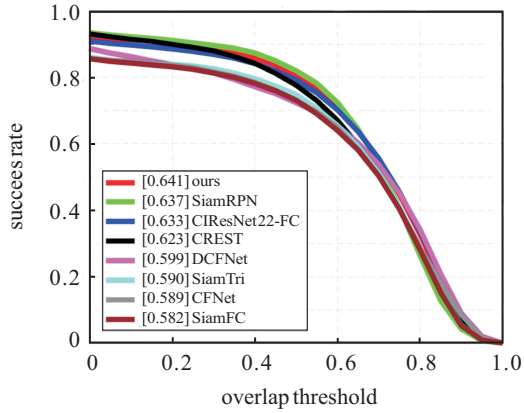
$$\mathrm{IoU} = \frac{B_A \bigcap B_G}{B_A \bigcup B_G}. \tag{11}$$

其中:B_A表示算法预测的目标框,B_G表示真实的目标框,$B_A \bigcap B_G$和$B_A \bigcup B_G$分别表示两者的交集和并集,IoU是最终的重叠率.

为验证所提出算法的有效性,与7种主流算法进行比较,测试结果如图5所示. 对比算法有SiamRPN、CIResNet22-FC、CREST[14]、DCFNet[15]、SiamTri[16]、



(a) 精度曲线



(b) 成功率曲线

图 5　本文算法与其他算法在OTB100上的总体精度和成功率对比

CFNet[17] 以及基准SiamFC. 在图5中,所提出算法在总体距离精度和重叠成功率上均表现最优,相比于基准SiamFC算法,更是在两项指标中分别获得了8.3 %和5.9 %的增益. 不同于SiamFC只用少量的数据集训练,SiamRPN使用大规模数据集进行网络训练以获得显著增益;而本文算法训练阶段仅在GOT-10 K数据集基础上引入少量负样本对,却在两项指标上都得到了可以比肩的结果. 同样,在SiamFC基础上改进的SiamDW (CIResNet 22-FC),虽然使用了较为深层的网络CIResNet 22提取特征,但是,所提出算法在改进ShuffleNet的基础上使用多尺度特征融合,得到了更高质量的响应图,因而收获了更高的跟踪精度.

表1展示了本文算法与其他对比算法在11项困难属性下的距离精度值,加粗表示在该项挑战下的最优值,显然,所提出算法在该评估指标中均处于领先地位. 在距离精度指标中,本文算法在遮挡、运动模糊、快动作、平面外旋转、离开视野以及背景复杂6项困难属性下位列第1;在尺度变化和形变2项困难属性中表现次优;尤其在快动作挑战下,本文算法的距离精度值高达0.842,超越第2名近4 %. 相比于使用相关滤波方法的CREST算法不进行离线训练,利用模板更新的方式来弥补训练数据的缺乏,本文算法通过在训练阶段构造不同种类负样本不仅增强了模型学习,同时避免了在线训练更新模板耗时长的问题,具备一定的发展潜力.

表 1　8种跟踪器在OTB100上所有困难属性下的距离精度值

algorithm	IV	SV	OCC	DEF	MB	FM	IPR	OPR	OV	BC	LR
ours	0.819	0.833	**0.816**	0.794	**0.857**	**0.842**	0.846	**0.857**	**0.787**	**0.840**	0.900
SiamRPN	0.864	**0.841**	0.785	**0.830**	0.821	0.793	**0.859**	0.855	0.728	0.803	**0.982**
CIResNet22-FC	0.795	0.818	0.800	0.765	0.841	0.808	0.823	0.830	0.780	0.762	0.902
CREST	**0.876**	0.786	0.786	0.776	0.813	0.792	0.853	0.842	0.734	0.829	0.866
DCFNet	0.752	0.758	0.778	0.730	0.715	0.745	0.783	0.764	0.697	0.734	0.832
SiamTri	0.746	0.748	0.726	0.680	0.727	0.763	0.774	0.763	0.723	0.715	0.900
CFNet	0.707	0.730	0.701	0.713	0.679	0.704	0.785	0.759	0.601	0.755	0.893
SiamFC	0.736	0.735	0.722	0.690	0.705	0.743	0.742	0.756	0.669	0.690	0.900

3.3　基于VOT2018实验分析

为进一步验证所提出跟踪器的性能,在数据集VOT2018上进行测试. VOT数据集包含60组视频序列,具备遮挡、光照变化、运动改变、尺寸改变以及相机运动共5项困难属性,利用平均期望重叠(expect average overlap rate, EAO)、准确率(accuracy, A)、鲁棒性(robustness, R)三项指标来衡量跟踪器的性能. 相较于OTB100使用矩形框来标定跟踪目标,VOT2018使用的是旋转矩形框,从而跟踪难度升级. 将本文算法与LWDNTthi[18]、CSRDCF[19]、DCFCF[18]、DSiam[20]、SiamFC、DCFNet以及DensSiam[21]算法进行对比,实验结果见表2.

表 2　8种算法在VOT2018上的对比结果

algorithm	A	R	EAO	FPS
ours	**0.525**	0.412	**0.266**	**114**
LWDNTthi	0.462	**0.332**	0.261	60
CSRDCF	0.491	0.356	0.256	13
DCFCF	0.485	0.342	0.247	–
DSiam	0.512	0.654	0.196	45
SiamFC	0.503	0.585	0.187	90
DCFNet	0.470	0.543	0.182	60
DensDiam	0.462	0.688	0.174	60

表 2 中, 所提出跟踪算法在平均期望重叠 (EAO) 和准确率 (A) 指标上表现最优, 相较于基准算法 SiamFC, 3 项指标都有大幅度提升. LWDNTthi 使用轻量型网络 ThiNet[22] 替换 SiamFC 的骨干网络, 精度得到大幅提升, 算法速度满足实时性, 但是依然不及所提出算法. DensSiam 同样是替换主干网络为 DenseNet[23], 而所提出算法使用轻量级 ShuffleNet, 具备更高精度的同时也满足了实时性, 相比于两种算法, EAO 分别提升了 0.5 % 和 9.2 %. CSRDCF 算法基于相关滤波方法, 引入空间信道置信度实现模板更新, 因此在鲁棒性指标上位居第 1, 但本文算法在精度方面还是略胜一筹. 对比同样使用多层深度特征融合的 DSiam 算法, 本文算法应用多尺度特征融合策略, 在鲁棒性指标上有 24 % 的提升, 且具备更高的精度.

当目标跟踪算法每秒可以给出 30 帧图片的跟踪结果即速度达到 30 FPS 时, 就能够称它满足实时性要求. 因为视频是由多张连续的图片构成, 而要使人眼观看流畅, 则须满足每秒的视频里至少要有 24 张图片, 所以如果算法能够在 1 s 内处理 30 张图片, 就能够对视频进行实时处理了. 而本文算法的跟踪速度可达 114 FPS, 远超实时性要求. 综上, 所提出跟踪器能够兼顾跟踪准确率和实时性, 在多种困难属性下完成对目标定位, 具有较大的发展前景.

图 6 展示了几种算法在 VOT2018 数据集上的一些序列跟踪的可视化结果. 第 1 个 fish 2 序列中, 背景颜色与目标对象颜色较为相近且存在同类干扰, 在第 27 帧时 SiamFC 与 DensSiam 算法都已跟丢目标, CSRDCF 虽然可以跟上目标, 但整个过程的尺度估计都偏大, 即使在第 239 帧时所有算法全部跟上目标的情况下, 本文算法预测的矩形框与真实标注的矩形框更为接近, 表明本文跟踪器对于目标中心的把握更佳. 第 2 个 motocross 1 序列中, 运动员在空中进行快速翻转运动, 图像出现一定程度的模糊, 在第 85 帧中 CSRDCF 和 DensSiam 算法预测的矩形框都出现了不同程度的漂移, 而所提出算法依然可以跟住目标, 相对于 SiamFC 的预测结果也更为精准. 在第 3 个 soldier 序列中, 士兵从草丛中爬出, 所戴头盔与背景十分相似都呈现出绿色, 在第 46 帧中 SiamFC 和 DensSiam 算法的预测结果都出现了轻微的偏移, 靠向左上角; 之后的第 74 帧中, 4 种算法都没能准确地框住目标, CSRDCF 算法甚至错误地将周围杂草视为

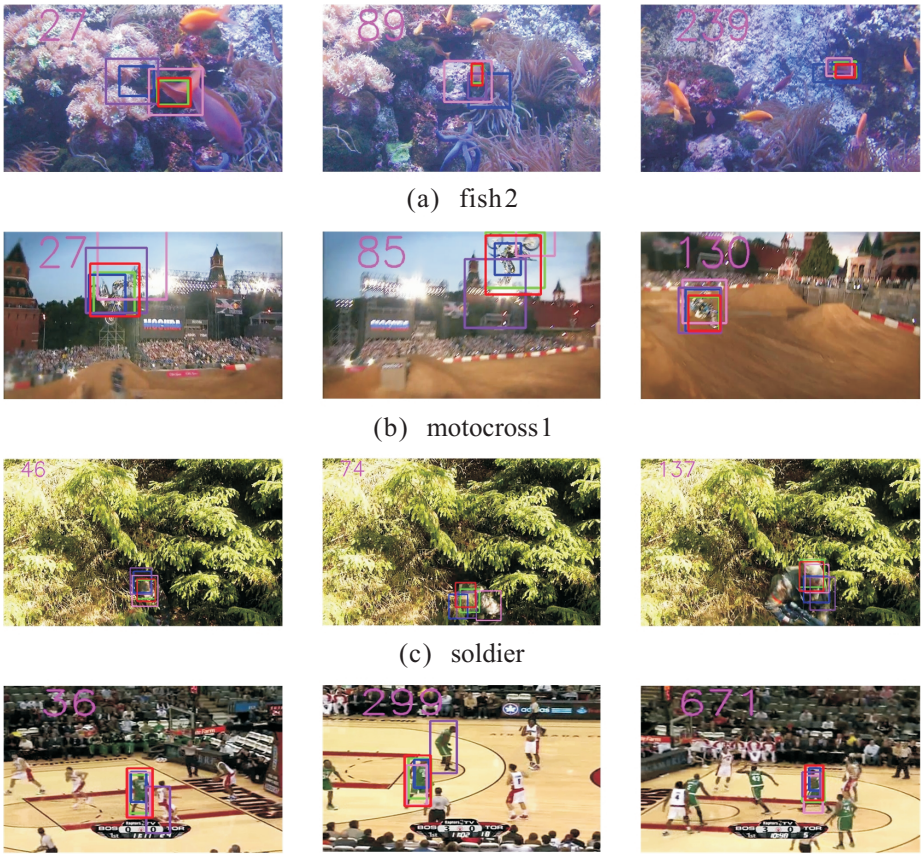


图 6 VOT2018 数据集上一些序列的跟踪结果

目标;到了第137帧,只有本文跟踪器成功锁定目标,其余3种算法的预测框都或多或少偏向右下方. 最后一个basketball序列中,运动员在场上打球进行快速运动,DensSiam算法在第36帧中的预测框已经严重偏离;到了第299帧中完全将目标错误预测为身穿相同球衣的干扰球员. 整个过程中只有本文算法和CSRDCF算法能够较为准确地跟住目标,SiamFC虽然没有跟丢,但是预测框都偏小,与真实结果存在一定差距. 即使在一定的困难因素下,所提出算法依然能够识别目标,优于其他对比算法,完成较为精准的目标跟踪.

3.4　消融研究与程度分析

为了验证本文算法所提出的3点改进策略是有效的,在OTB100数据集上针对距离精度(precision)、重叠成功率(AUC)以及速度(FPS)三项指标进行消融研究. 与基准算法SiamFC的对比结果见表3.

表3　本文算法与基础算法消融实验

algorithm	precision	AUC	FPS
ShuffleNetV2+Negative example+Fusion	0.854	0.641	114
ShuffleNetV2+Fusion	0.831	0.632	114
ShuffleNetV2+Negative example	0.842	0.628	134
ShuffleNetV2	0.827	0.621	135
SiamFC	0.771	0.582	90

将SiamFC骨干网络替换为轻量级ShuffleNetV2后,AUC指标由58.2%提升到62.1%,同时速度也得到显著增益. 在此基础上,训练阶段引入负样本对,或者单独采取融合策略,在距离精度指标上又分别收获了1.5%和0.4%的增益,在重叠成功率方面提升了0.7%和1.1%. 从速度角度上看,负样本挖掘几乎没有引入额外的开销却带来了不小的精度增益. 而在单独采用融合策略后,由于融合模块较为轻量,速度仅有些轻微下降,但实现的精度增长依然可观. 最终,将3点策略全部融入本文所提出的算法,在距离精度上得到了8.3%的提升且速度达到114 FPS,满足超越实时性要求.

本文所提出的跟踪器旨在兼顾跟踪的精确性和实时性,因此,继续对本文算法的轻量程度进行评估. 表4展示了与基准算法SiamFC在模型大小、主干网络参数量以及计算量3方面的对比结果. 模型大小是指通过深度网络训练得到最终模型需要占用的存储空间,参数量和计算量分别指深度网络中神经元的个数以及卷积层与归一化层的加法和乘法数目总和. 表4中,本文算法的模型大小只有SiamFC的四分之一不到,是HA-SiamVGG[24]的十分之一不到,而SiamFC的主干网络的参数量和计算量却是本文算法的近5倍,HA-SiamVGG则更多. 这也验证了本跟踪器主干网络的轻量型,体现了其在速度方面的优越性.

表4　3种模型大小、计算量及参数量比较

模型	模型大小/kB	参数量/M	计算量/M
ours	2 234	0.543 578	80.676
SiamFC	9 140	2.336 32	460.173
HA-SiamVGG	30 394	7.771 97	2 813.9

4　结　论

本文提出了一种基于负样本挖掘和多尺度特征融合的孪生网络高速跟踪算法. 该算法首先使用改进后的ShuffleNetV2作为主干网络提取更深层次的特征;然后在训练阶段构造来自相同与不同种类的负样本对,平衡数据分布,增强了模型的特征判别能力;最后引入多尺度特征融合策略,提高了最终响应图质量. 将所提出的算法在OTB100和VOT2018数据集上进行了实验,所得结果验证了该算法相比于其他算法具有更高的跟踪精度和速度.

参考文献(References)

[1] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, 2010: 2544-2550.

[2] 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, 2015, 37(3): 583-596.

[3] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 850-865.

[4] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 4586-4595.

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8971-8980.

[7] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese

networks for visual object tracking[C]. Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer, 2018: 101-117.

[8] 宋建辉, 张甲, 刘砚菊, 等. 基于条件对抗生成孪生网络的目标跟踪[J]. 控制与决策, 2021, 36(5): 1110-1118.
(Song J H, Zhang J, Liu Y J, et al. Conditional generative adversarial Siamese networks for object tracking[J]. Control and Decision, 2021, 36(5): 1110-1118.)

[9] 刘如浩, 张家想, 金辰曦, 等. 基于可变形卷积的孪生网络目标跟踪算法[J]. 控制与决策, 2022, 37(8): 2049-2055.
(Liu R H, Zhang J X, Jin C X, et al. Target tracking based on deformable convolution siamese network[J]. Control and Decision, 2022, 37(8): 2049-2055.)

[10] Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design[C]. European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 116-131.

[11] Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 6848-6856.

[12] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 936-944.

[13] Dai Y M, Gieseke F, Oehmcke S, et al. Attentional feature fusion[C]. 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, 2021: 3559-3568.

[14] Song Y B, Ma C, Gong L J, et al. CREST: Convolutional residual learning for visual tracking[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 2574-2583.

[15] Wang Q, Gao J, Xing J L, et al. DCFNet: Discriminant correlation filters network for visual tracking[J/OL]. 2017, arXiv: 1704.04057.

[16] Dong X, Shen J. Triplet loss in Siamese network for object tracking[C]. European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 459-474.

[17] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5000-5008.

[18] Kristan M, Matas J, Leonardis A, et al. The visual object tracking VOT2015 challenge results[C]. 2015 IEEE International Conference on Computer Vision Workshop. Santiago, 2015: 564-586.

[19] Lukeic A, Vojír T, Zajc L C, et al. Discriminative correlation filter with channel and spatial reliability[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4847-4856.

[20] Guo Q, Feng W, Zhou C, et al. Learning dynamic Siamese network for visual object tracking[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 1781-1789.

[21] Abdelpakey M H, Shehata M S, Mohamed M M. DensSiam: End-to-end densely-Siamese network with self-attention model for object tracking[C]. International Symposium on Visual Computing (ISVC). Berlin: Springer, 2018: 463-473.

[22] Luo J H, Wu J X, Lin W Y. ThiNet: A filter level pruning method for deep neural network compression[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 5068-5076.

[23] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 2261-2269.

[24] Zhang C Y, Wang H, Wen J W, et al. Deeper Siamese network with stronger feature representation for visual tracking[J]. IEEE Access, 2020, 8: 119094-119104.

作者简介

李虹瑾 (1997−), 女, 硕士生, 从事计算机视觉、目标跟踪等研究, E-mail: li.hongjin@foxmail.com;

彭力 (1967−), 男, 教授, 博士生导师, 从事视觉物联网、行为识别、深度学习等研究, E-mail: pengli@jiangnan.edu.cn.