基于多层深度卷积特征的抗遮挡实时跟踪算法

崔洲涓[1,2*]，安军社[1]，崔天舒[1,2]

[1]中国科学院国家空间科学中心复杂航天系统电子信息技术重点实验室，北京 100190；

[2]中国科学院大学，北京 100049

摘要 为提高复杂场景中目标跟踪算法的准确性与实时性，提出一种基于多层深度卷积特征的抗遮挡实时目标跟踪算法。针对目标跟踪任务，先对深度卷积网络 VGG-Net-19 进行微调，再提取目标区域的多层深度卷积特征。根据相关滤波框架构建位置相关滤波器，确定目标中心位置。设计尺度相关滤波器对目标区域进行不同尺度采样，确定目标尺度。目标遮挡时，采用阶段性评估策略进行模型更新与恢复，解决模型误差积累问题。选取目标跟踪评估数据集 OTB-2015(100 组视频序列)与 UAV123(123 组视频序列)进行测试。实验结果表明，本文算法具有更高的准确性，能够适应目标遮挡、外观变化及背景干扰等复杂情况，平均速度为 29.6 frame/s，满足目标跟踪任务的实时性要求。

关键词 机器视觉；目标跟踪；深度卷积特征；相关滤波；模型恢复

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/AOS201939.0715002

Real-Time and Anti-Occlusion Visual Tracking Algorithm Based on Multi-Layer Deep Convolutional Features

Cui Zhoujuan[1,2*], An Junshe[1], Cui Tianshu[1,2]

[1] *Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing* 100190, *China*;

[2] *University of Chinese Academy of Sciences, Beijing* 100049, *China*

Abstract In order to improve the accuracy and real-time performance of visual tracking in complex scenes, a real-time and anti-occlusion visual tracking algorithm based on multi-layer deep convolutional features is proposed. For the visual tracking task, the deep convolutional networks VGG-Net-19 are fine-tuned, and then the multi-layer deep convolutional features of the target region are extracted from the adjusted model. The location correlation filters are constructed to determine the target center position. In order to determine the target scale, a scale correlation filter is performed to sample multi-scale images surrounding the target region. When the target is occluded, the stage evaluation strategy is used to update and recover the model, which solves the problem of template error accumulation. The experimental results on the tracking benchmark OTB-2015 which concludes 100 video sequences and UAV123 which concludes 123 video sequences show that the proposed algorithm has higher accuracy and can adapt to complex situations such as target occlusion, appearance change and background clutters. The average speed is 29.6 frame/s, which meets the real-time requirements of the visual tracking task.

Key words machine vision; object tracking; deep convolutional features; correlation filters; model recovery

OCIS codes 150.0155; 150.1135; 100.4999

1 引 言

视觉目标跟踪是一个综合视觉特征提取、视觉信息分析、目标运动信息检测和识别等的交叉课题，是机器视觉领域一个重要的研究方向。随着目标跟踪理论研究的深入和计算机软硬件的发展，目标跟踪算法在商业和军事领域中的应用日益广泛。然而在实际应用中，设计一个可以处理好各种复杂多变场景的稳健算法依然具有很大的挑战。

近年来，源自信号处理理论的相关滤波视觉跟

收稿日期：2019-02-01；**修回日期**：2019-03-03；**录用日期**：2019-03-21

基金项目：中国科学院复杂航天系统电子信息技术重点实验室自主部署基金(Y42613A32S)

* **E-mail**：constance669@126.com

踪算法以优异的跟踪速度成为研究的热点方向。Bolme 等[1]将相关理论引入目标跟踪领域，设计了一个最小输出平方误差和(MOSSE)滤波器，在跟踪过程中通过提取图像灰度特征寻找目标最大响应值，实现了速度的飞跃。Henriques 等[2]提出的核循环结构算法(CSK)将训练阶段的密集采样问题转化为特征矩阵的循环移位运算。Danelljan 等[3]提出的颜色特征算法(CN)通过提取灰度特征与降维的颜色特征提升跟踪效果。Henriques 等[4]提出的核相关滤波算法(KCF)在 CSK 基础上，通过方向梯度直方图特征(HOG)将适用范围从灰度图扩大到多通道有色图，其跟踪速度达到 172 frame/s。Danelljan 等[5]针对 KCF 中利用循环矩阵求解损失函数时出现的边界效应问题，提出空间正则化算法(SRDCF)，通过在损失函数中引入惩罚项，抑制离中心较远的特征对跟踪算法的影响，进一步提高了跟踪精度。此外，国内一些研究人员在相关滤波框架上对多尺度适应问题进行了探索[6-9]。

随着深度学习方法在图像分类、目标检测等领域取得突破性进展，基于深度学习的跟踪算法也引起了广泛关注。Wang 等[10]提出的深度学习算法(DLT)，在大规模数据集上通过栈式降噪自编码器进行离线预训练得到通用物体表征能力，引用粒子滤波框架，对输入跟踪数据集的第 1 帧带标注的样本进行在线微调。Wang 等[11]又在 DLT 基础上改进，提出了结构化输出深度学习算法(SO-DLT)，跟踪时利用当前目标的有限样本信息对预训练卷积网络模型进行微调。Ma 等[12-13]将预训练深度网络中不同卷积层提取的特征与相关滤波的框架结合，提出分层卷积特征算法(HCFT)，经过优化更新得到稳健的分层卷积特征算法(HCFTstar)。Wang 等[14]提出全卷积网络算法(FCNT)，设计出特征筛选网络和互补的预测网络，提升了跟踪精度。Nam 等[15]使用大规模具有标注框的视频序列训练卷积网络得到通用的目标表观模型，提出多域网络(MDNet)结构，包括共享层及多分支的全连接层，解决了跟踪训练数据不足的问题。Bertinetto 等[16]引入基于对比性损失函数的孪生(Siamese)体系结构，训练一个完全端到端的跟踪模型，同时输入示例样本和候选样本，通过离线训练模型评估二者的相似程度，决策层选择适合的匹配算法计算相似度，匹配程度最高的候选样本作为目标当前最优区域。作者又在此基础上改进，得到了端到端相关滤波算法(CFNet)[17]。Tao 等[18]提出了一个基于 Siamese 的实例搜索算法(SINT)，通过学习匹配函数，对第 1 帧的初始块与当前帧候选样本进行相似度计算，返回最相似的候选样本作为目标当前状态，不再进行遮挡检测，无需更新模型，取得不错的跟踪效果。

综上，基于相关滤波框架的跟踪算法速度较快，但由于使用 HOG、CN 等单一特征，对遮挡问题没有做特别处理，对背景有强边缘、目标形变的场景表现不稳健。另外，基于卷积神经网络的跟踪算法精度较高，但难以预先获得大量样本进行训练，而且由于网络结构庞大而复杂，计算量大，直接影响跟踪算法的实时性。因此针对复杂场景下的快速稳健跟踪问题，本文提出一种利用深度卷积模型提取特征的抗遮挡实时目标跟踪算法。一方面，在针对目标任务改进的深度卷积模型上提取多层卷积特征，另一方面基于相关滤波框架，通过位置相关滤波器和尺度相关滤波器确定当前目标位置和目标尺度。同时，通过置信度指标判断目标当前遮挡状态，选择适宜的模型更新策略。本文算法在复杂环境下不但能够取得良好的跟踪精度，而且能够达到较快的跟踪速度，同时解决了跟踪过程中目标遮挡等问题。

2　算法概述

算法的整体框图如图 1 所示，主要分为 5 个部分。1)将视频序列第 1 帧中给定的目标候选位置区域的图像输入到针对目标跟踪任务改进的深度卷积网络(VGG-Net-19_OT)中，由相应的深度卷积层提取特征图，分别记作 Conv3_4_OT、Conv4_4_OT 和 Conv5_4_OT，进行训练并初始化 3 个相关滤波器 $W^{(3)}$、$W^{(4)}$、$W^{(5)}$。2) 对于输入视频的第 t 帧图像，以 t-1 帧图像中的目标预测结果为中心，确定搜索框，使用 VGG-Net-19_OT 网络模型获取搜索框区域内图像的深度卷积特征。3) 根据提取的深度卷积特征，与滤波器进行相关运算操作，根据快速傅里叶变换进行滤波器训练和响应图计算，从中获得响应值最大点的位置，即为跟踪目标在第 t 帧图像中的新位置，再通过尺度滤波器估计当前目标的最佳跟踪尺度。4)基于得到的跟踪目标的新位置，利用 VGG-Net-19_OT 网络模型在该中心位置区域内提取图像的深度特征，在线训练相关滤波器模型。5) 计算 3 个置信度评估指标，根据结果判断是否有遮挡，如有遮挡，将当前模型备份，对目标位置进行自适应更新，直到视频最后 1 帧。

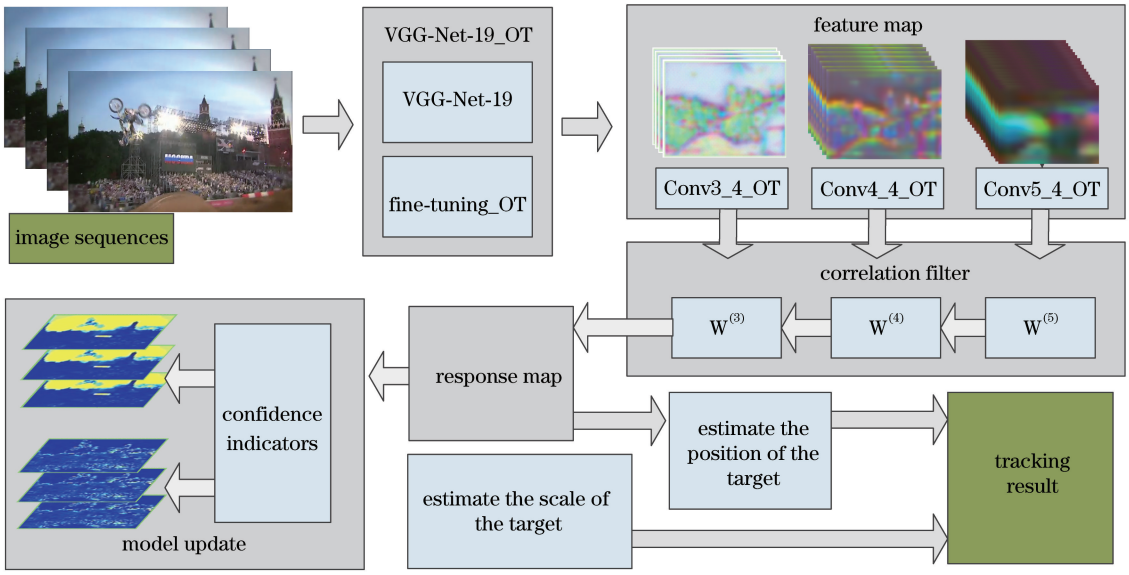

图 1　目标跟踪算法框架图

Fig. 1　Framework of the proposed visual tracking algorithm

3　基于多层深度卷积特征的目标跟踪算法

3.1　深度卷积特征

目标的特征表达是影响跟踪性能的重要因素，适用于目标跟踪任务的特征表达需要具备较高的区分性，能够对背景和非目标物体保持较好的判别性，还需要具备较强的泛化能力，能够适应各种遮挡、外观变化等不确定因素。同时，为达到跟踪的实时要求，特征表达的参数计算量需要尽可能地少。传统的目标特征提取如 HOG 特征、CN 特征等浅层特征，带有一定的先验知识，特征提取速度快，对于某些特定场景具有很好的表达能力和区分性，但在快速运动、遮挡、光照变化等复杂环境情况下稳健性较差。随着深度学习方法特别是卷积神经网络(CNNs)[19]在图像分类领域取得巨大的成功，出现了诸多性能优秀的网络模型，越来越多的研究者开始将其应用到目标跟踪领域。

3.1.1　卷积神经网络 VGG-Net

VGG-Net[20]是牛津大学计算机视觉组和 DeepMind 公司共同研发的深度卷积网络，探索了卷积神经网络的深度与性能之间的关系，通过反复堆叠 3×3 的小型卷积核和 2×2 的最大池化层，构筑了 16～19 层的卷积神经网络，证明小型卷积核可以通过增加网络的深度模仿较大卷积核，实现对图像的局部感知，减少网络训练参数，有效提升模型效果，影响网络最终的性能。卷积模型主要由 5 段卷积、2 个全连接特征层和 1 个全连接分类层组成。卷积核专注于扩大通道数，池化着重于缩小特征图的宽和高，逐渐忽略局部信息。网络结构复杂，参数占用空间与计算量大，以 VGG-Net-19[20]为例，各层参数量如表 1 所示。

VGG-Net 以物体分类作为回归训练标准，在层数递增时，仅对训练图片的判别性特征进行提取，对分类任务无贡献的背景等冗余信息将逐渐消失。目标跟踪与分类的目的不同，需要在跟踪时将目标从背景中区分出来，然而在实际场景中，背景包含同类型物体的可能性也存在。因此，VGG-Net 提取的深度特征并不完全适合直接用于目标跟踪任务。

3.1.2　目标跟踪的微调网络模型 VGG-Net-19_OT

深度卷积模型的优势来自于对大量标注训练数据的有效学习，而目标跟踪仅仅提供第 1 帧的边框作为训练数据。为了更好地在深度卷积模型中提取特征应用于目标跟踪任务，需要对 VGG-Net 的结构进行微调，微调过程的本质是针对特定数据集对卷积核进行微调，每个卷积核对应一个通道，提取一种更契合数据集的判别性特征。具有冗余的卷积层中，各通道的特征提取有很大重叠。减少特征提取能力差的卷积核，保留学习良好的卷积核作为微调初始值，以迭代删减的方式修剪冗余的网络，在保持网络精度前提下，有效减少存储空间、提升网络速度。调整后的网络 VGG-Net-19_OT 如图 2 所示。实线框中是基于大规模数据集 ImageNet[21]预训练的 VGG-Net-19 深度卷积模型，由于目标跟踪数据集的数据量相对较小，因此只选取 3 个不同层级 Conv3_4、Conv4_4、Conv5_4 卷积层输出的特征图

表 1　VGG-Net-19 的各层参数

Table 1　Parameters of VGG-Net-19

Structure	Filter	Output size /(pixel×pixel×pixel)	Memory /bit	Parameter
Image input		224×224×3	224×224×3=150528	0
Conv1_1	64	224×224×64	224×224×64=3211264	3×3×3×64=1728
Conv1_2	64	224×224×64	224×224×64=3211264	3×3×64×64=36864
POOL1		112×112×64	112×112×64=802816	0
Conv2_1	128	112×112×128	112×112×128=1605632	3×3×64×128=73728
Conv2_2	128	112×112×128	112×112×128=1605632	3×3×128×128=147456
POOL2		56×56×128	56×56×128=401408	0
Conv3_1	256	56×56×256	56×56×256=802816	3×3×128×256=294912
Conv3_2	256	56×56×256	56×56×256=802816	3×3×256×256=589824
Conv3_3	256	56×56×256	56×56×256=802816	3×3×256×256=589824
Conv3_4	256	56×56×256	56×56×256=802816	3×3×256×256=589824
POOL3		28×28×256	28×28×256=200704	0
Conv4_1	512	28×28×512	28×28×512=401408	3×3×256×512=1179648
Conv4_2	512	28×28×512	28×28×512=401408	3×3×512×512=2359296
Conv4_3	512	28×28×512	28×28×512=401408	3×3×512×512=2359296
Conv4_4	512	28×28×512	28×28×512=401408	3×3×512×512=2359296
POOL4		14×14×512	14×14×512=100352	0
Conv5_1	512	14×14×512	14×14×512=100352	3×3×512×512=2359296
Conv5_2	512	14×14×512	14×14×512=100352	3×3×512×512=2359296
Conv5_3	512	14×14×512	14×14×512=100352	3×3×512×512=2359296
Conv5_4	512	14×14×512	14×14×512=100352	3×3×512×512=2359296
POOL5		7×7×512	7×7×512=25088	0
FC6	4096	1×1×4096	1×1×4096=4096	7×7×512×4096=102760448
FC7	4096	1×1×4096	1×1×4096=4096	4096×4096=16777216
FC8	1000	1×1×1000	1×1×1000=1000	4096×1000=4096000

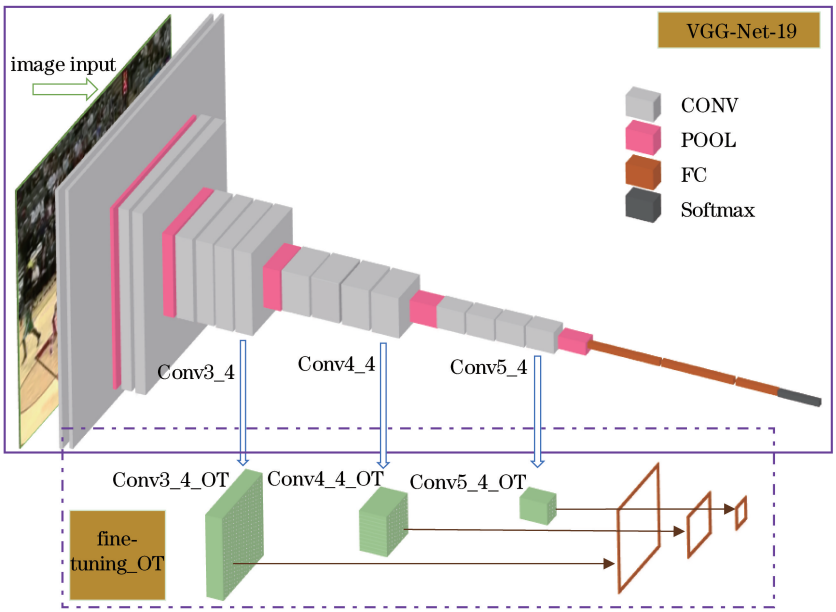

图 2　VGG-Net-19_OT 网络结构图

Fig. 2　Network structure of VGG-Net-19_OT

作为微调对象，虚线框中为针对目标跟踪任务添加的跟踪网络微调分支。

3.1.3　网络训练策略

首先对 VGG-Net-19 网络模型初始化，保持其结构固定不变。为适应目标在跟踪过程中表观模型的变化，将目标跟踪数据集中的图像筛选部分输入网络中进行训练。

采用 SSD[22] 的方式设定损失函数 $L(x,c,l,g)$，包括置信度损失 $L_{\mathrm{conf}}(x,c)$ 和位置损失 $L_{\mathrm{loc}}(x,l,g)$ 两部分：

$$L(x,c,l,g)=\frac{1}{N_{\mathrm{tr}}}\left[L_{\mathrm{conf}}(x,c)+\alpha L_{\mathrm{loc}}(x,l,g)\right], \tag{1}$$

式中，c 是默认框类别的置信度，α 表示权重系数，l 表示预测到的目标框，g 是真实的目标框，N_{tr} 为匹配的默认盒个数。则匹配类别 p 的第 i_{tr} 个默认框与第 j_{tr} 个真实目标框的置信度损失函数为

$$L_{\mathrm{conf}}(x,c)=-\sum_{i_{\mathrm{tr}}\in S_{\mathrm{Pos}}}^{N_{\mathrm{tr}}}x_{i_{\mathrm{tr}}j_{\mathrm{tr}}}^{(p)}\log\left[\hat{c}_{i_{\mathrm{tr}}}^{(p)}\right]-\sum_{i_{\mathrm{tr}}\in S_{\mathrm{Neg}}}\log\left[\hat{c}_{i_{\mathrm{tr}}}^{(0)}\right],\hat{c}_{i_{\mathrm{tr}}}^{(p)}=\frac{\exp\left[c_{i_{\mathrm{tr}}}^{(p)}\right]}{\sum_{p}\exp\left[c_{i_{\mathrm{tr}}}^{(p)}\right]}, \tag{2}$$

式中：S_{Pos}，S_{Neg} 分别代表正负样本集；$\hat{c}_{i_{\mathrm{tr}}}^{(0)}$ 表示预测值为背景的概率，概率越高则损失越小；$\hat{c}_{i_{\mathrm{tr}}}^{(p)}$ 表示预测值为目标的概率，p 为类别；概率越高则损失越小。

第 i_{tr} 个默认框与第 j_{tr} 个真实目标框的位置损失通过 smooth_{L1} 损失函数计算

$$L_{\mathrm{loc}}(x,l,g)=\sum_{i_{\mathrm{tr}}\in S_{\mathrm{Pos}}}^{N_{\mathrm{tr}}}\sum_{u\in\{c_x,c_y,w_{\mathrm{tr}},h_{\mathrm{tr}}\}}x_{i_{\mathrm{tr}}j_{\mathrm{tr}}}^{(k)}\cdot\mathrm{smooth}_{L1}\left[l_{i_{\mathrm{tr}}}^{(u)}-\hat{g}_{j_{\mathrm{tr}}}^{(u)}\right], \tag{3}$$

式中：$\{c_x,c_y,w_{\mathrm{tr}},h_{\mathrm{tr}}\}$ 为默认框的位置尺寸；$x_{i_{\mathrm{tr}}j_{\mathrm{tr}}}^{(k)}$ 表示第 i_{tr} 个默认框与第 j_{tr} 个真实目标框关于类别 k 是否匹配；$l_{i_{\mathrm{tr}}}^{(u)}$，$\hat{g}_{j_{\mathrm{tr}}}^{(u)}$ 分别表示第 u 个预测到的目标框与真实框。由于 $g^{(u)}$ 是真实框正规化的几何参数，在相关滤波算法中，目标大小和采样框的大小是固定的，为减少训练参数，仅进行粗略定位，(3)式可化简为

$$L_{\mathrm{loc}}(x,l,g)=\sum_{i_{\mathrm{tr}}\in S_{\mathrm{Pos}}}^{N_{\mathrm{tr}}}\sum_{u\in\{c_x,c_y\}}x_{i_{\mathrm{tr}}j_{\mathrm{tr}}}^{(k)}\cdot\mathrm{smooth}_{L1}\left[l_{i_{\mathrm{tr}}}^{(u)}-\hat{g}_{j_{\mathrm{tr}}}^{(u)}\right]。 \tag{4}$$

由于深度卷积网络中不同卷积核存在不同的稀疏度，部分卷积核的权重参数过于稀疏，对模型性能提升效果不高。为了更有针对性地提取目标特征，减少卷积核冗余，可以设定稀疏度阈值进行过滤。将小于阈值的卷积核权重设置为 0，逐层反复训练、微调，直至无法检测到冗余的卷积核，训练得到收敛的 VGG-Net-19_OT 网络。在应用场景中有同类背景干扰的情况下，当目标与背景是同一类物体时，调整后的网络训练出的特征也能进行区分，达到很好的跟踪效果。

3.1.4　深度卷积特征提取

预训练的深度卷积网络在不同卷积层提取到不同的特征，充分利用各个层次的特征可以提升目标跟踪的性能。底层特征具有较高的空间分辨率，包含丰富的空间特征和纹理信息，作为类内分类器时，可以剔除表观相似的干扰背景，同时易于捕捉位置的变化，进行精确定位。高层特征包含更多的语义信息，忽略物体的细节差异，难以识别或定位较小目标，作为类间分类器时，进行粗略定位，对于目标发生形变、遮挡等表观变化表现比较稳健[12]。

将视频图像目标候选区域输入改进后的预训练网络 VGG-Net-19_OT，并提取出适用于跟踪序列的多层深度卷积特征图，随着层数加深，特征图尺寸逐渐变小，空间分辨率逐步降低。对不同层级的特征图利用双线性插值进行上采样，获得一系列相同尺寸不同层级的深度特征图，进而层次化地构造目标外观模型，设 $\boldsymbol{h}$ 为原始特征图，$\boldsymbol{x}$ 为上采样后的特征图，β_{ik} 是插值系数，取决于位置 i 及其 k 邻域的特征向量，取决于第 i 个位置的特征向量：

$$x_i=\sum_k\beta_{ik}h_k。 \tag{5}$$

3.2　相关滤波框架

目标跟踪领域有很多主流的算法，如 KCF，都是基于核相关滤波框架。此滤波框架利用目标周围区域的循环矩阵采集正负样本，通过岭回归训练目标分类器，以循环矩阵在傅里叶空间可对角化的性质，将矩阵的运算转化为向量的 Hadamard 积，降低了运算量，提高了运算速度。本文算法在此相关滤波理论基础上进行构建。

3.2.1　训练阶段

选定跟踪目标，在给定的目标位置提取训练样本，将目标跟踪序列第 1 帧输入改进后的深度卷积网络模型 VGG-Net-19_OT，提取第 q 层深度卷积特征记为 $\boldsymbol{x}$，维度为 $M\times N\times D$，采用循环平移矩阵稠密采样的方法，在训练分类器时，根据样本 x_{ij}，

$i,j\in\{0,1,\cdots,M\}\times\{0,1,\cdots,N\}$，求得训练样本对应的二维高斯分布的回归标签 y_{ij}，构造相关滤波的目标函数为[13]

$$\boldsymbol{w}^{*}=\underset{\boldsymbol{w}}{\operatorname{argmin}}\sum_{i,j}\left\|\boldsymbol{w}\cdot x_{ij}-y_{ij}\right\|^{2}+\lambda\left\|\boldsymbol{w}\right\|_{2}^{2},\tag{6}$$

式中，λ 为正则化参数。利用傅里叶变换得到(6)式的闭式解，第 q 层 d 通道分类器权重的频域变换为

$$\boldsymbol{W}^{(d)}=\frac{\boldsymbol{Y}\odot\bar{\boldsymbol{X}}^{(d)}}{\sum_{d=1}^{D}\boldsymbol{X}^{(d)}\odot\bar{\boldsymbol{X}}^{(d)}+\lambda},\tag{7}$$

式中，$\boldsymbol{X}$、$\boldsymbol{Y}$ 分别为 $\boldsymbol{x}$、$\boldsymbol{y}$ 的频域变换，$\bar{\boldsymbol{X}}$ 为 $\boldsymbol{X}$ 的共轭，$\odot$ 表示 Hadamard 积。

3.2.2　检测阶段

将待检测样本提取出的第 q 层深度卷积特征构成的循环矩阵 $\boldsymbol{z}$ 输入训练好的分类器，得到回归函数 $f_q(\boldsymbol{z})$：

$$f_q(\boldsymbol{z})=\mathscr{F}^{-1}\left(\sum_{d=1}^{D}\boldsymbol{W}^{(d)}\odot\boldsymbol{Z}^{(d)}\right),\tag{8}$$

式中，$\mathscr{F}^{-1}$ 为傅里叶逆变换。首先计算位置 (m,n) 在第 q 层特征图的响应值，判断响应值最大的位置为跟踪目标在当前帧的精确位置 $(\hat{m},\hat{n})$：

$$(\hat{m},\hat{n})=\underset{m,n}{\operatorname{argmax}}f_q(\boldsymbol{z})。\tag{9}$$

再逐层向下搜索位置 $(\hat{m},\hat{n})$ 的 $r\times r$ 邻域，由于不同的卷积层的特征描述能力各不相同，每一层的响应峰值也不一样。可计算第 $q-1$ 层的响应图，作更细粒度的位置预测，逐层计算，以最低层的预测结果作为最后输出，计算式为

$$(\hat{m},\hat{n})=\underset{m,n}{\operatorname{argmax}}\sum_{q}\frac{1}{2^{5-q}\max(f_q(\boldsymbol{z}))}f_q(\boldsymbol{z})。\tag{10}$$

3.2.3　尺度估计

基于相关滤波框架的目标跟踪算法大都局限于对目标位置的预测，并未考虑针对运动目标的尺度变化进行估计。本文算法加入尺度相关滤波器，主要预测流程包括以下几步：首先，位置相关滤波器定位到目标后，在其周围采集不同尺度的图像，构成训练样本；接着，由于深度卷积特征计算量大，为保证算法的高效性，只提取 HOG 特征；然后，为保留目标的关键信息及平滑图像的边界效应，对提取到的特征进行加窗处理；最后，使用多尺度图像的特征训练核函数最小二乘分类器，得到尺度相关滤波器，寻找最大响应，其对应的尺度就是目标的新尺度。采用可变窗口大小的高斯窗函数代替余弦窗，通过 $\sigma_m=\frac{m}{w}$ 与 $\sigma_n=\frac{n}{h}$ 控制开窗大小，(m,n,w,h) 为位置尺度信息。其中二维高斯窗函数为

$$G(m,n,w,h)=\exp\left\{-\frac{1}{2}\left[\frac{i}{\sigma_m(m-1)}\right]^{2}\right\}\times\exp\left\{-\frac{1}{2}\left[\frac{j}{\sigma_n(n-1)}\right]^{2}\right\},\quad 0\leqslant i\leqslant m,0\leqslant j\leqslant n。\tag{11}$$

3.3　模型更新策略

为防止在跟踪过程中，由于目标表观模型的形变或外部条件的干扰，而造成漂移现象，需要对其进行模型更新。令第 t 帧 d 通道分类器权重为 $\boldsymbol{W}_t^{(d)}=\frac{\boldsymbol{A}_t^{(d)}}{\boldsymbol{B}_{\mathrm{t}}^{(d)}+\lambda}$，则

$$\boldsymbol{A}_t^{(d)}=(1-\eta)\boldsymbol{A}_{t-1}^{(d)}+\eta\boldsymbol{Y}\odot\bar{\boldsymbol{X}}_t^{(d)},\tag{12}$$

$$\boldsymbol{B}_t^{(d)}=(1-\eta)\boldsymbol{B}_{t-1}^{(d)}+\eta\sum_{d=1}^{D}\boldsymbol{X}_t^{(d)}\odot\bar{\boldsymbol{X}}_t^{(d)},\tag{13}$$

式中，学习速率 η 表征目标的外观模型对新视频图像帧的学习能力。η 值越小，学习速率越慢，无法及时捕捉目标外观模型变化；η 值越大，学习速率越快，易受外部噪声干扰。

对于相关滤波类跟踪算法，当目标被遮挡时，如果持续进行模型更新，会产生累积误差，导致模型污染，造成跟踪漂移，当遮挡逐步消除后，则很容易误判目标。因此模型更新策略需要重点解决局部遮挡判断问题。

本文算法采用高置信度的遮挡判断恢复机制，通过最大响应值 $F_{\max}$、平均峰值相关能量比 R_{APCE}、d 通道遮挡因子 E_{OCC_d} 三种指标进行评估。当这三种指标综合判定遮挡发生时，一方面，当前模型正常更新，以适应跟踪目标的表观变化。另一方面，由于当前模型包含较多目标信息，当目标重现时仍可对其进行识别，同时将其保存作为模型备份，等待遮挡结束后恢复。相关滤波器在判定遮挡已经发生时，计算当前模型和留存模型备份的响应，选择最优结果。当模型备份与当前模型响应差距较大且模型备份响应值足够好时，推断模型备份与当前模型识别不同的目标。若模型备份识别出目标与未遮挡时一致，此时响应值足够大，推断遮挡结束目标复现，用模型备份替换当前模型，完成整个模型恢复过程。

3.4　算法总体流程

结合上述对本文算法中关键部分的描述，给出算法的主要步骤，如图 3 所示。

Algorithm 1: Proposed tracking algorithm

Input: Initial target position $p_{t-1}=(x_{t-1},y_{t-1},w_{t-1},h_{t-1})$, the hierarchical correlation filters $\boldsymbol{W}_{t-1}^{(q)}(q=3,4,5)$

Output: Estimated object position $p_t=(x_t,y_t,w_t,h_t)$, $\boldsymbol{W}_t^{(q)}(q=3,4,5)$

1 Repeat

2 Crop out the searching window in frame t centered at (x_{t-1},y_{t-1}) and extract convolutional features with spatial interpolation using formula (5);

3 For each layer q do computing $\boldsymbol{W}_{t-1}^{(q)}$ and correlation response f_q using formulas (7) and (8);

4 Estimate the new position (x_t,y_t) on response map using formula (10);

5 Obtain the scale sample images around (x_t,y_t) and extract HOG features;

6 Compute scale correlation response;

7 Estimate the new scale (w_t,h_t) around (x_t,y_t);

8 Compute the occlusion indicators to update models using formulas (12) and (13);

9 Until end of video sequences

图 3　算法流程

Fig. 3　Flow chart of algorithm

4　实　　验

4.1　实验平台及参数配置

实验平台硬件配置为 CPU:Intel(R) Core(TM) i7-6700,3.4 GHz,16 GB 内存;GPU:NVIDIA GeForce GTX-1080。软件配置为 MATLAB 2016b 和 C++ 在 Matconvnet 深度学习库混合编程。算法的正则化参数 $\lambda=10^{-4}$,高斯核宽 $\sigma=0.1$。

4.2　评价标准

为评估算法的性能,在 OTB-2015[23] 与 UAV123[24] 数据集上进行测试。OTB-2015 有 100 组完全标注的视频,涵盖 11 个属性。UAV123 有 123 组完全标注的视频,涵盖 12 个属性。各属性包括视频序列个数分别如表 2、表 3 所示。

表 2　OTB-2015 视频属性

Table 2　Video attributes of OTB-2015

Video attribute	Value	Video attribute	Value
Background clutters (BC)	31	Motion blur (MB)	29
Deformation	44	Occlusion	49
Fast motion (FM)	39	Out-of-plane rotation (OPR)	63
Illumination variation (IV)	38	Out-of-view (OV)	14
In-plane rotation (IPR)	51	Scale variation (SV)	64
Low resolution (LR)	9		

表 3　UAV123 视频属性

Table 3　Video attributes of UAV123

Video attribute	Value	Video attribute	Value
Scale variation (SV)	109	Out of view (OV)	30
Aspect ratio change (ARC)	68	Background clutter (BC)	21
Low resolution (LR)	48	Illumination variation (IV)	31
Fast motion (FM)	28	Viewpoint change (VC)	60
Full occlusion (FOC)	33	Camera motion (CM)	70
Partial occlusion (POC)	73	Similar object (SOB)	39

选择一次通过评估(OPE)方法,绘制精确度图(Precision plot)和成功率图(Success plot)。采用 4 种常见的评估指标[25]:中心位置误差(CLE)、距离精度(DP)、重叠精度(OP)以及平均跟踪帧率(FPS)。

将 OTB-2015 代码库中自带的 ALSA[26] 等算法,以及近几年主流跟踪算法 CFNet[17]、CNN-

SVM[27]、DSST[28]、HCFT[12]、HCFTstar[13]、KCF[4]、LCTDeep[29]、SRDCF[5]等进行定性定量分析。针对 OTB-2015 数据集，将排名前 10 的跟踪算法在精确度图和成功率图上显示。针对 UAV123 数据集，仅选择 HCFTstar[13]、KCF[4]与本文算法共 3 种算法在精确度图和成功率图上比较。

4.3　定性分析

基于 OTB-2015 数据集测试中排名前 10 的算法的部分跟踪结果如图 4、图 5 所示，从 7 个方面进行分析。

4.3.1　背景复杂

以“Ironman 1”视频序列为例，如图 4(a)所示。目标运动过程中，背景与目标极为相似，多数算法都远离目标，本文算法与 HCFT、HCFTstar 算法得益于卷积网络提取的目标稳健性特征描述，能够跟上目标。

图 4　10 个跟踪算法在不同视频序列上的定性结果显示。(a) Ironman 1；(b) Ironman 2；(c) Doll；(d) MotorRolling；(e) Bolt2；(f) Skiing

Fig. 4　Qualitative results of the 10 tracking algorithms on different video sequences. (a) Ironman 1; (b) Ironman 2; (c) Doll; (d) MotorRolling; (e) Bolt2; (f) Skiing

4.3.2　光照变化

如图 4(b)所示，在“Ironman 2”中，由于不断有光束出现，光照剧烈变化，对算法的稳健性提出了极大的挑战，多数算法在跟踪开始就产生跟踪漂移直至失效，只有本文算法和 HCFT、HCFTstar 算法能够始终跟踪目标。

4.3.3　尺度变化

在“Doll”中，如图 4(c)所示，由于距离镜头时远时近，目标在跟踪过程中尺度不断变化，虽然 90% 的所选算法都能跟踪目标，但本文算法能够更好地根据尺度进行调整。

4.3.4　目标旋转

参考图 4(d)中“MotorRolling”视频序列，目标在跟踪过程中经历了超过 360°的旋转，对算法提出了很大的挑战。除了本文算法和 HCFT、HCFTstar、LCTDeep、CFNet 算法，其余都出现了跟踪失败。

4.3.5　快速运动

图 4(e)中，视频序列“Bolt2”是短跑比赛场景，跟踪目标为其中一名运动员。目标姿态不断变化，且随着镜头的转动图像中运动员也从正面逐渐转向背面。本文算法由于提取了高层特征，受目标外观变化的影响不大，可以始终跟踪目标。

4.3.6　目标分辨率低

以视频“Skiing”为例，如图 4(f)所示，目标分辨率低且尺寸小，对特征提取的性能提出了更高的要求。由于本文算法从微调后的预训练网络提取到多层深度卷积特征，能够更好地跟踪到弱小目标。

4.3.7　目标遮挡

选取 4 组典型的序列“Jogging-1”、“Walking2”、“Coke”和“Soccer”，目标在跟踪过程中受到部分或全部遮挡，如图5所示。当目标被遮挡后，本文算法

图 5　10 个跟踪算法在部分遮挡视频序列上的定性结果显示。(a) Jogging-1；(b) Walking2；(c) Coke；(d) Soccer

Fig. 5　Qualitative results of the 10 tracking algorithms on different occluded video sequences. (a) Jogging-1; (b) Walking2; (c) Coke; (d) Soccer

通过微调后的深度卷积网络提取特征，使遮挡物和目标的区分性更强，跟踪上目标，未出现跟踪漂移。另外，利用置信度指标判断模型更新机制，避免在遮挡因素下产生错误模型更新，在遮挡消失后，通过恢复备份模型，更新模型。

4.4　定量分析

为进一步全面评估所提算法的性能，对 OTB-2015 与 UAV123 的测试视频序列的综合性能进行定量分析。

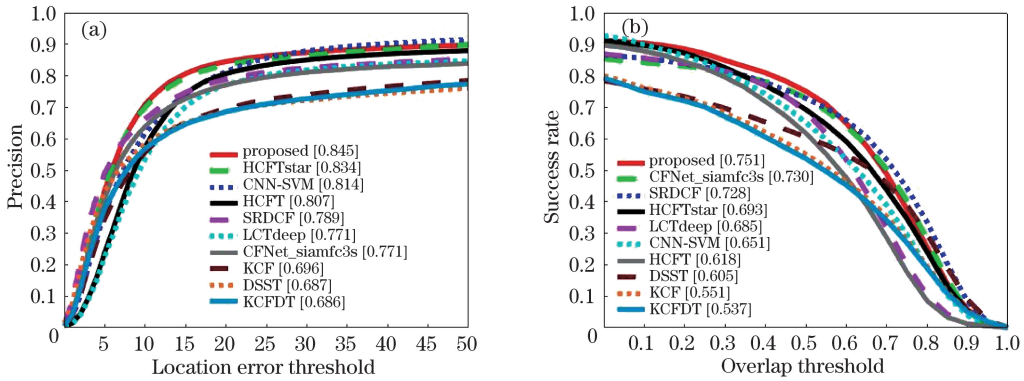

图 6　基于 OTB-2015 评估基准 OPE 的跟踪算法。(a)精度曲线图；(b)成功率曲线图

Fig. 6　Algorithm of OPE on OTB-2015. (a) Precision plot; (b) overlap success plot

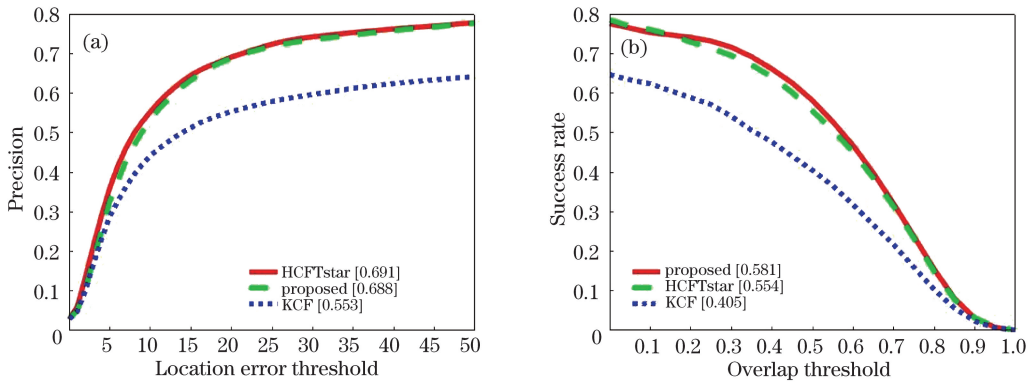

图 7　基于 UAV123 评估基准 OPE 的跟踪算法。(a)精度曲线图；(b)成功率曲线图

Fig. 7　Algorithm of OPE on UAV123. (a) Precision plot; (b) overlap success plot

4.4.1　算法综合性能的定量分析

图 6 是基于 OTB-2015 数据集排名前 10 的跟踪算法的精度曲线和成功率曲线，图例中标注的是每种算法的性能评分。本文算法在所有 100 段视频上的跟踪精度评分 0.845、成功率评分 0.751，在对比算法中表现最好。

图 7 是基于 UAV123 数据集的精度曲线和成功率曲线，本文算法跟踪精度评分 0.688、成功率评分 0.581，与 HCFTstar 算法性能持平。

4.4.2　基于 OTB-2015 的不同视频属性的定量分析

面对不同挑战，平均精度和平均成功率结果分别如图 8 和图 9 所示。

本文算法在 11 种不同属性的跟踪挑战中，跟踪精度始终取得最优或次优的成绩。跟踪成功率在 LR、OV 属性排名第三，在 SV 属性处于次优，其余属性均排名第一。由此表明，本文算法可以较好地适应于复杂场景下的目标跟踪任务。

4.4.3　基于 UAV123 的不同视频属性的定量分析

如表 4 所示，本文算法在 12 种不同属性的跟踪挑战中，跟踪精度、跟踪成功率与 HCFTstar 算法持平，特别是在遮挡和部分遮挡序列挑战中可以较好地适应。

4.5　算法跟踪速率

跟踪算法对实时性的要求比较高，任何冗余计算都会影响算法的实用性。由于高维卷积特征需要进行复杂的计算，这导致 VGG-Net 提取目标特征用时较多，本文算法在其模型结构基础上增加目标跟踪分支，优化筛选更适用于目标跟踪任务的卷积核，将小于阈值的卷积核权重设置为 0，移除，逐层训练，得到 VGG-Net-19_OT 深度卷积网络模型，各层有效卷积核的个数约降低为VGG-Net-19的1/9。

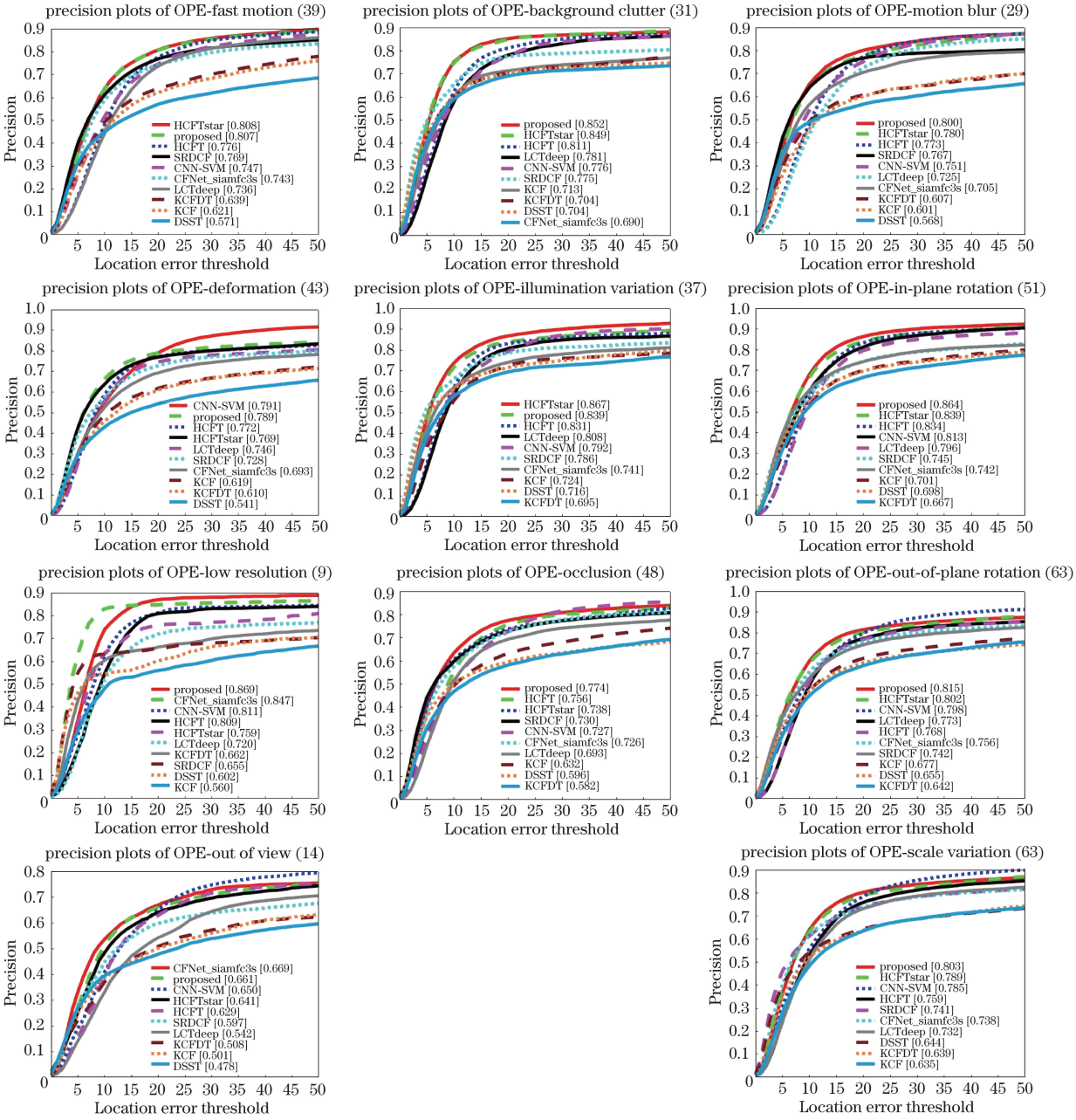

图 8　OTB-2015 11 种不同属性视频序列跟踪精度曲线

Fig. 8　Precision plots on 11 different attributes video sequences of OTB-2015

表 4　UAV123 12 种不同属性视频序列跟踪精度与跟踪成功率

Table 4　Precision values and success rates on 12 different attributes video sequences of UAV123

Sequence	Proposed algorithm		KCF		HCFTstar	
	Precision value	Success rate	Precision value	Success rate	Precision value	Success rate
Aspect ratio change (ARC)	0.619	0.464	0.447	0.292	0.610	0.434
Background clutter (BC)	0.585	0.447	0.536	0.413	0.584	0.470
Camera motion (CM)	0.677	0.556	0.502	0.366	0.682	0.543
Fast motion (FM)	0.544	0.402	0.301	0.200	0.516	0.377
Full occlusion (FOC)	0.567	0.358	0.420	0.243	0.561	0.381
Illumination variation (IV)	0.627	0.506	0.464	0.334	0.614	0.451
Low resolution (LR)	0.555	0.333	0.435	0.251	0.579	0.346
Out of view (OV)	0.609	0.500	0.406	0.277	0.603	0.467
Partial occlusion (POC)	0.632	0.499	0.497	0.365	0.628	0.491
Scale variation (SV)	0.644	0.534	0.497	0.339	0.646	0.498
Similar object (SOB)	0.691	0.566	0.616	0.418	0.693	0.552
Viewpoint change (VC)	0.637	0.494	0.450	0.302	0.625	0.440

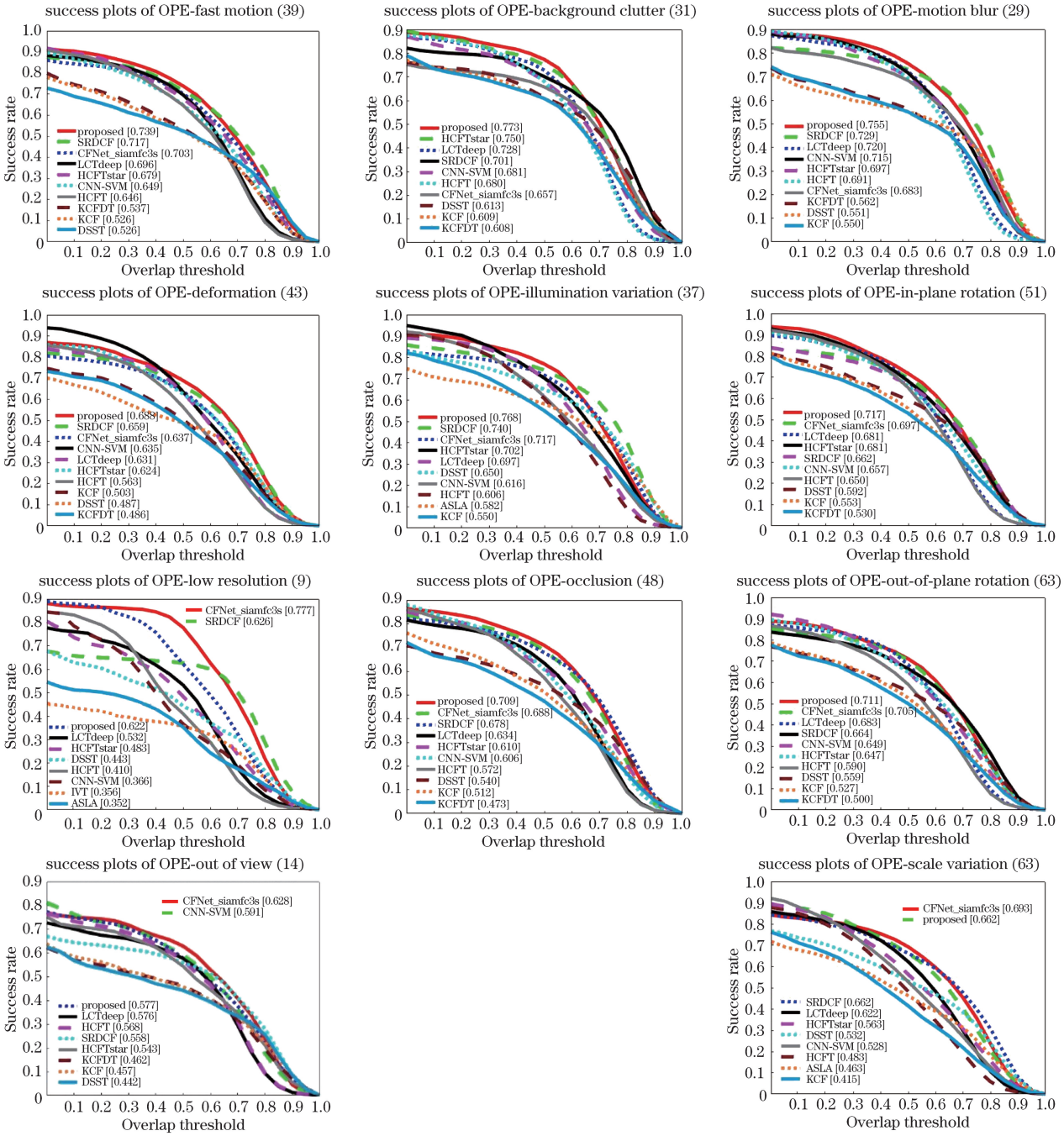

图 9　OTB-2015 11 种不同属性视频序列跟踪成功率曲线

Fig. 9　Success plots on 11 different attributes video sequences of OTB-2015

目标特征提取层参数的减少直接减少了跟踪过程的计算量，提升了跟踪的速度。本文算法在 OTB-2015 的 100 组视频的测试中，CPU 模式下平均速度为 16.3 frame/s，GPU 加速条件下，平均跟踪速率为 29.6 frame/s，达到实时跟踪要求。表 5 列举了部分视频序列的跟踪速率。

表 6 列出了 FCNT、MDNet、HCFT 等当前基于深度学习的主流跟踪算法的跟踪速率，并与本文算法对比。从表中可以看出，较传统的基于深度学习的跟踪算法，本文跟踪算法在跟踪速率上有较大的提升，基本可以满足实时要求。

综上分析，本文算法在跟踪过程中，遇到背景干扰、光照变化、遮挡、尺度变化、快速运动等变化时，均表现出良好的跟踪性能，且速度优于其他深度学习类算法。

5　结　　论

本文提出一种结合深度卷积特征和相关滤波框

表 5 跟踪速率

Table 5 Tracking speeds frame /s

Sequence	Basketball	FaceOcc1	Football1	Girl	Jogging1	Jumping	Soccer	Sylvester	Trellis
Speed	31.3	34.5	26.1	35.9	28.2	26.1	25.1	32.6	27.9

表 6 基于深度学习的跟踪算法的平均跟踪速率对比

Table 6 Average tracking speed comparison for the deep learning-based tracking algorithm frame /s

Algorithm	Proposed	FCNT	MDNet	HCFT
Tracking speed	29.6	3	1	10

架的抗遮挡实时算法。对卷积网络模型 VGG-Net-19 结构进行调整，加入针对目标跟踪任务训练的卷积层，优化冗余卷积核，减轻高维卷积特征结构复杂度。通过微调后的深度卷积模型提取目标表观特征，在维持跟踪精度的同时，降低了特征提取的计算量，从而加快了算法运行速度。同时本文算法引入阶段式模型更新恢复机制，解决了相关滤波跟踪误差随时间积累导致模型污染的问题。实验结果表明，与近年来的主流算法相比，本文算法不仅得到较高的跟踪精度，在目标遮挡、外观变化、背景干扰等复杂场景下也有稳健的表现，而且平均跟踪速度可达到 29.6 frame/s，满足实际应用的要求。

参 考 文 献

[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, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE, 2010: 2544-2550.

[2] Henriques J F, Caseiro R, Martins P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels[M]// Fitzgibbon A, Lazebnik S, Perona P, *et al*. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2012, 7575: 702-715.

[3] Danelljan M, Khan F S, Felsberg M, *et al*. Adaptive color attributes for real-time visual tracking[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 1090-1097.

[4] 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.

[5] Danelljan M, Häger G, Khan F S, *et al*. Learning spatially regularized correlation filters for visual tracking[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4310-4318.

[6] Wang X, Hou Z Q, Yu W S, *et al*. Target scale adaptive robust tracking based on fusion of multilayer convolutional features[J]. Acta Optica Sinica, 2017, 37(11): 1115005.
王鑫, 侯志强, 余旺盛, 等. 基于多层卷积特征融合的目标尺度自适应稳健跟踪[J]. 光学学报, 2017, 37(11): 1115005.

[7] Cai Y Z, Yang D D, Mao N, *et al*. Visual tracking algorithm based on adaptive convolutional features[J]. Acta Optica Sinica, 2017, 37(3): 0315002.
蔡玉柱, 杨德东, 毛宁, 等. 基于自适应卷积特征的目标跟踪算法[J]. 光学学报, 2017, 37(3): 0315002.

[8] Li C, Lu C Y, Zhao X, *et al*. Scale adaptive correlation filtering tracing algorithm based on feature fusion[J]. Acta Optica Sinica, 2018, 38(5): 0515001.
李聪, 鹿存跃, 赵珣, 等. 特征融合的尺度自适应相关滤波跟踪算法[J]. 光学学报, 2018, 38(5): 0515001.

[9] Wang H Y, Wang L, Yin W R, *et al*. Multi-scale correlation filtering visual tracking algorithm combined with target detection[J]. Acta Optica Sinica, 2019, 39(1): 0115004.
王红雨, 汪梁, 尹午荣, 等. 结合目标检测的多尺度相关滤波视觉跟踪算法[J]. 光学学报, 2019, 39(1): 0115004.

[10] Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems, December 5-10, 2013, Lake Tahoe, Nevada. USA: Curran Associates Inc., 2013, 1: 809-817.

[11] Wang N, Li S, Gupta A, *et al*. Transferring rich feature hierarchies for robust visual tracking[EB/OL]. (2015-04-23)[2019-01-06]. https://arxiv.org/abs/

1501.04587.

[12] Ma C, Huang J B, Yang X K, *et al*. Hierarchical convolutional features for visual tracking[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3074-3082.

[13] Ma C, Huang J B, Yang X K, *et al*. Robust visual tracking via hierarchical convolutional features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access), (2018-08-13)[2019-01-06]. DOI: 10.1109/TPAMI.2018.2865311.

[14] Wang L J, Ouyang W L, Wang X G, *et al*. Visual tracking with fully convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 3119-3127.

[15] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 4293-4302.

[16] Bertinetto L, Valmadre J, Henriques J F, *et al*. Fully-convolutional Siamese networks for object tracking[M]// Hua G, Jégou H. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5000-5008.

[18] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1420-1429.

[19] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2019-01-06]. https://arxiv.org/abs/1409.1556.

[21] Deng J, Dong W, Socher R, *et al*. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 248-255.

[22] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C]// Leibe B, Matas J, Sebe N, *et al*. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[23] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[24] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[M]// Leibe B, Matas J, Sebe N, *et al*. Lecture notes in computer science. Cham: Springer, 2016, 9905: 445-461.

[25] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 2411-2418.

[26] Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 1822-1829.

[27] Hong S, You T, Kwak S, *et al*. Online tracking by learning discriminative saliency map with convolutional neural network[C]// Proceedings of the 32nd international Conference on Machine Learning, July 6-11, 2015, Lille, France. Massachusetts: JMLR. org, 2015, 37: 597-606.

[28] Danelljan M, Häger G, Khan F S, *et al*. Accurate scale estimation for robust visual tracking[C]// Proceedings of the British Machine Vision Conference 2014, September 1-5, 2014, Nottingham. Durham, England, UK: BMVA Press, 2014: 1-11.

[29] Ma C, Huang J B, Yang X K, *et al*. Adaptive correlation filters with long-term and short-term memory for object tracking[J]. International Journal of Computer Vision, 2018, 126(8): 771-796.