基于空间可靠性约束的鲁棒视觉跟踪算法

蒲 磊*① 冯新喜② 侯志强③ 余旺盛②
①(空军工程大学研究生院 西安 710077)
②(空军工程大学信息与导航学院 西安 710077)
③(西安邮电大学计算机学院 西安 710121)

摘 要：针对复杂背景下目标容易发生漂移的问题，该文提出一种基于空间可靠性约束的目标跟踪算法。首先通过预训练卷积神经网络(CNN)模型提取目标的多层深度特征，并在各层上分别训练相关滤波器，然后对得到的响应图进行加权融合。接着通过高层特征图提取目标的可靠性区域信息，得到一个二值注意力矩阵，最后将得到的二值矩阵用于约束融合后响应图的搜索范围，范围内的最大响应值即为目标的中心位置。为了处理长时遮挡问题，该文提出一种基于首帧模板信息的随机选择更新策略。实验结果表明，该算法在应对相似背景干扰、遮挡、超出视野等多种场景均有良好的性能表现。

关键词：视觉跟踪；空间可靠性约束；深度特征；相关滤波；模型更新
中图分类号：TP391.4 **文献标识码**：A **文章编号**：1009-5896(2019)07-1650-08
DOI: 10.11999/JEIT180780

Robust Visual Tracking Based on Spatial Reliability Constraint

PU Lei① FENG Xinxi② HOU Zhiqiang③ YU Wangsheng②
①(*Graduate College*, *Air Force Engineering University*, *Xi'an* 710077, *China*)
②(*Institute of Information and Navigation*, *Air Force Engineering University*, *Xi'an* 710077, *China*)
③(*School of Computer Science and Technology*, *Xian University of Posts and Telecommunications*, *Xi'an* 710121, *China*)

Abstract: Because of the problem that the target is prone to drift in complex background, a robust tracking algorithm based on spatial reliability constraint is proposed. Firstly, the pre-trained Convolutional Neural Network (CNN) model is used to extract the multi-layer deep features of the target, and the correlation filters are respectively trained on each layer to perform weighted fusion of the obtained response maps. Then, the reliability region information of the target is extracted through the high-level feature map, a binary matrix is obtained. Finally, the obtained binary matrix is used to constrain the search area of the response map, and the maximum response value in the area is the target position. In addition, in order to deal with the long-term occlusion problem, a random selection model update strategy with the first frame template information is proposed. The experimental results show that the proposed algorithm has good performance in dealing with similar background interference, occlusion, and other scenes.

Key words: Visual tracking; Spatial reliability constraint; Deep features; Correlation filter; Model update

1 引言

视觉目标跟踪是计算机视觉领域的关键问题之一[1,2]，广泛应用于视频监控、人机交互、自动控制以及医学成像等领域。视觉跟踪一般指的是在首帧目标信息给定的情况下，去估计后续帧目标的状态，如位置、尺度等。近年来，随着视觉跟踪技术的发展，简单场景下的跟踪任务已经得到了很好的解决。目前的相关研究主要是关注复杂环境下的跟踪问题，比如光照变化、严重遮挡、背景杂波、快速运动以及大范围尺度变化等等，这些挑战的存在限制了目标跟踪算法性能的进一步提升。为了获取大量外观变化下目标的鲁棒特征描述，研究人员先后设计了颜色直方图、HOG、Harr特征、SURF、ORB、子空间表示和超像素等手工特征。最近从卷积神经网络(Convolutional Neural Network,

收稿日期：2018-08-07；改回日期：2019-01-21；网络出版：2019-02-15
*通信作者： 蒲磊 warmstoner@163.com
基金项目：国家自然科学基金(61571458, 61473309, 41601436)
Foundation Items: The National Natural Science Foundation of China (61571458, 61473309, 41601436)

CNN)中学到的特征被广泛用于图像分类[3]、目标识别[4]和图像分割[5]等视觉任务，并取得了优异的性能表现。因此，如何更好地将CNN的深度特征用于目标跟踪任务便成为一个值得研究的方向。

已有的基于深度学习的跟踪算法通常将跟踪任务表述为每帧中的目标检测问题[6–10]。这些方法首先在估计的目标位置附近提取大量的正负样本去训练CNN分类器，再将其用于检测后续帧的目标。但是这些算法存在两个问题，首先是都遵循图像分类方法仅使用CNN的最后一层的输出来表示目标，对于视觉识别、分类任务是有效的，因为分类层的特征和类别语义信息密切相关。但是视觉跟踪任务的目的在于准确地对目标进行定位而不是推断出他们的语义类别。因此单纯使用最后一层的特征对视觉跟踪任务并不是理想的。第2个问题是训练样本的提取，训练一个强大的分类器需要大量的正负样本，但是视觉跟踪只给定了第1帧的数据，数据量严重不足，如果通过在目标附近采集大量的样本，样本之间会存在高度的相关性，难以确定最优的决策边界。

本文通过以下方式解决上述的两个问题：(1)首先去掉了网络的分类层，采用具有空间分辨率的卷积特征对目标进行表示，为了增强特征的鲁棒性，本文采取具有互补性的多层特征进行融合而不是仅使用最后一层来表示目标；(2)为了解决训练样本不足的问题，本文在各层特征上独立地训练相关滤波器，将所有特征的移位版本视为训练样本，并将它们回归到由高斯函数生成的范围在[0, 1]的软标签，从而减轻训练二元判别分类器的采样模糊度，在获取对目标位置的多个估计后，采用加权融合的方式获取目标的最终位置。

但是传统的相关滤波依然存在很多问题，研究人员先后提出了CSK[11], CN[12], KCF[13], DSST[14], SRDCF[15], C-COT[16]等一系列优秀的方法，从多个角度对算法进行了改进。本文在大量实验的基础上，针对跟踪过程中不真实样本造成的背景杂乱问题，通过构建空间可靠区域对边界区域进行抑制；针对遮挡情况下造成的目标漂移问题，本文引入了一种随机更新策略，将首帧模板的引入使得模型在污染时可以得到适度的修正，依然保持一定的分辨力。另外，针对目标在跟踪过程中的尺度变化问题，本文通过位置滤波器之外再构建了一个尺度滤波器进行处理，取得了很好的效果。

本文从特征提取和分类器训练两个方面入手，提出了基于空间可靠性约束的鲁棒视觉跟踪算法。在大规模基准数据集上测试结果表明，本文算法不仅可以处理遮挡、目标旋转、尺度变化以及背景干扰等问题，同时也具有较高的跟踪速度，并在数据集上有着优异的性能表现。

2　本文算法

本文提出了基于空间可靠性约束的鲁棒视觉跟踪算法。首先通过提取多层特征训练多个相关滤波器，再对得到的多个响应图进行加权融合，接着建立目标可靠性区域约束范围并获得最终的响应图，取得响应最大值便是目标的中心位置，最后提出一种随机更新策略对模型进行更新。

2.1　多层卷积特征

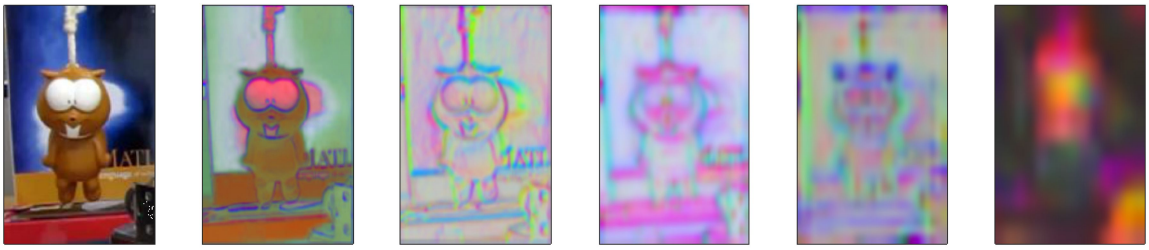
CNN是一个经典的网络架构，近年来随着计算机性能的大幅度提升以及大规模标记数据库ImageNet[17]的出现，出现许多性能优异的卷积神经网络模型，如AlexNet[18], VGGNet[19]和ResNet[20]等，并被广泛应用于计算机视觉领域。

近年来的研究表明，从深度网络提取的不同卷积层特征具有不同的视觉任务特性。高层的CNN特征捕捉目标的抽象和语义特征，这些特征可以用来区分不同类别的对象，并且对目标外观的剧烈变化有着良好的适应性，但是难以区分相似目标的干扰。相反，浅层的特征对于目标外观变化的稳健性较差，但是可以提供目标更加详细的局部信息，这些特征对于区分相似外观的目标非常有利，同时也可以获得对目标更加准确的定位。这些特性就自然而然地启发我们采用深度特征和浅层特征进行融合来提高跟踪器的性能。

本文采用VGGNet-19作为特征提取器，其中的“19”表示网络中需要学习的权重的层数。在图1中可见，提取了第1个到第5个卷积层的特征并进行了可视化。从图中可以看出，早期卷积层的特征图保留了更高的分辨率，而靠后层提取的特征捕获了更多的语义信息，逐渐丢失了细粒度的空间细节信息。本文方法主要在于用靠后层的语义信息来处理大的外观变化，并通过浅层的特征来进行精确定位来缓解目标漂移。

2.2　相关滤波器

本文采用在ImageNet上训练的VGGNet-19模型作为目标的特征提取器，将目标在Conv1-2, Conv4-4和Conv5-4的卷积特征分别训练多个相关滤波器，实现对目标的鲁棒跟踪。首先给定目标的待搜索区域，提取深度特征$\boldsymbol{z} \in \boldsymbol{R}^{M\times N\times D}$，为了更好地融合多层特征，所有的特征图都通过双线性插值上采样到固定的大小224×224。将目标的多维特征$\boldsymbol{z}$沿着垂直和水平方向上循环样本作为训练样本，每个样本可以表示$\boldsymbol{z}_{m,n}$, $(m,n) \in \{0,1,\cdots,$



(a)输入 (b) cov1-2 (c) cov2-2 (d) cov3-4 (e) cov4-4 (f) cov5-4

图 1 卷积深度特征可视化

$M-1\}\times\{0,1,\cdots,N-1\}$。每个样本的期望输出标签为$g(m,n)$，采用2维高斯函数

$$g(m,n)=\exp\left(-\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma_T^2}\right) \tag{1}$$

通过岭回归来最小化输出误差，可以得到最优滤波器为

$$\boldsymbol{h}^*=\arg\min_{\boldsymbol{h}}\left(\sum_{m,n}\left\|\sum_{d=1}^{D}\boldsymbol{h}^d\boldsymbol{z}^d_{m\times n}-g(m,n)\right\|^2+\lambda\|\boldsymbol{h}\|_2^2\right) \tag{2}$$

其中，λ为正则化参数且$\lambda\geq 0$。令$\varepsilon=\sum_{m,n}\left\|\sum_{d=1}^{D}\boldsymbol{h}^d\boldsymbol{z}^d_{m\times n}-g(m,n)\right\|^2+\lambda\|\boldsymbol{h}\|_2^2$，根据帕萨瓦尔定理可以得到$\varepsilon$的频域表示

$$\tilde{\varepsilon}=\frac{1}{MN}\left(\left\|\sum_{d=1}^{D}\boldsymbol{H}^d\circ\left(\boldsymbol{Z}^d\right)^*-\boldsymbol{G}\right\|^2+\lambda\sum_{d=1}^{D}\left\|\boldsymbol{H}^d\right\|^2\right) \tag{3}$$

其中，$\boldsymbol{Z}$, $\boldsymbol{G}$和$\boldsymbol{H}$分别为$\boldsymbol{z}$, $\boldsymbol{g}$和$\boldsymbol{h}$的离散傅里叶变换；$\boldsymbol{Z}^*$为$\boldsymbol{Z}$的复数共轭；$\circ$表示元素的点乘运算。通过对计算$\partial\tilde{\varepsilon}/\left(\partial\boldsymbol{H}^d\right)=0$，可以求得每个通道$d(d\in\{1,2,\cdots,D\})$上的最优相关滤波器为

$$H^d=\frac{G\circ\left(Z^d\right)^*}{\sum_{d=1}^{D}Z^i\circ\left(Z^i\right)^*+\lambda} \tag{4}$$

因此，给定第t+1帧中目标的在某一层上的卷积特征图$\boldsymbol{z}_t$, $\boldsymbol{z}_t\in\boldsymbol{R}^{M\times N\times D}$，其DFT变换为$\boldsymbol{Z}_t$，可以得到第$t$帧的相关响应图$E$

$$\boldsymbol{E}=\mathrm{F}^{-1}\left(\sum_{d=1}^{D}\boldsymbol{H}^d\circ\left(\boldsymbol{Z}_t{}^d\right)^*\right) \tag{5}$$

其中，$(\boldsymbol{Z}_t)^*$为$\boldsymbol{Z}_t$的复数共轭；F^{-1}表示逆傅里叶变换。按照上述步骤，便得到了不同层的多个相关响应图$\{\boldsymbol{E}_1,\boldsymbol{E}_4,\boldsymbol{E}_5\}$，接着通过加权融合获得最终的响应图

$$\boldsymbol{E}_{\mathrm{f}}=\gamma_1\cdot\boldsymbol{E}_1+\gamma_2\cdot\boldsymbol{E}_4+\gamma_3\cdot\boldsymbol{E}_5 \tag{6}$$

其中，γ_1, γ_2和γ_3代表不同响应图的权值。

2.3 构建空间可靠性区域矩阵

为了减轻背景干扰对算法性能的影响，本文提出了一种简单且有效的方法用于确定目标在空间中的位置。虽然单个特征图并不具有一定的分辨能力，但是如果很多特征通道都在某一区域具有较大的响应，便有理由相信该区域是一个目标而非背景。本文通过对Cov5-4的深度特征$\boldsymbol{z}_5\in\boldsymbol{R}^{M\times N\times D}$在通道$D$的维度上进行相加运算，这样便得到一个2维的矩阵$\boldsymbol{M}=\sum_{n=1}^{D}\boldsymbol{S}_n$，其中$\boldsymbol{S}_n$表示Cov5-4的$n$个特征图。接着考虑更高的激活响应区域更有可能是目标，通过计算$\boldsymbol{M}$中所有位置响应的均值$\bar{m}$来作为判断目标和背景的阈值，高于阈值的便是目标大致区域

$$\hat{\boldsymbol{M}}_{i,j}=\begin{cases}1, & \boldsymbol{M}_{i,j}>\bar{m}\\ 0, & \text{其他}\end{cases} \tag{7}$$

由于背景的复杂性，仍然可能存在一些小的部分响应较高，对此，本文获取$\hat{\boldsymbol{M}}$的最大连通部分，作为目标的最终可靠性区域，采用$\tilde{\boldsymbol{M}}$表示。接着通过得到可靠性区域矩阵来限制响应的搜索范围

$$\tilde{\boldsymbol{E}}_{\mathrm{f}}=\boldsymbol{E}_{\mathrm{f}}\circ\tilde{\boldsymbol{M}} \tag{8}$$

其中，$\circ$表示元素的点乘运算。最后在得到的最终响应图上可以定位出当前目标的中心位置(x_t,y_t)

$$(x_t,y_t)=\arg\max_{m,n}\tilde{\boldsymbol{E}}_{\mathrm{f}}(m,n) \tag{9}$$

2.4 随机更新策略

在跟踪过程中，目标外观和背景由于许多原因而不断变化，为了提高相关滤波器的鲁棒性和判别能力，必须对模型进行一定程度的更新。但是当目标受到诸如严重遮挡或外观的巨大变化等相当具有挑战性的场景时，常规的更新方法很容易将错误的信息引入模型。尤其是当目标存在严重遮挡时，传统的更新方法会导致错误积累导致跟踪漂移甚至失败。在视觉跟踪中，只有第1帧的信息为目标的外观和位置提供了最可靠的信息，因此本文考虑将第

1帧作为原始目标模板，然后依靠某种判决准则近乎随机地将其引入当前模型中，缓解了长时严重遮挡对目标模板的破坏。

在实验中，本文发现目标的遮挡和最大响应值τ之间确实存在一定的关系，但是并不是最大响应值的数值变化就可以对目标的遮挡情况进行判定。因为很多场景的变化都可以造成目标响应值的跳跃性改变。通过大量的实验，并对最终的响应值进行归一化后，得出了两个观察结果：

(1)当目标被长时间遮挡且背景不变的情况下，模型将会完全将背景看做目标，此时的响应值将一直保持为最大值1。

(2)由于进行了归一化，并且本文采用了3个相关滤波器进行跟踪，只有当三者的峰值响应完全在同一位置时，才会出现最大值响应为1，而这种情况是以一种近乎随机的方式出现。

通过上述两点观察，最大响应值τ是否等于1可以作为加入首帧模板进行更新的一个简单的判决准则。这样处理，使得模型始终具有一定的判别力，可以在一定程度上实现对目标的长时跟踪。

对于相关滤波器$\boldsymbol{H}^d$，令$\boldsymbol{A}_{t-1}^d$和$\boldsymbol{B}_{t-1}^d$分别表示滤波器在$t-1$帧的分子项和分母项，则在第t帧模型的更新策略为

$$\boldsymbol{A}_t^d=\begin{cases}(1-\eta)\,\boldsymbol{A}_{t-1}^d+\eta\boldsymbol{G}\odot\bar{\boldsymbol{Z}}_t(t)\,, \\ \qquad\qquad \tau<1 \\ (1-\eta)\,\boldsymbol{A}_{t-1}^d+\eta\left[\omega\boldsymbol{A}_1^d+(1-\omega)\left(\boldsymbol{G}\odot\bar{\boldsymbol{Z}}_t(t)\right)\right], \\ \qquad\qquad \tau=1\end{cases} \tag{10}$$

$$\boldsymbol{B}_t^d=\begin{cases}(1-\eta)\,\boldsymbol{B}_{t-1}^d+\eta\sum\limits_{i=1}^{D}\boldsymbol{Z}_t^i(t)\odot\bar{\boldsymbol{Z}}_t^i(t)\,,\tau<1 \\ (1-\eta)\,\boldsymbol{B}_{t-1}^d+\eta\left[\omega\boldsymbol{B}_1^d+(1-\omega)\right. \\ \qquad\cdot\left.\left(\sum\limits_{i=1}^{D}\boldsymbol{Z}_t^i(t)\odot\bar{\boldsymbol{Z}}_t^i(t)\right)\right], \qquad \tau=1\end{cases} \tag{11}$$

$$\boldsymbol{H}_t^d=\boldsymbol{A}_t^d/(\boldsymbol{B}_t^d+\lambda) \tag{12}$$

其中，η为学习率，ω为加权因子，τ为最大响应值，$\boldsymbol{A}_1^d$和$\boldsymbol{B}_1^d$为首帧模板。

通过将首帧模板信息结合到当前模型更新中，减轻了滤波器模型在跟踪过程中的错误积累，并且在一些严重场景中依然可以很好地跟踪上目标。

2.5 算法流程

本文跟踪算法的主要流程如表1和图2所示。采用在ImageNet数据集上训练的VGGNet-19作为特征提取器。首先移除全连接层，并使用conv1-2, conv4-4和conv5-4卷积层的输出作为深层特征。为了在每个卷积层上保留更大的空间分辨率，不使用池化层的输出。通过训练多个相关滤波器得到多个响应图，并采用固定权值进行融合。接着将构建的01二值矩阵加载在融合后的响应图上约束搜索范围，取响应最大值即为目标位置。最后依据响应值的大小变化对当前模型进行更新，用于下一帧的跟踪。

3　仿真实验

为验证本文算法的有效性，在Windows10操作系统下，采用MATLAB和C++混合编程实现本文算法，并采用了MatConvNet工具箱[21]实现对预训练深度网络的前向传播。在Intel Xeon 2.4 GHz的

表 1　基于空间可靠性约束的鲁棒视觉跟踪算法

输入：图像序列$I_1, I_2, \cdots, I_n$，目标初始位置$p_0=(x_0, y_0)$，目标初始尺度$s_0=(w_0, h_0)$。

输出：每帧图像的跟踪结果$p_t=(x_t, y_t)$, $s_t=(w_t, h_t)$。

对于$t=1, 2, \cdots, n$, do:

(1) 定位目标中心位置

(a) 利用前一帧目标位置p_{t-1}确定第t帧ROI区域，并提取其分层卷积特征；

(b) 对于每一层的卷积特征，利用式(4)和式(5)计算其相关响应图；

(c) 利用式(6)对多个相关响应图进行融合，得到最终的相关响应图；

(d)通过式(7)和式(8)提取空间可靠性区域图并将用于约束响应图搜索范围；

(e) 利用式(9)确定第t帧中目标的中心位置p_t。

(2) 确定目标最佳尺度

(a) 利用p_t和前一帧目标尺度s_{t-1}进行多尺度采样，得到采样图像集$I_s=\{I_{s_1}, I_{s_2}, \cdots, I_{s_m}\}$；

(b) 采用文献[14]中的尺度估计方法确定第t帧中目标的最佳尺度s_t。

(3) 模型更新

(a) 通过得到响应图计算最大响应值；

(b) 依据响应值大小和式(10)—式(12)对滤波器进行更新。

结束

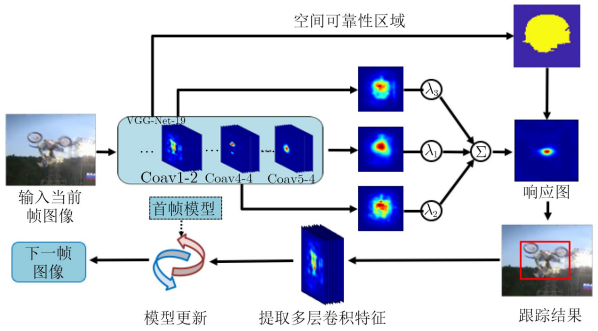


图 2 算法流程图

处理器上对本文算法进行测试，并利用了GPU (TITANX)进行加速。

测试参数设置如下：目标搜索框大小按照目标的大小进行设置；对于平移估计，将每个卷积层上的参数保持相同以用于训练相关滤波器。为了消除边界不连续性，通过余弦框滤波器对每个卷积层的提取特征通道进行加权。对于式(2)中的正则化参数$\lambda = 10^{-4}$；高斯核宽$\sigma_T = 0.1$；本文将式(10)—式(12)中的学习率设置为$\eta = 0.005$，加权系数为$\omega = 0.4$。在对多层特征得到的目标位置估计进行融合时，我们在大量的对比实验和分析的基础上，将各层的融合权重设置为$\gamma_1 = 0.5, \gamma_2 = 0.5, \gamma_3 = 1$。值得一提的是，该权值是次优折中的结果，后续将考虑自适应的方式进行权值的确定。尺度估计采用和文献[14]相同的参数设置。

3.1 在OTB2015上的性能评测

本文在OTB2015数据集[22]上将所提算法与其他4种主流且相关的跟踪算法进行比较，这些方法包括：DeepSRDCF[23], HDT[24], HCF[25], MEEM[26]。这些算法均使用默认参数。

3.1.1 算法整体性能

本文采用跟踪精度、成功率两个指标在OTB数据集上对算法的性能进行评估，图3显示了跟踪精度和成功率的对比曲线图。总的来说，本文算法取得了非常优异的性能表现。在对比的算法中，虽然DeepSRDCF在成功率上优于本文算法，但是不到1 fps/s的跟踪速度使得此类算法在实际应用中有着更大限制。另外DeepSRDCF是基于SRDCF框架的，而本文算法是在DCF上进行改进，虽然性能上有所欠缺，但是有着更好的速度优势。在比较算法中，HCF采用深度特征来来对目标进行表示，精度为0.837，成功率为0.566，HDT采用Hedge算法自适应地融合多个基于深度特征的弱跟踪器，精度达到了0.848，成功率为0.569。相比于这两个同样基于多层深度特征进行跟踪的方法，本文算法获得了更好的跟踪性能表现，精度达到了0.864，成功率为0.624，相比于HCF在精度上提升了2.7%，在成功率上提升了5.8%。

3.1.2 算法各属性性能分析

本文采用OTB100数据集中的11个标注属性对算法性能进行分析，表2和表3分别列出了各属性下算法的跟踪精度和成功率，红色代表最优结果，绿色代表次优结果，表中的字母缩写分别表示不同的跟踪条件，括号内的数字表示其包含的视频数目。11种属性分别为：尺度变化(Scale Variation, SV)、光照变化(Illumination Variation, IV)、目标遮挡(OCClusion, OCC)、背景杂波(Background Clutters, BC)、目标形变(DEFormation, DEF)、运动模糊(Motion Blur, MB)、快速运动(Fast Motion, FM)、平面内旋转(In-Plane Rotation, IPR)、平面外旋转(Out-of-Plane Rotation, OPR)、目标超出视野(Out-of-View, OV)、低分辨率(Low Resolution, LR)。为了更好地说明本文算法的有效性，重点选取同样采用多层深度特征进行跟踪的算法HCF, HDT进行对比分析。

从表2和表3可以看出，本文算法在几乎所有属性上均取得了最优的跟踪结果。尤其是在成功率上，本文算法在所有属性上均获得了提升。另外从

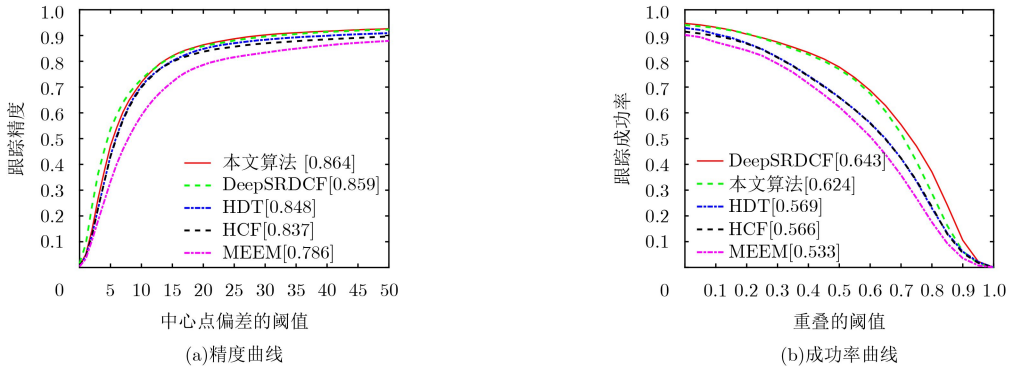


图 3 OTB100测试结果的精度曲线和成功率曲线

表 2 不同属性下算法的跟踪精度对比结果

算法	SV(60)	OCC(45)	IV(34)	BC(27)	DEF(42)	MB(29)	FM(37)	IPR(46)	OPR(57)	OV(13)	LR(8)
本文算法	0.827	0.799	0.855	0.872	0.801	0.813	0.800	0.879	0.844	0.756	0.870
HDT	0.811	0.753	0.803	0.855	0.817	0.764	0.800	0.851	0.804	0.663	0.749
HCF	0.800	0.748	0.805	0.857	0.788	0.772	0.788	0.863	0.807	0.680	0.778

表 3　不同属性下算法的跟踪成功率对比结果

算法	SV(60)	OCC(45)	IV(34)	BC(27)	DEF(42)	MB(29)	FM(37)	IPR(46)	OPR(57)	OV(13)	LR(8)
本文算法	0.580	0.594	0.635	0.627	0.570	0.624	0.609	0.605	0.597	0.556	0.510
HDT	0.491	0.528	0.540	0.593	0.546	0.545	0.549	0.557	0.533	0.541	0.376
HCF	0.490	0.526	0.547	0.602	0.532	0.557	0.550	0.599	0.534	0.542	0.383

表2和表3中，可以注意到本文算法有更多的细节表现。首先，对于背景杂波问题，本文算法相比同样基于深度特征的HCF, HDT算法都有了更好的处理能力，这表明本文的空间可靠性区域提取方法对背景杂波起到了很好地过滤作用。另外针对遮挡以及超出视野问题，在跟踪精度和成功率上，相对于HCF都有着4%～8%的性能提升，这充分证明了本文算法的随机更新策略具有很好的可用性和稳健性，可以在一定程度上缓解目标暂时丢失造成模型污染问题。另外针对旋转和低分辨率场景，主要得益于本文采用多层深度特征进行融合的方式，高层提取语义信息，底层提取细节信息，使得对于目标的表示具有很好的不变性。

3.2　在TempleColor128上的性能评测

为更加充分地验证本文算法的跟踪性能，进一步在TempleColor128[27]数据集中对算法进行评测。

TempleColor128包括128组不同跟踪场景下的彩色视频序列，本文采用和OTB数据集相同的评测指标对算法的性能进行比较。

图4显示了6个跟踪算法的精度和成功率曲线图。在这些对比算法中，C-COT在两个指标上都获得了最好结果(0.775, 0.569)。本文算法分别在精度(0.736)和成功率(0.533)上排名第2。同时，与CF2算法相比，本文算法在精度上获得了3%的性能提升，成功率上获得了5%的提升，并且显著优于KCF算法。总体而言，本文所提算法在

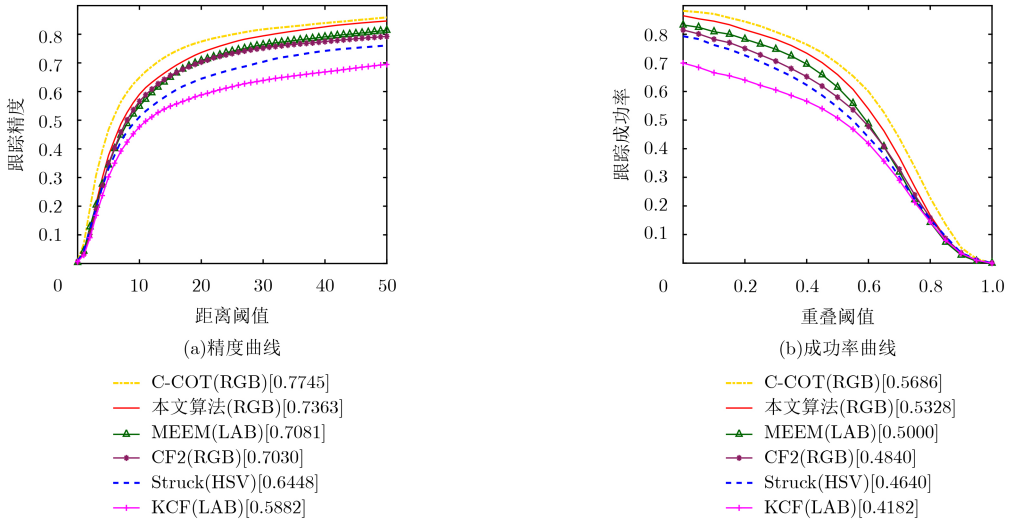


图 4 TempleColor128测试结果的精度曲线和成功率曲线

TempleColor128数据集上与最先进的跟踪算法相比，依然具有良好的性能竞争力。

3.3　各部分对跟踪性能的影响

为了进一步分析算法各部分对跟踪性能的影响，我们对所提算法进行拆分并在OTB2015数据集上做了4组对比实验，实验结果如表4所示。其中，SRCT代表本文算法，SRCT-S表示不加空间可靠性约束，SRCT-R表示不进行首帧模板的随机更新，SRCT-S-R代表以上两种策略都不采用。

表 4　算法各部分对跟踪性能影响对比实验

	SRCT	SRCT-S	SRCT-R	SRCT-S-R
成功率	0.624	0.618	0.610	0.603
跟踪精度	0.864	0.856	0.841	0.838

从表4中可以看出，单独去掉空间可靠性约束或者去掉随机更新策略都将对跟踪性能有较大影响。其中，相比于空间可靠性约束，随机更新策略对算法的性能影响更大，精度下降接近3%。

4　结束语

本文提出了一种基于空间可靠性约束的鲁棒视觉跟踪方法，在多个数据集上均取得了很好的跟踪性能表现。算法首先提取卷积网络的多层互补特征

用于训练多个相关滤波器，通过加权的方式对得到的多个响应图进行融合。同时考虑到背景杂波的影响，通过第5层的所有通道特征得到一个关于目标区域的二值矩阵，用于约束目标在响应图上的搜素范围。另外，考虑到长时遮挡以及跟踪过程中出现的漂移现象对跟踪模型的影响，采用一种随机加入首帧模型信息的方式使得跟踪模型始终保持在一个良好的状态，并可以有效地处理遮挡、背景干扰等问题。虽然本文算法取得了优于大部分算法的性能表现，但是在实验过程中，依然发现存在很多工作值得下一步继续深入地研究。

本文算法速度依然较慢，这主要可以从两个方面寻找原因，一是多层特征的提取使得特征通道数较高，在用于相关滤波算法时增加了运算量，另一方面主要是算法在处理一帧数据先后提取了两次特征，使得网络进行了两次前向传播。另外目前采用的特征提取模型大多是从分类数据库ImageNet上训练得到的。考虑到分类问题和跟踪问题的差异性，这类模型并不是处理跟踪问题的最优选择。近年来，一种基于端到端学习的Siamese网络[28,29]受到广泛的关注和大量的研究，主要在视频数据库VID上对模型进行训练，获得了较好的跟踪性能表现。目标训练网络主要是整体的方式，也可以通过对图像分块的方式进行训练[30]，提高模型的鲁棒性。

参 考 文 献

[1] 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*, 2014, 36(7): 1442–1468. doi: 10.1109/TPAMI.2013.230.

[2] WANG Naiyan, SHI Jianping, YEUNG D Y, *et al*. Understanding and diagnosing visual tracking systems[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 3101–3109. doi: 1109/ICCV.2015.355.

[3] RAWAT W and WANG Zenghui. Deep convolutional neural networks for image classification: A comprehensive review[J]. *Neural Computation*, 2017, 29(9): 2352–2449. doi: 10.1162/neco_a_00990.

[4] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 580–587. doi: 10.1109/CVPR.2014.81.

[5] SHELHAMER E, LONG J, and DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640–651. doi: 10.1109/TPAMI.2016.2572683.

[6] WANG Naiyan and YEUNG D Y. Learning a deep compact image representation for visual tracking[C]. Proceedings of the 26th International Conference on Neural Information Processing Systems, South Lake Tahoe, Nevada, USA, 2013: 809–817.

[7] 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 International Conference on Machine Learning, Lille, France, 2015: 597–606.

[8] NAM H and HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4293–4302. doi: 10.1109/CVPR.2016.465.

[9] 李寰宇, 毕笃彦, 杨源, 等. 基于深度特征表达与学习的视觉跟踪算法研究[J]. 电子与信息学报, 2015, 37(9): 2033–2039. doi: 10.11999/JEIT150031.

LI Huanyu, BI Duyan, YANG Yuan, *et al*. Research on visual tracking algorithm based on deep feature expression and learning[J]. *Journal of Electronics & Information Technology*, 2015, 37(9): 2033–2039. doi: 10.11999/JEIT150031.

[10] 侯志强, 戴铂, 胡丹, 等. 基于感知深度神经网络的视觉跟踪[J]. 电子与信息学报, 2016, 38(7): 1616–1623. doi: 10.11999/JEIT151449.

HOU Zhiqiang, DAI Bo, HU Dan, *et al*. Robust visual tracking via perceptive deep neural network[J]. *Journal of Electronics & Information Technology*, 2016, 38(7): 1616–1623. doi: 10.11999/JEIT151449.

[11] HENRIQUES J F, CASEIRO R, MARTINS P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels[C]. Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 2012: 702–715. doi: 10.1007/978-3-642-33765-9_50.

[12] DANELLJAN M, KHAN F S, FELSBERG M, *et al*. Adaptive color attributes for real-time visual tracking[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1090–1097. doi: 10.1109/CVPR.2014.143.

[13] 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. doi: 10.1109/tpami.2014.2345390.

[14] DANELLJAN M, HÄGER G, KHAN F S, *et al*. Accurate scale estimation for robust visual tracking[C]. Proceedings of British Machine Vision Conference, Nottingham, UK, 2014: 65.1–65.11. doi: 10.5244/C.28.65.

[15] DANELLJAN M, HÄGER G, KHAN F S, *et al*. Learning spatially regularized correlation filters for visual tracking[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4310–4318. doi: 10.1109/ICCV.2015.490.

[16] DANELLJAN M, ROBINSON A, KHAN F S, *et al*. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]. Proceedings of the 14th European Conference, Amsterdam, the Netherlands, 2016: 472–488. doi: 10.1007/978-3-319-46454-1_29.

[17] RUSSAKOVSKY O, DENG Jia, SU Hao, *et al*. Imagenet large scale visual recognition challenge[J]. *International Journal of Computer Vision*, 2015, 115(3): 211–252. doi: 10.1007/s11263-015-0816-y.

[18] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, USA, 2012: 1097–1105. doi: 10.1145/3065386.

[19] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. International Conference on Learning Representations, San Diego,USA,2015.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[21] VEDALDI A and LENC K. Matconvnet: Convolutional neural networks for matlab[C]. Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 2015: 689–692. doi: 10.1145/2733373.2807412.

[22] WU Yi, LIM J, and YANG M H. Object tracking benchmark[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1834–1848. doi: 10.1109/TPAMI.2014.2388226.

[23] DANELLJAN M, HÄGER G, KHAN F S, *et al*. Convolutional features for correlation filter based visual tracking[C]. Proceedings of 2015 IEEE International Conference on Computer Vision Workshop, Santiago, Chile, 2015: 58–66. doi: 10.1109/ICCVW.2015.84.

[24] QI Yuankai, ZHANG Shengping, QIN Lei, *et al*. Hedged deep tracking[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4303–4311. doi: 10.1109/CVPR.2016.466.

[25] MA Chao, HUANG Jiabin, YANG Xiaokang, *et al*. Hierarchical convolutional features for visual tracking[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 3074–3082. doi: 10.1109/ICCV.2015.352.

[26] ZHANG Jianming, MA Shugao, and SCLAROFF S. MEEM: Robust tracking via multiple experts using entropy minimization[C]. Proceedings of the 13th European Conference, Zurich, Switzerland, 2014: 188–203.

[27] LIANG Pengpeng, BLASCH E, and LING Haibin. Encoding color information for visual tracking: Algorithms and benchmark[J]. *IEEE Transactions on Image Processing*, 2015, 24(12): 5630–5644. doi: 10.1109/TIP.2015.2482905.

[28] TAO Ran, GAVVES E, and SMEULDERS A W M. Siamese instance search for tracking[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1420–1429. doi: 10.1109/CVPR.2016.158.

[29] BERTINETTO L, VALMADRE J, HENRIQUES J F, *et al*. Fully-convolutional siamese networks for object tracking[C]. European Conference on Computer Vision, Amsterdam, the Netherlands, 2016: 850–865.

[30] 侯志强, 张浪, 余旺盛, 等. 基于快速傅里叶变换的局部分块视觉跟踪算法[J]. 电子与信息学报, 2015, 37(10): 2397–2404. doi: 10.11999/JEIT150183.

HOU Zhiqiang, ZHANG Lang, YU Wangsheng, *et al*. Local patch tracking algorithm based on fast fourier transform[J]. *Journal of Electronics & Information Technology*, 2015, 37(10): 2397–2404. doi: 10.11999/JEIT150183.

蒲　磊：男，1991年生，博士生，研究方向为计算机视觉、目标跟踪.

冯新喜：男，1964年生，教授，研究方向为信息融合、模式识别.

侯志强：男，1973年生，教授，研究方向为图像处理、计算机视觉.

余旺盛：男，1985年生，讲师，研究方向为图像处理、模式识别.