基于多层卷积特征融合的目标尺度自适应稳健跟踪

王 鑫, 侯志强, 余旺盛, 金泽芬芬, 秦先祥
空军工程大学信息与导航学院, 陕西 西安 710077

摘要 针对复杂跟踪条件下目标的稳健跟踪和精确尺度估计问题,提出了一种基于多层卷积特征融合的目标尺度自适应稳健跟踪算法。算法首先利用 VGG-Net-19 深层卷积网络架构提取目标候选区域的多层卷积特征,通过相关滤波算法构建二维定位滤波器,得到多层卷积特征并进行加权融合,从而确定目标的中心位置;然后通过对目标区域进行多尺度采样,提取其梯度方向直方图特征构建一维尺度相关滤波器,确定目标的最佳尺度。实验结果表明,与 6 种当前主流跟踪算法相比,该算法取得了最好的跟踪成功率与精度,同时在跟踪过程中较好地实现了对目标快速尺度变化的自适应跟踪,且具有较快的跟踪速率。

关键词 机器视觉; 稳健跟踪; 深度学习; 卷积特征; 相关滤波; 尺度估计
中图分类号 TP391.4 **文献标识码** A
doi: 10.3788/AOS201737.1115005

Target Scale Adaptive Robust Tracking Based on Fusion of Multilayer Convolutional Features

Wang Xin, Hou Zhiqiang, Yu Wangsheng, Jin Zefenfen, Qin Xianxiang
Information and Navigation College, Air Force Engineering University, Xi'an, Shaanxi 710077, *China*

Abstract For the problems about robust tracking and precision scale estimation of the visual objects in the complex tracking conditions, a target scale adaptive robust tracking algorithm based on the fusion of multilayer convolutional features is proposed. First, the multilayer convolutional features are extracted from the target candidate area using VGG-Net-19 deep convolutional network architecture. By constructing the two-dimensional location filters by correlation filtering algorithm and fusing the multilayer convolutional features, the center location of the target is determined. Then, through the multi-scale sampling of target, the histogram of oriented gradient features are extracted to construct the one-dimensional scale filter to achieve the optimal scale estimation. The experimental results show that the proposed algorithm gains the best success rate and precision compared with the six state-of-the-art methods. Meanwhile, this algorithm achieves an adaptive tracking to the fast scale changing of target effectively, and possesses a fast tracking speed.

Key words machine vision; robust tracking; deep learning; convolutional features; correlation filtering; scale estimation
OCIS codes 150.1135; 100.4996; 100.2000; 100.2960

1 引 言

视觉跟踪是计算机视觉领域研究的热点[1-3],其广泛应用于视频监控、智能交通、无人机制导等领域。近年来,随着计算机硬件和软件的快速发展,跟踪算法的性能得到了显著的提升。但在实际跟踪环境中,由于目标受到光照变化、相似背景、尺度变化等复杂条件的影响,稳健的视觉跟踪系统的设计与应用仍然面临着严峻的挑战。

随着深度学习[4-5]技术的快速发展,其在图像分类[6]、目标识别[7]、显著性检测[8]等计算机视觉领域取得了出色的成绩。在视觉跟踪中,基于深度学习的跟踪算法在性能上逐渐超越了传统方法,成为了研究的热点方向。Wang 等[9]基于消噪自编码器原理,将离线训练与在线微调相结合,提出了深度学习跟踪器(DLT)跟踪算法;Zhang 等[10]设计了一个两层的卷积神经网络,利用局部特征映射得到目标的稳健性全局特征,提升

收稿日期: 2017-06-21; **收到修改稿日期**: 2017-07-19
基金项目: 国家自然科学基金(61473309,61703423,41601436)、陕西省自然科学基础研究计划项目(2016JM6050)
作者简介: 王 鑫(1992—),男,硕士研究生,主要从事计算机视觉、机器学习方面的研究。E-mail: wangxiin@foxmail.com
导师简介: 侯志强(1973—),男,博士,教授,主要从事计算机视觉、模式识别方面的研究。E-mail: hou-zhq@sohu.com

了算法的稳健性；Wang 等[11]提出一种基于全卷积神经网络的跟踪算法，通过转换机制联合使用多层卷积特征用于目标跟踪，得到了更加准确的跟踪结果；Ma 等[12]利用卷积神经网络的分层卷积特征，提出了一种由粗到精的跟踪框架，提升了算法的跟踪精度；Nam 等[13]通过对卷积神经网络增加目标分类层，提出了一种多域网络（MDNet）跟踪算法，在跟踪性能上取得了显著的提升。

尽管上述跟踪算法在跟踪成功率和精度上取得了显著的提升，但相比传统的基于人工特征的跟踪方法，这些算法的跟踪速度普遍较慢。与深度学习相比，基于相关滤波[14-18]的跟踪算法在跟踪速度上具有明显的优势。Bolme 等[14]提出误差最小平方和滤波器（MOSSE）跟踪算法，取得了 600 frame/s 的跟踪速度；在 MOSSE 基础上，Danelljan 等[15]提出颜色特征（CN）跟踪算法；Henriques 等[16-17]提出循环结构相关滤波跟踪器（CSK）、核相关滤波跟踪器（KCF）跟踪算法，其跟踪速度也都达到了 100 frame/s 以上。

另外，对于跟踪过程中目标的快速尺度变化问题，基于深度学习的跟踪算法没有取得令人满意的结果。而在自动驾驶、智能监控等实际应用领域，目标的尺度变化在很大程度上影响了算法的跟踪成功率。近年来，一些尺度估计方法[18-21]逐渐应用在跟踪领域，特别是文献[19]提出一种基于尺度金字塔滤波的尺度精确估计方法，在目标尺度快速变化问题上取得了不错的效果。受其启发，本文将深层卷积神经网络与相关滤波跟踪相结合，提出了一种基于多层卷积特征融合的目标尺度自适应跟踪算法。

针对视觉跟踪过程中的稳健跟踪和精确尺度估计问题，本文的主要研究工作有：1）针对复杂场景下的快速稳健跟踪问题，将深度学习与相关滤波相结合，提出了一种融合多层卷积特征的视觉跟踪算法，有效缓解了不同目标的跟踪漂移问题，提高了目标跟踪的精度；2）针对目标尺度的快速变化问题，将尺度金字塔应用到跟踪算法中，通过构造目标的一维尺度相关滤波器，从而确定目标的最佳跟踪尺度。与目前常用的粒子滤波中的尺度估计方法相比，其将目标的定位与尺度估计独立计算，减少了粒子尺度估计的随机性，避免了目标漂移造成的尺度估计不准确问题，提高了尺度估计的准确性。同时，尺度滤波器的频域计算优势提高了尺度估计的计算速度，保证了跟踪算法尺度估计的实时性；3）利用本文算法在 OTB2013[22]跟踪数据库上进行了大量的实验验证，并与 6 种近年来性能优越的主流算法进行了对比分析，实验结果表明，与同类算法相比，本文算法具有更好的跟踪性能，同时很好地解决跟踪过程目标的尺度变化问题，且具有较快的跟踪速率。

2　相关基础理论

2.1　分层卷积特征

卷积神经网络[23-24]是一种典型的深度学习架构，其受生物自然视觉认知机制启发而来，能够提取具有图像平移、旋转、形变等不变性的稳健特征。近年来，随着计算机性能的大幅度提升和训练数据集的快速发展，出现了 AlexNet[25]、VGG-Net[26]和 ResNet[27]等诸多性能优秀的卷积神经网络深层架构，直接将多维图像作为网络输入，避免了传统识别算法中复杂的特征提取和数据重建过程，广泛应用于图像处理和计算机视觉领域。

卷积神经网络作为一个多层感知器，其每个卷积层都可以得到输入图像的不同特征表达。以 VGG-Net-19 深层卷积网络为例，其中的 19 表示网络中需要学习的权重的层数。如图 1 所示，VGG-Net-19 主要由 5 组（共 16 层）卷积层（Conv1～Conv5）、2 个全连接特征层（FC-4096）和 1 个全连接分类层（FC-1000）组成。

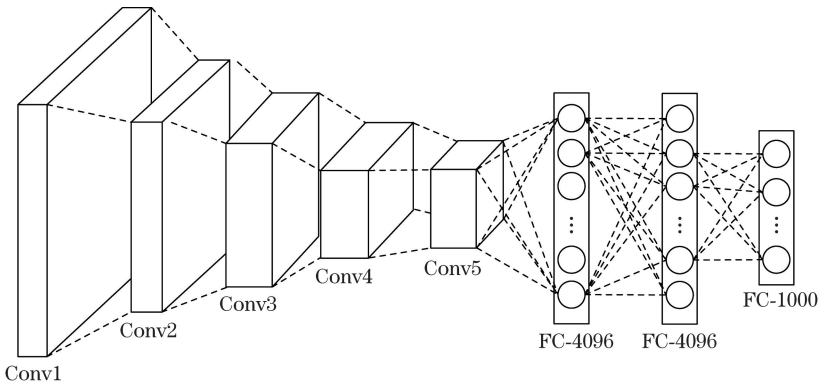

图 1　VGG-Net-19 深层卷积网络结构示意图

Fig. 1　Schematic of deep convolution network of VGG-Net-19

其中，从 Conv1 到 Conv5 的每组卷积层分别包含了 2、2、4、4、4 层卷积，所有卷积层均使用相同大小（3×3）的卷积核。通过在 ImageNet 数据集上进行训练，VGG-Net-19 中的每一个卷积层都可以得到目标的不同层级的特征表达。图 2 为 VGG-Net-19 中不同卷积层的多通道特征图的可视化表示，可以看出，层级越深，提取特征的语义信息越丰富，对不同类别的物体的分类性能越好，但对目标细节的体现越少；层级越浅，特征对目标的细节表示得越明显，但也会引入更多的背景杂波。因此，为了能更好地将卷积神经网络的卷积特征应用到视觉跟踪领域，目前很多学者将目光转向了多层卷积特征的融合与联合应用方面，取得了一些令人满意的跟踪结果[11-12]。

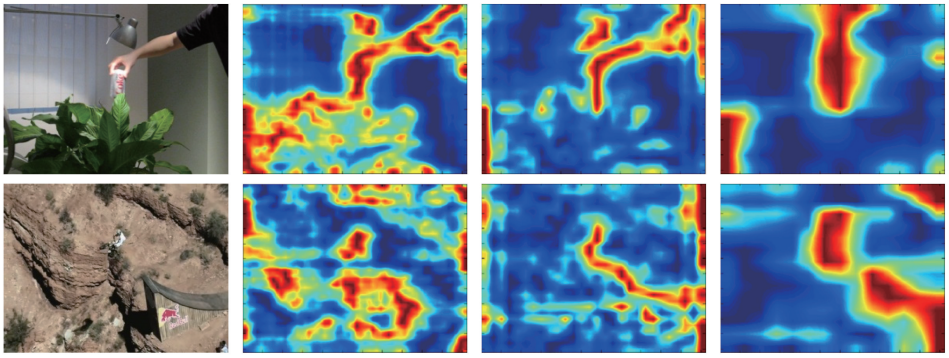

图 2　VGG-Net-19 中不同卷积层特征的可视化结果。(a) 输入图像；(b) Conv3-4；(c) Conv4-4；(d) Conv5-4

Fig. 2　Visualizations for different convolutional layers of VGG-Net-19.
(a) Input images; (b) Conv3-4; (c) Conv4-4; (d) Conv5-4

2.2　相关滤波跟踪

近年来，基于相关滤波[14-18]的跟踪算法将信号从时域转换到频域，利用快速傅里叶变换（FFT）进行滤波器训练和响应图计算，极大地提高了跟踪速度，具有很好的跟踪实时性。因此，本文将相关滤波与深层卷积神经网络相结合，显著提高了算法的跟踪速度。

相关性表示两个信号之间的联系，两个信号越相似，其相关性越大。给定跟踪目标的多维特征输入 f，基于相关滤波的跟踪算法通过学习训练数据得到一个最优相关滤波器 h^*，利用 h^* 寻找目标候选区域中的最大相关响应值来进行目标的跟踪。

在第 t 帧中，目标的多维特征输入为 $f\in\mathbf{R}^{M\times N\times D}$，将 f 沿垂直和水平方向上的所有循环移位作为训练样本，每个样本可表示为 $f_{m,n}$，$(m,n)\in\{0,1,\cdots,M-1\}\times\{0,1,\cdots,N-1\}$。给定每个样本 $f_{m,n}$ 的期望输出 $g(m,n)$，通过最小化输出误差，可以得到此时的最优相关滤波：

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

式中 λ 为正则化参数且 $\lambda\geqslant 0$，$g(m,n)$ 通常为峰值在 $f_{m,n}$ 中心位置的二维高斯核函数。令 $\varepsilon=\sum_{m,n}\left\|\sum_{d=1}^{D}h^d f^d_{m\times n}-g(m,n)\right\|^2+\lambda\|h\|_2^2$，根据帕萨瓦尔定理可以得到 ε 的频域表示：

$$\tilde{\varepsilon}=\frac{1}{MN}\left(\left\|\sum_{d=1}^{D}H^d\circ\bar{F}^d-G\right\|^2+\lambda\sum_{d=1}^{D}\|H^d\|^2\right),\tag{2}$$

式中 F、G 和 H 分别为 f、g 和 h 的离散傅里叶变换（DFT），$\bar{F}$ 为 F 的复数共轭，$\circ$ 表示元素的点乘运算。

由 $\frac{\partial\tilde{\varepsilon}}{\partial H^d}=0$，可以求得每个通道 d，$(d\in\{1,2,\cdots,D\})$ 上的最优滤波为

$$H^d=\frac{G\circ\bar{F}^d}{\sum_{i=1}^{D}F^i\circ\bar{F}^i+\lambda}。\tag{3}$$

因此，给定第 $t+1$ 帧中目标候选区域的多维特征图 z，$z\in\mathbf{R}^{M\times N\times D}$，其 DFT 变换为 Z，可以得到第 t 帧的相关响应图：

$$E=\mathscr{F}^{-1}\left(\sum_{d=1}^{D}H^d\circ\bar{Z}^d\right),\tag{4}$$

式中 $\bar{Z}$ 为 Z 的共轭复数，$\mathcal{F}^{-1}$ 表示离散傅里叶逆变换操作。在相关响应图 E 中寻找最大响应值即可以得到第 $t+1$ 帧中目标的跟踪结果。

3　本文算法

基于卷积神经网络的多层卷积特征和相关滤波算法，针对视觉跟踪中的稳健跟踪和精确尺度估计问题，提出了一种基于多层卷积特征融合的目标尺度自适应跟踪算法。通过构建定位滤波器和尺度滤波器，分别进行目标定位和尺度估计，较好地解决了目标尺度变化条件下的稳健跟踪问题。

3.1　基于多层卷积特征融合的目标定位

在跟踪过程中，利用训练好的 VGG-Net-19 卷积神经网络架构[26]，利用多层卷积特征构建定位滤波器进行目标的准确定位。

在 VGG-Net-19 架构中，每个卷积层的输出都是一组多通道的特征图 $f\in\mathbf{R}^{M\times N\times D}$，$M$、$N$ 和 D 分别表示特征图的宽、高和通道数。但因为卷积网络的池化操作，所以不同层级之间特征图的尺寸大小存在差异，层级越深，特征图的尺寸越小。因此，为了更好地融合不同层级之间的卷积特征图，利用双线性插值方法对特征图进行上采样操作，使得所有卷积层的特征图都具有相同的尺寸。

对于第 l 层的多通道卷积特征图 $f_l\in\mathbf{R}^{M\times N\times D}$，利用第 2.2 节介绍的相关滤波算法，构建每个特征通道 d 上的定位滤波器：

$$H_l^d=\frac{G_l\circ\bar{F}_l^d}{\sum_{i=1}^{D}F_l^i\circ\bar{F}_l^i+\lambda_{\mathrm{p}}},\tag{5}$$

式中 λ_{p} 为定位滤波器的正则化参数，F_l 和 G_l 分别为 f_l 和 g_l 的 DFT 变换，$\bar{F}_l$ 为 F_l 的复数共轭。这里的 g_l 是 f_l 进行循环移位采样的样本 $f_l(m,n)$ 的期望输出，为二维高斯核函数，其表达式为

$$g_l(m,n)=\exp\left[-\frac{(m-M/2)^2+(n-N/2)^2}{2\sigma_{\mathrm{p}}^2}\right],\tag{6}$$

式中 $(m\times n)\in\{0,1,\cdots,M-1\}\times\{0,1,\cdots,N-1\}$，$\sigma_{\mathrm{p}}$ 表示高斯核的宽度。

因此，给定跟踪目标的感兴趣区域(ROI)，利用 VGG-Net-19 得到其在第 l 层的卷积特征图 $z_l\in\mathbf{R}^{M\times N\times D}$，利用(4)式可以得到其在第 l 层上的相关响应图：

$$E_l=\mathcal{F}^{-1}\left(\sum_{d=1}^{D}H_l^d\circ\bar{Z}_l^d\right),\tag{7}$$

式中 $\bar{Z}_l$ 为 z_l 的 DFT 变换的共轭复数。

因为不同层级的卷积特征包含着不同层次的信息[11-12]，所以为了更加准确地实现目标的定位，分别计算 VGG-Net-19 中的 Conv3-4、Conv4-4 和 Conv5-4 卷积层的相关响应图 E_3、E_4 和 E_5，然后对其进行加权融合得到融合后的相关响应图：

$$E=\gamma_1E_5+\gamma_2E_4+\gamma_3E_3,\tag{8}$$

式中 γ_1、γ_2 和 γ_3 分别是不同卷积层对应的融合加权值。

此时，通过搜索最大响应值即可以定位当前跟踪目标的中心位置 $p_t=(x_t,y_t)$：

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

式中 (m,n) 为 ROI 中的像素点位置。

3.2　基于尺度滤波器的目标尺度估计

在确定目标的中心位置 p_t 后，利用已知的目标尺度大小 $s_{t-1}=(w_{t-1},h_{t-1})$，对目标区域进行多尺度采样[19]，构建目标尺度金字塔，并通过提取其梯度方向直方图(HOG)特征，构建尺度滤波器，进行目标尺度的估计。这里之所以选用 HOG 特征，是因为其不仅能够较好地表征特征，而且计算速度快，有助于实现跟踪过程中的快速尺度估计。而深度卷积特征虽然具有较好的稳健性能，但其计算复杂，在对多个样本进行特征提取时会占用大量时间，不适合目标的快速尺度估计。

给定尺度因子 a 和采样个数 S，对于尺度等级 $n\in\left\{\left[-\frac{S-1}{2}\right],\cdots,\left[\frac{S-1}{2}\right]\right\}$，以 p_t 为采样中心，提取目

标的多尺度图像块 J_n，其尺寸大小为 $a^n \cdot w_{t-1} \times a^n \cdot h_{t-1}$，采样过程如图 3 所示。对每个图像块 J_n 提取其 HOG 特征 $\boldsymbol{f}_s(n)$，$\boldsymbol{f}$ 它是一个 K 维的特征向量。给定其一维高斯形式的样本标签 $g_s(n)=\exp\left(-\frac{n^2}{2\sigma_s^2}\right)$，$\sigma_s$为一维高斯核宽度，可以得到尺度滤波器：

$$H_s^k=\frac{G_s\circ\bar{F}_s^k}{\sum_{k=1}^{K}F_s^k\circ\bar{F}_s^k+\lambda_s}, \tag{10}$$

式中 λ_s为尺度滤波器的正则化参数，F_s和 G_s分别为 f_s和 g_s的 DFT 变换，$\bar{F}_s$为 F_s的复数共轭。

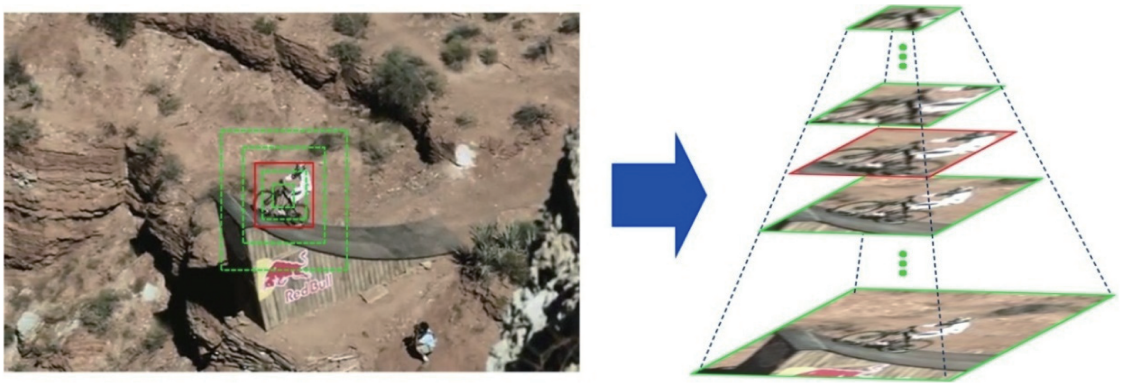

图 3　多尺度采样构建目标尺度金字塔

Fig. 3　Construct the scale pyramid of the target by multi-scale sampling

因此，在跟踪第 t 帧图像序列的目标时，已知第 t-1 帧中目标的尺度滤波器 $H_s(t\text{-}1)$，通过多尺度采样和 DFT 变换得到第 t 帧中目标的多尺度 HOG 特征 $F_s(t)$，则由(4)式可以得到其一维相关响应值：

$$E_s=\mathscr{F}^{-1}\left\{\sum_{k=1}^{K}H_s^k(t\text{-}1)\circ\bar{F}_s^k(t)\right\}。 \tag{11}$$

此时，通过寻找 E_s中的最大响应值，即可得到当前跟踪目标的最优尺度估计为

$$s_t=(w_t,h_t)=\arg\max_n E_s(n)。 \tag{12}$$

3.3　模型更新策略

目标在跟踪过程中不仅会出现形变、旋转等变化，而且还可能受到光照变化、背景遮挡等复杂条件的干扰，因此跟踪过程中必须对跟踪模型进行更新。在本文算法中，模型的更新主要体现在对相关滤波器的参数更新上。

对于定位滤波器 H_l^d，令 A_p^{t-1}和 B_p^{t-1}分别表示其在第 $t-1$ 帧时的分子项和分母项，则在第 t 帧中，定位滤波器的更新策略为

$$\begin{cases}A_p^t=(1-\eta_p)A_p^{t-1}+\eta_p G_l\circ\bar{F}_l^d(t)\\ B_p^t=(1-\eta_p)B_p^{t-1}+\eta_p\sum_{i=1}^{D}F_l^i(t)\circ\bar{F}_l^i(t)\\ H_l^d(t)=\dfrac{A_p^t}{B_p^t+\lambda_p}\end{cases}。 \tag{13}$$

同理，对于尺度滤波器 H_s，令 A_s^{t-1}和 B_s^{t-1}分别表示其在第 $t-1$ 帧时的分子项和分母项，则在第 t 帧中，尺度滤波器的更新策略为

$$\begin{cases}A_s^t=(1-\eta_s)A_s^{t-1}+\eta_s G_s\circ\bar{F}_s(t)\\ B_s^t=(1-\eta_s)B_s^{t-1}+\eta_s\sum_{k=1}^{K}F_s^k(t)\circ\bar{F}_s^k(t)\\ H_s(t)=\dfrac{A_s^t}{B_s^t+\lambda_s}\end{cases}, \tag{14}$$

式中 η_p和 η_s分别为定位滤波器和尺度滤波器的学习率。

3.4　算法整体流程

综合上述对本文算法关键部分的描述，主要跟踪步骤如表 1 所示。算法整体流程如图 4 所示。

表 1　基于多层卷积特征融合的目标尺度自适应稳健跟踪算法

Table 1　Scale adaptive robust tracker based on fusion of multilayer convolutional features

Input: Image sequence: I_1, I_2, …, I_n. Initial target position: $\boldsymbol{p}_0=(x_0, y_0)$, and initial target scale: $\boldsymbol{s}_0=(w_0, h_0)$.

Output: The estimated position of target: $\boldsymbol{p}_t=(x_t, y_t)$, and estimated scale: $\boldsymbol{s}_t=(w_t, h_t)$.

For $t=1,2,\cdots,n$, ***do***:

1	**Locate the Center of Target**
1.1	Crop out the ROI image in frame $\#t$ centered at $\boldsymbol{p}_{t-1}$, and extract the hierarchical convolutional features;
1.2	Learn the correlation response map using Eq. (5) and Eq. (7) for each convolutional layer;
1.3	Fuse the multiple correlation response maps using Eq. (8), and obtain the compositive response map;
1.4	Locate the center of the target $\boldsymbol{p}_t$ in frame $\#t$ using Eq. (9).
2	**Estimate the Scale of Target**
2.1	Obtain the multi-scale sample images $I_s=\{I_{s1},\cdots, I_{sm}\}$ in frame $\#t$ based on $\boldsymbol{p}_t$ and $\boldsymbol{s}_{t-1}$;
2.2	Build scale filters by extracting HOG features from the above multi-scale sample images;
2.3	Compute the correlation response score using Eq. (10) and Eq. (11);
2.4	Estimate the optimal scale $\boldsymbol{s}_t$ of the target in frame $\#t$ using Eq. (12).
3	**Model Update**
3.1	Update the position filters using Eq. (13);
3.2	Update the scale filters using Eq. (14).

Until End of the image sequence.

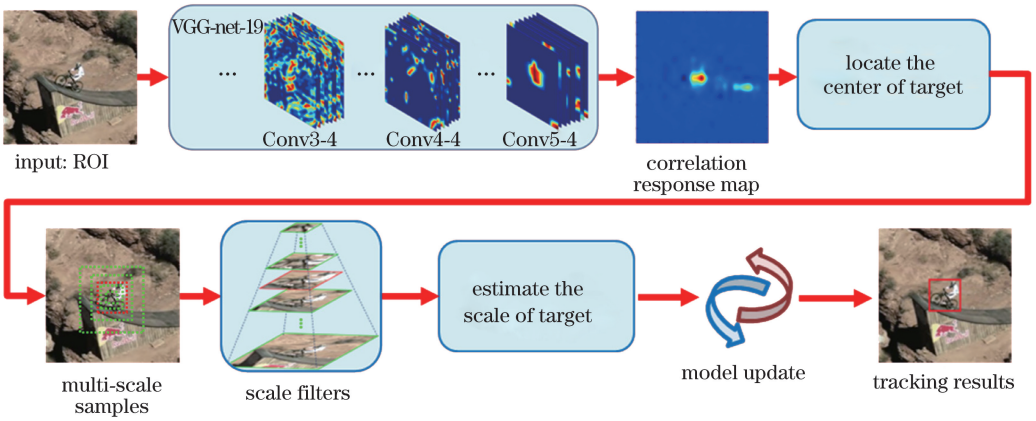

图 4　本文算法流程图

Fig. 4　Flow chart of proposed algorithm

4　实验结果与分析

在 Windows7 操作系统下，采用 MATLAB 和 C++混合编程实现本文算法。算法在实验中利用图形处理器(GPU)加速运算，实验平台为 Intel Xeon 2.4 GHz，NVIDIA GTX-TITAN X，32 GB 内存。算法参数的具体设置如下：目标 ROI 的大小为目标大小的 2.8 倍；正则化参数 $\lambda_p=10^{-4}$，$\lambda_s=10^{-2}$；高斯核宽 $\sigma_p=0.1$，$\sigma_s=0.25$；学习率 $\eta_p=0.01$，$\eta_s=0.025$。尺度滤波器的尺度因子 a 和采样个数 S 分别设定为 $a=1.02$，$S=33$。定位滤波器中，通过大量的对比实验和分析，得到融合各卷积层特征的最优加权值分别为 $\gamma_1=1$，$\gamma_2=0.5$，$\gamma_3=0.2$，该融合策略保证了算法在实验中实现了最好的跟踪性能。在实验过程中，所有参数的设置始终是固定不变的。

为充分验证本文跟踪算法的跟踪稳健性和尺度估计的准确性，对 OTB2013[22] 测试数据集中具有尺度变化的 28 组视频进行测试分析，并将本文算法与 6 种当前主流的跟踪算法进行对比分析。选取的对比算法有：卷积网络跟踪器(CNT)[10]、全卷积网络跟踪器(FCNT)[11]、分层卷积相关滤波跟踪器(HCF)[12]、判别式尺度空间

跟踪器(DSST)[19]、KCF[17]、卷积神经网络-支持向量机跟踪器(CNN-SVM)[28]。其中,CNT、FCNT、CNN-SVM、HCF 算法为基于深度学习的跟踪算法,KCF 和 DSST 为基于相关滤波的跟踪算法,FCNT 和 DSST 算法考虑了目标的尺度变化问题。这些算法均是近 3 年以来的主流跟踪算法,其跟踪结果数据均由原文提供。

4.1　定性分析

图 5 给出了实验中 7 种算法的部分跟踪结果,其中,不同跟踪算法用不同的颜色表示,红色为本文算法,左上角数字为当前图像帧数。从以下 6 个方面对算法进行定性分析。

1) 快速尺度变化。以视频 CarScale、Dog1 和 Doll 为例,目标在跟踪过程中出现了快速、剧烈的尺度变化,虽然 7 种算法都能始终跟踪目标,但只有本文算法能够很好地适应目标的尺度变化。特别是在视频 Doll 中,目标尺度大小交替变化,本文算法能很好地适应目标尺度的变化。

2) 光照变化。以视频 Ironman 和 MotorRolling 为例,跟踪过程中背景光照条件出现了剧烈的变化,要求算法对光照变化具有较好的稳健性。在视频 Ironman 中,光照的剧烈变化使得 CNT、FCNT、CNN-SVM、KCF、DSST 算法在跟踪开始就产生跟踪漂移直至失效,只有基于分层卷积特征的本文算法和 HCF 算法能够始终跟踪目标。

3) 目标旋转。以视频 MotorRolling 和 Skiing 为例,目标在跟踪过程中出现了明显的旋转变化,要求算法具有高度旋转不变性。在这 2 个视频序列中,KCF 算法和 DSST 算法都出现了跟踪失败,而本文算法和同样基于卷积神经网络跟踪的 FCNT、CNN-SVM、HCF 算法能够较好地跟踪目标。

4) 相似背景。以视频 Ironman 和 Soccer 为例,跟踪过程中出现了与目标极为相似的背景,对算法跟踪的准确性提出了挑战。在视频 Soccer 中,本文算法能够始终准确跟踪目标并且自适应地调整跟踪框的尺寸大小,而其他算法都出现了不同程度的漂移现象。

5) 目标遮挡。以视频 Soccer 和 Walking2 为例,目标在跟踪过程中被不同程度的遮挡。在视频 Soccer 第 356 帧中,当目标被部分遮挡时,FCNT 算法和 CNT 算法都出现了跟踪漂移;在视频 Walking2 第 331 帧中,当遮挡背景消失时,KCF 算法出现了跟踪漂移。而本文算法对目标遮挡问题具有较好的稳健性,能够始终准确地跟踪目标。

6) 低分辨率目标。以视频 Skiing 和 Walking2 为例,跟踪目标的尺度较小、分辨率较低,检验了算法在低分辨率条件下的特征提取能力。从这 2 个视频序列的跟踪结果可以看出,本文算法能够始终稳健地跟踪目标,对低分辨率图像具有较好的处理能力。

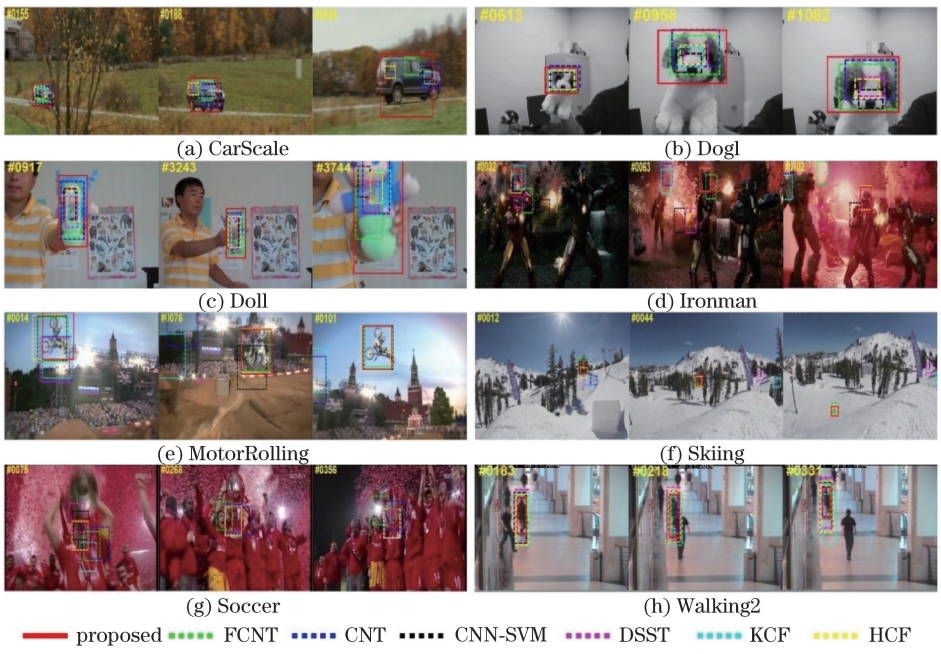

图 5　7 种跟踪算法的部分跟踪结果对比

Fig. 5　Comparison of partial tracking results of seven trackers

4.2　定量分析

从算法对单个视频的跟踪性能和对 28 组测试视频序列的综合性能两个方面对算法进行定量分析。

4.2.1　对单个视频的定量分析

针对上述 8 组视频序列，采用中心位置误差和覆盖率 2 个评价指标[22]对算法进行对比分析。中心位置误差是指跟踪结果与真实目标的中心位置的欧式距离，误差越小表示算法的跟踪精度越高。图 6 为 7 种跟踪算法在 8 组测试视频中的中心位置误差曲线。由图 6 可以看出，相比其他对比算法，本文算法的中心位置误差始终保持在较低水平，由此表明，本文算法在不同的测试视频中都能保持较高的跟踪精度，具有较好的跟踪性能。

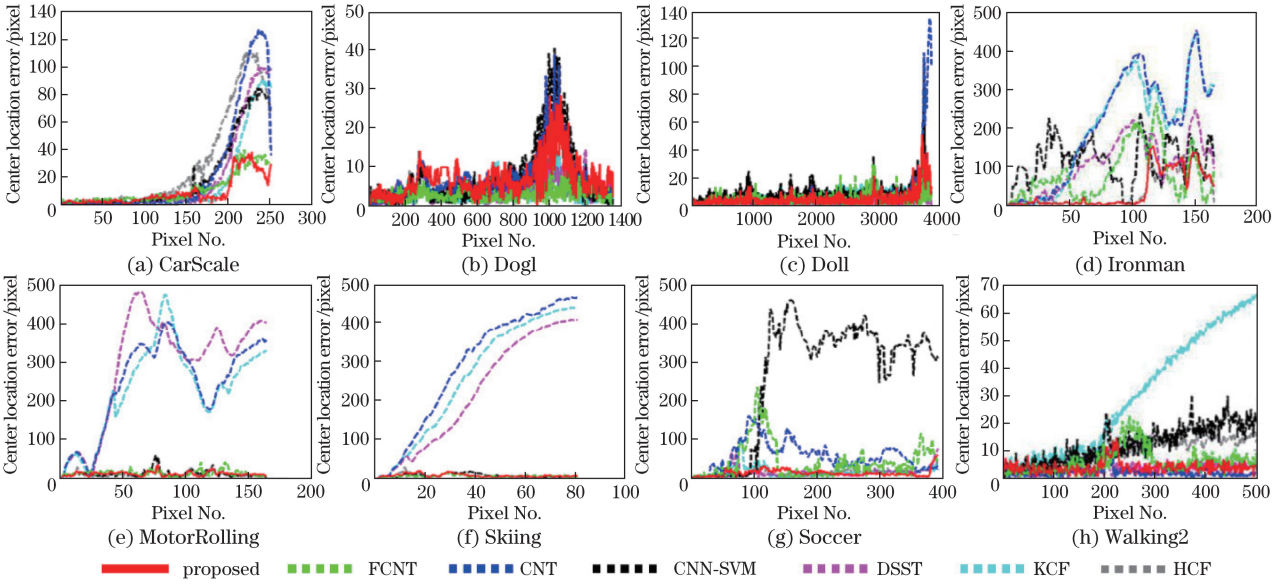

图 6　8 组测试序列的中心位置误差曲线

Fig. 6　Center location error curves of eight test sequences

覆盖率指的是跟踪结果与真实目标的重叠率，其值越大表示算法的尺度适应性越好，计算公式为 $R_{\text{overlap}}=|S_{\text{T}}\cap S_{\text{G}}|/|S_{\text{T}}\cup S_{\text{G}}|$。其中，$S_{\text{T}}$ 和 S_{G} 分别表示跟踪结果与真实目标区域，$\cap$ 和 $\cup$ 分别表示区域的交集和并集操作，$|\cdot|$ 表示计算区域的面积。图 7 为 7 种跟踪算法在 8 组测试视频中的覆盖率曲线，从曲线可以看出，本文算法在不同的测试视频中始终保持了较高的覆盖率，由此表明，本文算法在跟踪过程中具有较好的尺度适应性。

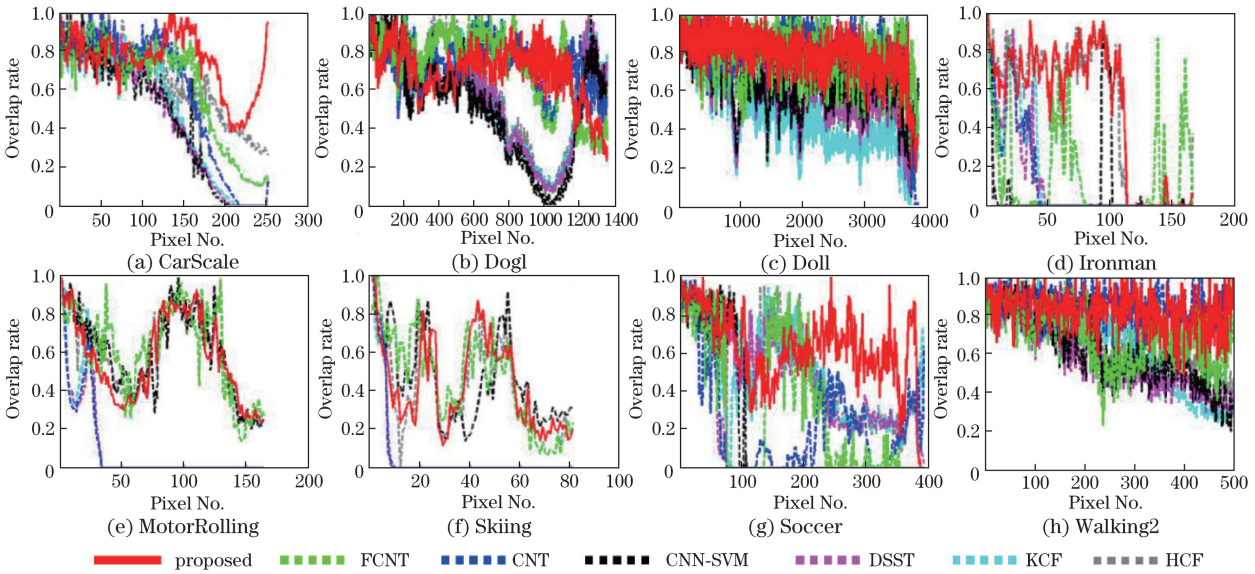

图 7　8 组测试序列的覆盖率曲线

Fig. 7　Overlap rate curves of eight test sequences

4.2.2　算法综合性能的定量分析

为综合评判算法在目标尺度变化条件下的跟踪性能，选用 OTB2013 中具有尺度变化的 28 组视频进行测试分析，并且采用跟踪成功率和跟踪精度 2 个通用的评价指标[22]来进行定量分析。其中，跟踪成功率是指当覆盖率 $R_{\text{overlap}}>0.5$ 时，算法成功跟踪的帧数与视频总帧数的比值；跟踪精度是指当平均中心位置误差小于给定阈值时，算法成功跟踪的帧数与视频总帧数的比值。

图 8 表示 7 种跟踪算法对于 28 组测试视频的整体成功率曲线和精度曲线，图例中的数字分别为成功率曲线的曲线下面积(AUC)值和平均中心位置误差为 20 pixel 时的跟踪精度值。由图 8 可以看出，对于 28 组目标尺度变化的视频，本文算法的跟踪成功率最高，且比同样基于分层卷积特征跟踪的 HCF 算法提高了 13.0%。在图 8(a)中，当覆盖率阈值处于中间范围时，本文算法的成功率高于其他对比算法，尤其是当阈值处于 0.4～0.6 时，本文算法的成功率明显高于 HCF 算法；在阈值大于 0.7 时，本文算法的成功率虽然稍逊于 CNT 算法，但仍高于 HCF 算法。这种成功率的明显提升，是因为本文算法通过对目标进行准确的尺度估计，提高了其跟踪覆盖率，从而提升了算法在同一覆盖率阈值下的成功率。对于跟踪精度，在中心位置误差阈值为 20 pixel 时，本文算法与 HCF 算法同时取得了 0.880 的最优精度值，但从图 8(b)中的曲线可以看出，当中心位置误差小于 20 pixel 时，本文算法的精度值要高于 HCF 算法，这说明了在高精度约束条件下，本文算法的稳健性比 HCF 算法更好。

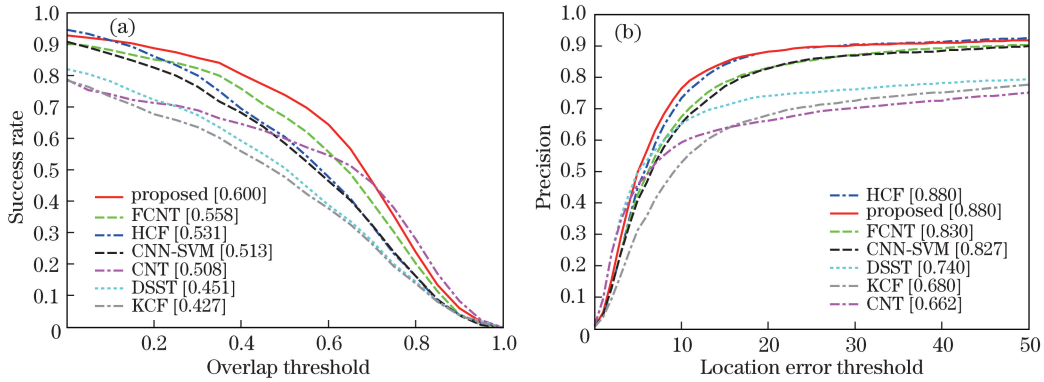

图 8　28 个测试序列的(a)成功率曲线和(b)精度曲线

Fig. 8　(a) Success rate curves and (b) precision curves of 28 test sequences

为了进一步分析跟踪算法在不同跟踪条件下的跟踪性能，表 2 和表 3 分别列出了 11 种不同属性[22]的跟踪条件下算法的跟踪精度和成功率，其中最优结果加粗表示，次优结果加下划线表示。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(SV)、光照变化(IV)、目标遮挡(OCC)、背景杂波(BC)、目标形变(DEF)、运动模糊(MB)、快速运动(FM)、平面内旋转(IPR)、平面外旋转(OPR)、目标超出视野(OV)、低分辨率(LR)。

表 2　不同属性下算法的跟踪精度对比结果

Table 2　Comparison of the tracking precisions of the algorithm of different attributes

Algorithm	SV(28)	IV(15)	OCC(16)	BC(11)	DEF(9)	MB(8)	FM(12)	IPR(18)	OPR(23)	OV(4)	LR(3)
Proposed	**0.880**	**0.838**	0.841	0.861	**0.932**	**0.870**	**0.772**	**0.879**	0.855	**0.702**	**0.873**
HCF	**0.880**	**0.858**	**0.847**	**0.867**	0.927	0.844	0.757	0.873	**0.857**	0.656	0.863
FCNT	0.830	0.779	0.737	0.713	0.925	0.740	0.715	0.774	0.798	0.691	0.686
CNN-SVM	0.827	0.751	0.733	0.689	0.890	0.725	0.685	0.793	0.800	0.650	0.606
CNT	0.662	0.521	0.667	0.463	0.686	0.479	0.477	0.583	0.630	0.481	0.410
DSST	0.740	0.681	0.785	0.610	0.733	0.635	0.539	0.714	0.725	0.453	0.402
KCF	0.680	0.632	0.744	0.578	0.734	0.679	0.586	0.619	0.678	0.639	0.233

表 3　不同属性下算法的跟踪成功率对比结果

Table 3　Comparison of the tracking success rates of the algorithm of different attributes

Algorithm	SV(28)	IV(15)	OCC(16)	BC(11)	DEF(9)	MB(8)	FM(12)	IPR(18)	OPR(23)	OV(4)	LR(3)
Proposed	**0.600**	**0.556**	**0.582**	**0.586**	**0.629**	0.591	**0.554**	**0.591**	**0.579**	0.527	**0.574**
HCF	0.531	0.509	0.514	0.573	0.589	**0.594**	0.545	0.532	0.525	0.522	0.497
FCNT	0.558	0.551	0.517	0.506	0.628	0.552	0.533	0.504	0.539	**0.573**	0.451
CNN-SVM	0.513	0.477	0.473	0.500	0.594	0.535	0.513	0.480	0.504	0.536	0.373
CNT	0.508	0.425	0.506	0.372	0.541	0.426	0.411	0.442	0.475	0.417	0.342
DSST	0.451	0.412	0.462	0.421	0.491	0.457	0.411	0.441	0.446	0.405	0.238
KCF	0.427	0.389	0.458	0.398	0.501	0.512	0.450	0.383	0.425	0.520	0.209

由表 2 和表 3 可以看出，在 11 种不同属性的跟踪条件中，本文算法的跟踪精度均处于最优或次优位置，跟踪成功率在除 OV 属性外的其他属性中也均处于最优或者次优位置。由此表明，本文算法不仅对于目标的尺度变化具有较好的跟踪性能，而且对于其他复杂条件下的跟踪也具有较好的稳健性。

4.3　算法跟踪速率

表 4 给出了本文算法在 8 组展示视频中的跟踪速率和平均速率，可以看出，目标的尺寸大小对跟踪速度有着较大的影响，目标尺寸越小，算法的跟踪速度越快。这是由于本文算法采用对目标候选区域进行卷积特征提取和相关滤波，而候选区域的大小主要受目标尺寸大小的影响，因此其跟踪速率随目标尺寸大小的不同而存在差异。

表 4　本文算法对于 8 组展示视频的跟踪速度

Table 4　Tracking speed of proposed algorithm for the eight videos　　frame /s

Video	CarScale	Dog1	Doll	Ironman	MotorRolling	Skiing	Soccer	Walking2	Average
Tracking speed	9.0	8.3	9.7	6.7	3.1	12.1	4.7	9.6	7.9

在 GPU 加速条件下，本文算法在 28 组测试视频中的跟踪速率为 2～15.5 frame/s，平均跟踪速率为 8.5 frame/s。表 5 列出了本文算法与 7 种当前主流的基于深度学习的跟踪算法的跟踪速率对比，并分别列出了其算法的编程方式和实验平台，“—”表示没有给出算法的平均速率，M 代表 MATLAB，C 代表 C/C++。从表 5 可以看出，与传统的基于深度学习的跟踪算法相比，本文跟踪算法在跟踪速率上有较大的提升，但由于其跟踪过程中要进行尺度估计，因此速度稍慢于 HCF 算法。

表 5　基于深度学习的跟踪算法的平均跟踪速率对比

Table 5　Comparison of average tracking speed of the trackers based on deep learning　　frame /s

Tracker	Proposed	CNT	FCNT	CNN-SVM	HCF	MDNet	DeepTrack[29]	STCT[30]
Code	M+C	M	M	C+M	M+C	M	M	C+M
Platform	CPU+GPU	CPU	CPU+GPU	CPU+GPU	GPU	CPU+GPU	CPU+GPU	CPU+GPU
Average tracking speed	8.5	5	3	—	10	1	2.5	2.5

4.4　算法的特征分析

为进一步分析不同卷积层特征对本文算法的影响，通过组合不同卷积层特征进行实验，可以得到算法在不同特征组合情况下的跟踪结果。图 9 展示了算法利用 VGG-Net-19 中的 Conv5-4、Conv4-4、Conv3-4、Conv2-2 卷积层进行不同组合时的跟踪成功率与精度曲线。从图 9 可以看出，本文算法利用 Conv5-4、Conv4-4 和 Conv3-4 卷积层进行跟踪，在成功率和精度上都取得了最好的跟踪结果；而当增加一层或减少一层卷积特征时，算法的跟踪性能出现了下降。实验进一步验证了本文算法中多层卷积特征融合方法的有效性和稳健性。

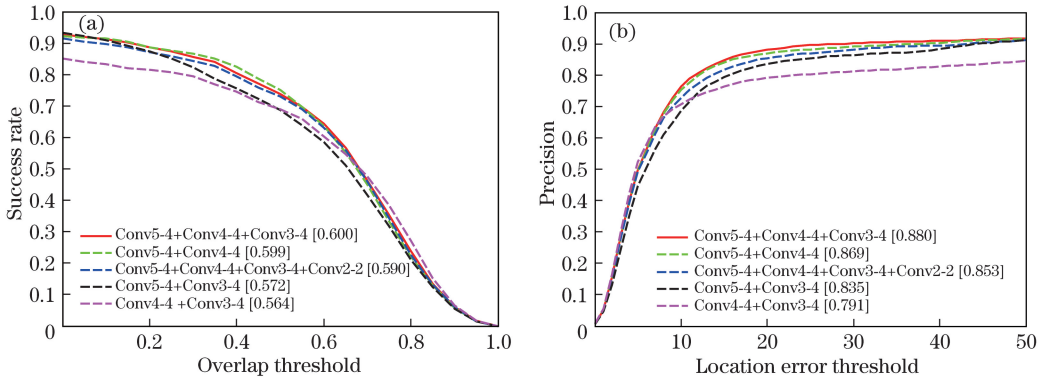

图 9　不同特征组合下的跟踪性能分析。(a)成功率曲线;(b)精度曲线

Fig. 9　Tracking performance analysis in different combinations of feature. (a) Success rate curves; (b) precision curves

5　结　　论

提出了一种基于多层卷积特征融合的目标尺度自适应稳健跟踪算法。该算法利用相关滤波算法融合多层卷积特征用于跟踪过程中的目标准确定位;同时为解决目标跟踪中的尺度变化问题,对目标候选区域进行多尺度采样,提取其 HOG 特征,构建尺度滤波器,进行目标的精确尺度估计。大量实验结果表明,在复杂跟踪条件下,与 6 种性能优越的对比算法相比,本文算法不仅得到了精度高、稳健性好的跟踪结果,而且较好地解决了跟踪过程中的目标尺度变化问题,有效提高了目标跟踪的成功率。

同时在实验中发现,对于 Lemming 序列,当在跟踪过程中出现长时间、大范围的目标遮挡时,本文算法容易出现跟踪漂移甚至失败的情况。因此,如何进一步改进算法的跟踪策略,将其与检测跟踪思想[31]相结合,以提高算法处理目标遮挡问题的能力,将是下一步研究的重点工作。

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.

[2] Hou Zhiqiang, Han Chongzhao. A survey of visual tracking[J]. Acta Automatica Sinica, 2006, 32(4): 603-617.
侯志强, 韩崇昭. 视觉跟踪技术综述[J]. 自动化学报, 2006, 32(4): 603-617.

[3] Wang N Y, Shi J P, Yeung D Y, *et al*. Understanding and diagnosing visual tracking systems[C]. IEEE International Conference on Computer Vision, 2015: 3101-3109.

[4] Schmidhuber J. Deep learning in neural networks: an overview[J]. Neural Network, 2015, 61: 85-117.

[5] Guan Hao, Xue Xiangyang, An Zhiyong. Advances on application of deep learning for video object tracking[J]. Acta Automatica Sinica, 2016, 42(6): 834-847.
管皓, 薛向阳, 安志勇. 深度学习在视频目标跟踪中的应用进展与展望[J]. 自动化学报, 2016, 42(6): 834-847.

[6] Zhang Y, Zhang E, Chen W. Deep neural network for halftone image classification based on sparse auto-encoder[J]. Engineering Applications of Artificial Intelligence, 2016, 50(1): 245-255.

[7] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection[J]. Advances in Neural Information Processing Systems, 2013, 26: 2553-2561.

[8] Wang L Z, Wang L J, Lu H C, *et al*. Saliency detection with recurrent fully convolutional networks[C]. European Conference on Computer Vision, 2016: 825-841.

[9] Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking[C]. Advances in Neural Information Processing Systems, 2013: 809-817.

[10] Zhang K H, Liu Q S, Wu Y, *et al*. Robust visual tracking via convolutional networks without training[J]. IEEE Transactions on Image Processing, 2016, 25(4): 1779-1792.

[11] Wang L J, Ouyang W L, Wang X G, *et al*. Visual Tracking with fully convolutional networks[C]. IEEE International Conference on Computer Vision, 2015: 3119-3127.

[12] Ma C, Huang J B, Yang X K, *et al*. Hierarchical convolutional features for visual tracking[C]. IEEE International Conference on Computer Vision, 2015: 3074-3082.

[13] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4293-4302.

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

[15] Danelljan M, Khan F S, Felsberg M, *et al*. Adaptive color attributes for real-time visual tracking[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1090-1097.

[16] Henriques J F, Caseiro R, Martins P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels[C]. European Conference on Computer Vision, 2012: 702-715.

[17] Henriques J F, Caseiro R, Martins P, *et al*. High-speed tracking with kernelized correlation filters[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[18] Shen Qiu, Yan Xiaole, Liu Linfeng, *et al*. Multi-scale correlation filtering tracker based on adaptive feature selection[J]. Acta Optica Sinica, 2017, 37(5): 0515001.
沈秋, 严小乐, 刘霖枫, 等. 基于自适应特征选择的多尺度相关滤波跟踪[J]. 光学学报, 2017, 37(5): 0515001.

[19] Danelljan M, Häger G, Khan F, *et al*. Accurate scale estimation for robust visual tracking[C]. British Machine Vision Conference, 2014: 1-11.

[20] Yu W S, Tian X H, Hou Z Q, *et al*. Multi-scale mean shift tracking[J]. IET Computer Vision, 2015, 9(1): 110-123.

[21] Li Shuangshuang, Zhao Gaopeng, Wang Jianyu. Distractor-aware object tracking based on multi-feature fusion and scale-adaption[J]. Acta Optica Sinica, 2017, 37(5): 0515005.
李双双, 赵高鹏, 王建宇. 基于特征融合和尺度自适应的干扰感知目标跟踪[J]. 光学学报, 2017, 37(5): 0515005.

[22] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2013, 9(4): 2411-2418.

[23] Lecun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[24] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]. European Conference on Computer Vision, 2014: 818-833.

[25] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. Computer Science, 2014, 1(2): 3.

[27] He K, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[28] Hong S, You T, Kwak S, *et al*. Online tracking by learning discriminative saliency map with convolutional neural network[C]. International Conference on Machine Learning, 2015: 597-606.

[29] Li H X, Li Y, Porikli F. Deeptrack: learning discriminative feature representations by convolutional neural networks for visual tracking[C]. British Machine Vision Conference, 2014: 1-14.

[30] Wang L J, Ouyang W L, Wang X G, *et al*. STCT: sequentially training convolutional networks for visual tracking[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1373-1381.

[31] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409-1422.