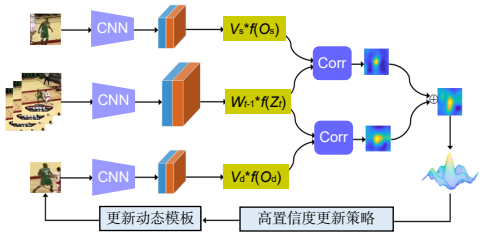
Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 4 期

DOI: 10.12086/oee.2021.200140

基于在线学习的 Siamese 网络视觉跟踪算法



张成煜[1,2]，侯志强[1,2*]，蒲　磊[3]，
陈立琳[1,2]，马素刚[1,2]，余旺盛[3]

[1]西安邮电大学计算机学院，陕西 西安 710121；
[2]西安邮电大学陕西省网络数据分析与智能处理重点实验室，陕西 西安 710121；
[3]空军工程大学信息与导航学院，陕西 西安 710077

摘要：基于 Siamese 网络的视觉跟踪算法是近年来视觉跟踪领域的一类重要方法，其在跟踪速度和精度上都具有良好的性能。但是大多数基于 Siamese 网络的跟踪算法依赖离线训练模型，缺乏对跟踪器的在线更新。针对这一问题，本文提出了一种基于在线学习的 Siamese 网络视觉跟踪算法。该算法采用双模板思想，将第一帧中的目标当作静态模板，在后续帧中使用高置信度更新策略获取动态模板；在线跟踪时，利用快速变换学习模型从双模板中学习目标的表观变化，同时根据当前帧的颜色直方图特征计算出搜索区域的目标似然概率图，与深度特征融合，进行背景抑制学习；最后，将双模板获取的响应图进行加权融合，获得最终跟踪结果。在 OTB2015、TempleColor128 和 VOT 数据集上的实验结果表明，本文算法的测试结果与近几年的多种主流算法相比均有所提高，在目标形变、相似背景干扰、快速运动等复杂场景下具有较好的跟踪性能。

关键词：视觉跟踪；Siamese 网络；双模板；快速变换学习模型

中图分类号：TP391　　**文献标志码：**A

张成煜，侯志强，蒲磊，等. 基于在线学习的 Siamese 网络视觉跟踪算法[J]. 光电工程，2021，**48**(4): 200140
Zhang C Y, Hou Z Q, Pu L, *et al*. Siamese network visual tracking algorithm based on online learning[J]. *Opto-Electron Eng*, 2021, **48**(4): 200140

Siamese network visual tracking algorithm based on online learning

Zhang Chengyu[1,2], Hou Zhiqiang[1,2*], Pu Lei[3], Chen Lilin[1,2], Ma Sugang[1,2], Yu Wangsheng[3]

[1]Institute of Computer, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[3]Information and Navigation Institute, Air Force Engineering University, Xi'an, Shaanxi 710077, China

Abstract: Visual tracking algorithm based on a Siamese network is an important method in the field of visual tracking in recent years, and it has good performance in tracking speed and accuracy. However, most tracking algorithms based on the Siamese network rely on an off-line training model and lack of online update to tracker. In order to solve

收稿日期：2020-04-24； 收到修改稿日期：2020-10-20
基金项目：国家自然科学基金资助项目(61473309，61703423)
作者简介：张成煜(1995-)，男，硕士研究生，主要从事计算机视觉、视觉跟踪的研究。E-mail：chengy_z@163.com
通信作者：侯志强(1973-)，男，博士，教授，博士生导师，主要从事图像处理、计算机视觉和信息融合的研究。E-mail：hou_qz@163.com

this problem, we propose an online learning-based visual tracking algorithm for Siamese networks. The algorithm adopts the idea of double template, treats the target in the first frame as a static template, and uses the high confidence update strategy to obtain the dynamic template in the subsequent frame; in online tracking, the fast transform learning model is used to learn the apparent changes of the target from the double template, and the target likelihood probability map of the search area is calculated according to the color histogram characteristics of the current frame, and the background suppression learning is carried out. Finally, the response map obtained by the dual templates is weighted, and the final prediction result is obtained. The experimental results on OTB2015, TempleColor128, and VOT datasets show that the test results of this algorithm are improved compared with the mainstream algorithms in recent years and have better tracking performance in target deformation, similar background interference, fast motion, and other scenarios.

Keywords: target tracking; Siamese network; dual templates; fast transformation learning model

1 引 言

视觉目标跟踪是计算机视觉领域的一个重要研究方向，其主要任务是在视频序列初始帧中给定目标位置和大小，在后续帧中预测该目标的位置和大小。其在智能监控、无人驾驶、军事侦查等领域都有广泛的应用[1-2]。视觉跟踪中，目标通常会面临尺度变化、运动模糊、目标形变、遮挡等问题，如何在这些复杂场景中准确预测目标仍然是一个具有挑战性的问题[3-4]。

近年来，目标跟踪领域取得了飞速的发展，尤其是深度学习技术的使用，使目标跟踪算法的性能得到了很大的提升，其中，基于 Siamese 网络的视觉跟踪算法因为有着较高的速度和精度，在视觉跟踪领域得到了众多研究人员的青睐。Siamese 网络能够很好地计算两路输入的相似度，在这类方法中，通常有 SiamFC[5]、SiamFC-Tri[6]、DCFNet[7]等算法，其中 SiamFC 作为大多数算法的基准算法，得到了广泛关注。在 SiamFC 中，选取初始帧中的目标作为模板，把跟踪任意目标当作一种相似性学习任务，利用离线训练的全卷积神经网络学习一个深度相似性函数，以此来预测目标的位置。

由于训练数据中存在样本不平衡问题，简单的相似匹配在复杂场景中难以对目标进行跟踪；另外，Siamese 网络能否和传统的相关滤波方法相结合，进一步提高算法的跟踪性能也是一个重要的研究方向。针对上述问题，SiamFC-Tri[6]基于 SiamFC[5]的跟踪框架提出了不同于 SiamFC 的损失函数(triplet loss)，充分利用了训练数据中模板、正样本、负样本三者之间的关系，使得网络可以提取更具表现力的特征。DCFNet[7]在 Siamese 网络中加入了相关滤波层，将相关滤波和 Siamese 网络相结合，实现了端到端的设计。SiamRPN[8]不同于传统跟踪算法中的多尺度测试和在线跟踪，它将目标检测领域中的 RPN 模块融入 Siamese 网络结构当中，在跟踪阶段将跟踪任务构造成局部单目标检测任务，有效提升了跟踪器的精度和速度。

大多数基于 Siamese 网络的跟踪算法都是使用 ILSVRC[9]和 Youtube-BB[10]等大型数据集离线训练出模型然后用于在线跟踪，缺乏类似于相关滤波框架中的在线更新。为了能够在跟踪过程中进行模型更新，Guo 等人提出了 DSiam[11]算法，该算法构建了一个动态 Siamese 网络结构，其中包含一个快速变换学习模型，在跟踪阶段能够在线学习目标的表观变化和背景抑制，但是它仍然存在以下不足：

1) 在跟踪阶段，历史帧中存在大量与目标有关的时空信息，包括形变、运动、背景等信息，而该算法只是从第一帧模板学习目标的表观变化，并没有利用到历史帧中的丰富信息，无法应对剧烈形变、目标遮挡等情况；

2) 在通过快速变换模型学习背景抑制时，在搜索区域上仅使用高斯权重图，并不能有效地凸显目标，抑制背景。

针对以上问题，本文主要做了以下工作：

首先，将第一帧中的目标当作静态模板，动态模板同样初始化为第一帧中的目标区域。然后在后续跟踪过程中，通过快速变换模型，从两个模板中学习目标的表观变化，并结合高置信度更新策略更新动态模板。其次，在背景抑制模块中，通过计算当前帧的颜色直方图特征获取搜索区域的目标似然概率图[12]，与深度特征融合，进行背景抑制学习。最后，对搜索区域和双模板进行相似性计算，将获取到的响应图加权融合求出最终响应，获得跟踪结果。

本文算法在一定程度上解决了基于 Siamese 网络跟踪算法中的模板更新问题，双模板的引入可以构建

更加稳健和丰富的目标模板，使模型在线学习到更多的目标变化；引入颜色直方图特征获取的目标似然概率图可以有效抑制背景，突出目标。在 OTB2015[13]，TempleColor128[14]和 VOT[15]数据集上的实验结果表明，与近几年的主流跟踪算法相比，本文算法的测试结果在跟踪精度和成功率上均有提升。

2 基于 Siamese 网络的跟踪算法

本文算法基于 Siamese 网络框架，在 DSiam[11]的基础上，引入动态模板和颜色直方图特征，通过快速变换模型学习目标的表观变化并进行背景抑制，有效地提升了视觉跟踪算法的鲁棒性。

基于 Siamese 网络的跟踪算法框架如图 1 所示，它包含两路并行的全卷积网络，分别对 127×127×3 的模板图像和 255×255×3 的搜索图像进行特征提取，然后对提取到的两路特征进行相似性计算。用 O_1 和 Z_t 分别表示模板图像和搜索图像，则响应图的计算：

$$S_t = \mathrm{corr}(f(O_1), f(Z_t)) \quad , \tag{1}$$

其中：corr 为相似性计算，$f(\cdot)$ 为特征提取函数，S_t 表示最终得到的响应图搜索区域中各部分与模板的相似度，选取响应最高的点作为预测的目标位置。

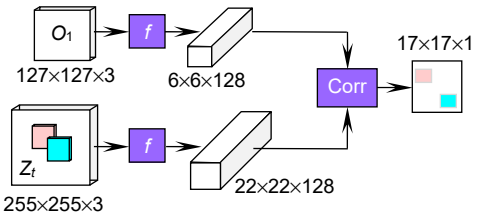


图 1 基于 Siamese 网络跟踪算法示意图

Fig. 1 Schematic diagram of tracking algorithm based on Siamese network

Simaese 网络对搜索区域进行了正负样本的区分，将搜索区域中的每一个候选窗口作为一个样本，它的输出就是其属于正负样本的概率，也就是一个二分类问题，则损失函数可表示为

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

其中：v 是候选区域的得分，$y\in\{1,-1\}$，是其真实类别。采用一个模板和一个搜索区域图像来训练 SiameseFC 网络，最终损失函数为每一个损失的均值：

$$L(y,v)=\frac{l}{|D|}\sum_{u\in D} l(y[u],v[u]) \quad , \tag{3}$$

其中：u 为响应图中的位置，标签 $y[u]\in\{1,-1\}$ 由位置 u 与响应图中心的距离确定。在某一阈值内时，将其作为正样本，否则为负样本。网络使用 ILSVRC 数据集进行离线训练，通过随机梯度下降方法调整网络参数。

为了能够应对在跟踪过程中出现的目标形变、背景干扰等情况，Guo 等人提出的 DSiam[11]算法在 SiamFC[5]的基础上构建了一个动态 Siamese 网络，通过在网络中嵌入一个快速变换模型，在每一帧学习目标的表观变化和背景抑制。DSiam[11]将目标跟踪视为一个模板匹配和在线变换学习的联合问题，在 SiamFC[5]的匹配机制上针对目标的表观变化和背景抑制引入了快速变换学习模型 V 和 W，将响应图的求解转换为

$$S_t = \mathrm{corr}(V_{t-1} * f(O_1), W_{t-1} * f(Z_t)) \quad , \tag{4}$$

其中：S_t 表示第 t 帧搜索区域与目标模板进行相关操作后得到的响应图，V_{t-1} 和 W_{t-1} 是在跟踪过程中学习到的表观变化和背景抑制，$*$ 表示循环卷积操作，$f(\cdot)$ 表示特征提取。

通常的模型更新会直接使用 $t-1$ 帧的目标区域 O_{t-1} 来替换第一帧目标模板 O_1，这种方式通常会导致模型漂移，造成目标丢失。在求解张量 $\boldsymbol{X}$ 相似于张量 $\boldsymbol{Y}$ 的最优线性变换矩阵 $\boldsymbol{R}$ 时，常使用正则化线性回归，求解方式如下：

$$\boldsymbol{R} = \arg\min_{\boldsymbol{T}} \|\boldsymbol{T} * \boldsymbol{X} - \boldsymbol{Y}\|^2 + \lambda \|\boldsymbol{T}\|^2 \quad , \tag{5}$$

其中：$\boldsymbol{T}$ 为线性变换矩阵，可以通过让第 $t-1$ 帧中目标区域特征相似于模板特征求解出变换矩阵，即可以在线学习到目标的表观变化 V_{t-1}。

$$V_{t-1} = \arg\min_{V} \|V * F_1 - F_{t-1}\|^2 + \lambda_v \|V\|^2 \quad , \tag{6}$$

其中：F_1 表示初始模板的特征，F_{t-1} 表示 $t-1$ 帧中目标区域的特征，λ_v 是正则化参数。将求解出的 V_{t-1} 与目标模板的特征 $f(O_1)$ 进行卷积操作，使模板学习到目标的表观变化。

在第 t 帧，应当选择与目标模板具有较高相似度的区域输入网络，但通常搜索区域中含有大量背景信息，为减少背景干扰，DSiam[10]引入背景抑制学习。跟踪完 $t-1$ 帧时，在该帧图像 I_{t-1} 上以目标为中心裁剪出一个与搜索区域相同大小的区域 G_{t-1}，并加上高斯权重图得到 $\overline{G}_{t-1}$，通过式(7)学习背景抑制 W_{t-1}。

$$W_{t-1} = \arg\min_{W} \left\|W * F_{G_{t-1}} - F_{\overline{G}_{t-1}}\right\|^2 + \lambda_w \|W\|^2 \quad 。 \tag{7}$$

3 本文算法

3.1 总体框架

在 DSiam[11]中，将第一帧中的目标区域作为模板，后续跟踪中利用快速变换模型从该模板中学习目标的

表观变化。但当目标出现剧烈形变，复杂背景等情况时，只从单一模板中学习这些变化容易造成跟踪失败，也没有充分利用到丰富的历史帧信息。其次，模型在学习背景抑制时，只是使用了高斯权重图，并不能有效地突出目标，抑制背景。

本文在 DSiam[11]算法的基础上进行了以下改进，其流程图如图 2 所示。

1) 在跟踪阶段，将第一帧中的目标当作静态模板，在后续帧中使用高置信度更新策略获取动态模板，使得快速变换模型可以从双模板中学习目标的表观变化；

2) 根据当前帧的颜色直方图特征计算出搜索区域的目标似然概率图，与深度特征融合，进行抑制背景学习；

3) 将搜索区域与双模板得到的响应图加权融合，得到最终响应，实现对目标的定位。

3.2 双模板的建立

3.2.1 静态模板

静态模板是跟踪过程中使用的基准模板，提供目标的初始信息，所以在视频序列的第一帧中，首先以目标位置 P 为中心裁剪一个正方形区域。假设目标的宽和高记为 w 和 h，则需要裁减的正方形区域边长：

$$S_z=\sqrt{w_z\times h_z}\quad , \tag{8}$$

其中：$w_z=w+c\times(w+h)$，$h_z=h+c\times(w+h)$，模板需要引入一定的背景信息，可以通过上式中的影响因子 c 进行调整。然后将裁剪后的区域放缩为 127×127×3 大小作为基准模板输入特征提取网络。

3.2.2 动态模板

在跟踪过程中仅使用静态模板去学习目标的表观变化无法有效应对目标形变、遮挡等情况，逐帧对静态模板进行替换又容易造成模型漂移。因此本文引入动态模板，通过在跟踪过程中逐帧计算响应图的置信度，选取置信度较高的帧作为动态模板，并结合更新策略对动态模板进行实时更新。

在 LMCF(large margin object tracking with circulant feature maps)[16]中，Wang 等人提出了平均峰值相关能量指标，可以用来衡量响应图的波动程度。文献[17]中，Chen 等人提出了一种对动态模板的更新方式。根据文献[16]和[17]，本文提出了一种高置信度更新策略，在跟踪过程中通过计算响应图的峰值和波动程度，获取当前帧的置信度 H，并依据更新策略对动态模板进行更新。通过实验，我们发现在跟踪过程中，当响应图峰值尖锐或波动程度较低时，可以很好地对目标进行定位；而当响应图波动剧烈或出现多峰时，通常会发生目标遮挡或丢失。所以在当前帧的置信度 H 和最大峰值 $V_{\max}$ 均以一定比例大于各自的历史均值 mH 和 $mV_{\max}$ 时，对动态模板进行更新。

$$H>\alpha\cdot mH\quad , \tag{9}$$

$$V_{\max}>\beta\cdot mV_{\max}\quad , \tag{10}$$

其中：参数 α 和 β 控制动态模板的更新频率，更新过快，易导致模型漂移，更新过慢，使模板无法适应目标变化。如表 1 所示，在 OTB2015[13]数据集上的实验表明，阈值 α 和 β 分别设置为 0.8 和 0.9 时，效果最优。

动态模板的引入，一方面可以构造更加稳健的目标模型，使跟踪器充分利用到丰富的历史帧的信息，另一方面，通过从双模板中学习目标的表观变化，可以有效应对跟踪中出现的目标形变、遮挡等情况。

3.3 快速变换学习模型

3.3.1 表观变化学习

虽然动态网络的引入能够在线学习目标的表观变化，但也仅仅利用了第一帧模板的信息，无法应对复杂环境下的跟踪问题。为了能够充分利用历史帧中的丰富信息，在跟踪过程中获取更多的目标变化，我们通过前面建立的双模板机制，让表观变化学习模型从静态模板和动态模板中同时学习目标的表观变化。

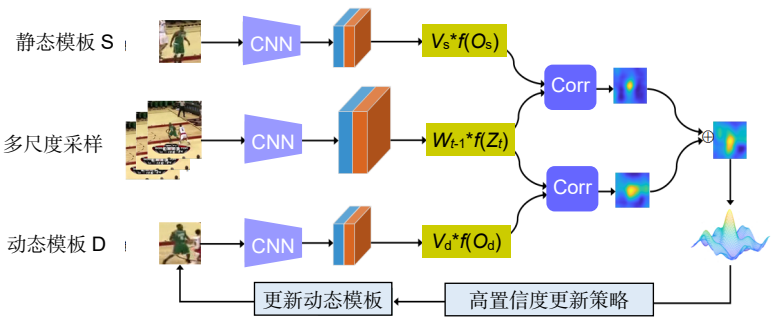


图 2 基于在线学习的视觉跟踪

Fig. 2 Visual tracking based on online learning

表 1 参数α、β的取值对成功率的影响(OTB2015)

Table 1 Influence of parameter values on success rate (OTB2015)

α	β				
	0.75	0.80	0.85	0.90	0.95
0.75	0.592	0.596	0.599	0.607	0.594
0.80	0.602	0.604	0.607	**0.612**	0.603
0.85	0.594	0.596	0.603	0.609	0.598
0.90	0.591	0.598	0.601	0.608	0.602
0.95	0.589	0.593	0.599	0.602	0.597

在跟踪完$t-1$帧后，以预测位置为中心，在$t-1$帧裁剪出与模板相同大小的区域O_{t-1}，并计算出其深度特征$f(O_{t-1})$，我们可以让$f(O_{t-1})$相似于静态模板O_s和动态模板O_d的特征$f(O_s)$和$f(O_d)$，通过求解式(11)和式(12)，学习目标表观变化V_s和V_d。

$$V_s=\arg\min_V\left\|V*F_s-F_{t-1}\right\|^2+\lambda_v\left\|V\right\|^2 \quad , \tag{11}$$

$$V_d=\arg\min_V\left\|V*F_d-F_{t-1}\right\|^2+\lambda_v\left\|V\right\|^2 \quad , \tag{12}$$

其中：$F_s=f(O_s)$，$F_d=f(O_d)$，$F_{t-1}=f(O_{t-1})$。在跟踪过程中，目标的表观变化可视为平滑变化，将学习到的目标表观变化V_s和V_d分别和$f(O_s)$和$f(O_d)$进行循环卷积操作，可以使模板适应目标形变、目标遮挡等情况。

3.3.2 背景抑制学习

在线跟踪时，为减少搜索区域中背景对跟踪器的干扰，在跟踪完$t-1$帧后，我们以预测位置为中心裁剪出一个与搜索区域大小相同的区域G_{t-1}，并计算其颜色直方图特征得到该区域的目标似然概率图，如图 3 所示。因为深度特征含有较多的语义信息，缺乏空间表征，所以将目标似然概率图与深度特征$F_{G_{t-1}}$融合，可以有效抑制背景干扰，突出目标的空间细节，弥补深度特征的不足。

将G_{t-1}划分为目标区域 O 和背景区域 B，并结合贝叶斯公式计算G_{t-1}区域内每个像素 x 属于目标区域 O 的概率：

$$P(x\in \mathrm{O}\,|\,G_{t-1},x)=\frac{P(x\in \mathrm{O}\,|\,G_{t-1})}{\sum\limits_{\Omega\in\{\mathrm{O},\mathrm{B}\}}P(x\in\Omega\,|\,G_{t-1})} \quad 。 \tag{13}$$

令 b 表示第 b 个颜色空间，b_x表示像素 x 属于第 b 个颜色空间，并计算G_{t-1}中目标和背景区域的颜色直方图$H_{\mathrm{O}}^{t-1}(b)$和$H_{\mathrm{B}}^{t-1}(b)$，根据文献[18]，该区域像素点的目标似然概率可以化简为

$$P(x\in \mathrm{O}\,|\,G_{t-1},b_x)=\frac{H_{\mathrm{O}}(b_x)}{H_{\mathrm{O}}(b_x)+H_{\mathrm{B}}(b_x)} \quad 。 \tag{14}$$

利用双线性插值对G_{t-1}区域的深度特征$P_{G_{t-1}}$进行上采样，使其与目标似然概率图 P 的尺寸相同，然后对两者逐元素相乘进行融合[19]。

$$\bar{F}_{G_{t-1}}=P\odot\uparrow F_{G_{t-1}} \quad , \tag{15}$$

其中：$\odot$ 表示逐元素相乘，$\uparrow$ 表示上采样。本文使用 AlexNet[20]作为特征提取网络，选取 Conv5 层特征与目标似然概率图融合，弥补了深层特征仅包含语义信息，缺乏空间信息的不足，有效抑制了搜索区域中的背景干扰。然后令G_{t-1}的深度特征$F_{G_{t-1}}$相似于与目标似然概率图融合后的深度特征$\bar{F}_{G_{t-1}}$，可以学习到背景抑制W_{t-1}。

$$W_{t-1}=\arg\min_W\left\|W*F_{G_{t-1}}-\bar{F}_{G_{t-1}}\right\|+\lambda_w\left\|W\right\|^2 \quad , \tag{16}$$

其中：$F_{G_{t-1}}=f(G_{t-1})$，λ_w是正则化参数。

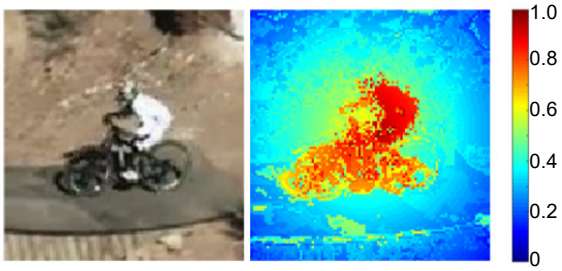


图 3 搜索区域及其目标似然概率图

Fig. 3 Search area and its target likelihood probability graph

3.4 目标定位

在线跟踪时，跟踪器可以通过快速变换模型从双模板中学习到目标的表观变化V_s和V_d，并与模板图像特征$f(O_s)$和$f(O_d)$进行循环卷积操作，由背景抑制模块学习到的W_{t-1}也与搜索区域特征$f(Z_t)$进行循环卷积操作，之后对模板分支和搜索分支提取到的特征进行相似性计算得到两个响应图，最终当前帧中目标位置由两个响应图加权融合得到：

$$R_s = \mathrm{corr}(V_s * f(O_s), W_{t-1} * f(Z_t)) \quad , \tag{17}$$

$$R_d = \mathrm{corr}(V_d * f(O_d), W_{t-1} * f(Z_t)) \quad , \tag{18}$$

$$R_t = \lambda R_s + (1-\lambda) R_d \quad , \tag{19}$$

其中：corr 表示相似性计算，$*$ 表示循环卷积操作，R_s 和 R_d 分别表示通过静态模板和动态模板得到的响应图，R_t 为当前帧加权融合后的响应图，R_t 中最高响应位置即为预测目标位置，λ 作为衡量两路网络重要性的超参数，应对静态模板赋予较大的权重，在保证跟踪性能的同时又不会导致模型漂移。在表 2 中给出了参数 λ 取不同值时对算法成功率的影响，通过大量实验并结合对响应值的综合分析，确定权重因子 λ 为 0.75。

3.5 尺度估计

获取目标定位后，以预测位置为中心进行多尺度采样，本文算法采用 3 个尺度因子，$\gamma = [0.9639, 1, 1.0375]$，并将得到的多尺度候选区域转换为 255×255×3 大小输入网络，把产生最大响应的作为最佳尺度因子。

最终目标尺度更新式：

$$(w,h) = \tau(w_{t-1}, h_{t-1}) + (1-\tau)(w_t, h_t) \quad , \tag{20}$$

其中：(w_t, h_t) 表示当前帧响应最大的尺度，(w_{t-1}, h_{t-1}) 表示前一帧目标尺度，τ 控制目标尺度的变化速度，在文献[21]算法 CREST 中，该参数被设置为 0.4，本文采用相同的参数设置进行尺度估计。

3.6 算法流程

本文算法流程如下。

输入：图像序列:I_1，I_2，…，I_n;初始目标位置:$P_0 = (x_0, y_0)$，初始目标尺度：$S_0 = (w_0, h_0)$

输出：第 t 帧预测目标位置：$P_t = (x_t, y_t)$，预测目标大小：$S_t = (w_t, h_t)$；

从 2 到 n:

1) 在线跟踪

1.1) 以上一帧预测位置 P_{t-1} 为中心在第 t 帧中裁剪出搜索区域；

1.2) 提取静态模板、动态模板和搜索区域的特征；

1.3) 通过快速变换学习模型学习目标的表观变化 V_s、V_d 和背景抑制 W_{t-1}；

1.4) 结合式(17)和式(18)对搜索区域和双模板进行相似度计算，将得到的响应图进行加权融合，得到最终响应。

2) 模板更新

2.1) 计算响应图的置信度；

2.2) 计算指标 H 和 $V_{\max}$ 的平均值 mH 和 $mV_{\max}$；

2.3) 假若 $H > \alpha \cdot mH$ 且 $V_{\max} > \beta \cdot mV_{\max}$

更新动态模板 D；

结束。

结束。

4 实验结果与分析

本文算法的实验平台为 Windows 10 系统下的 MATLAB 2017b+Visual Studio 2015，所有的实验均在配置为 Intel(R) Core(TM)i7-6850k 3.6 GHz 处理器、内存为 16 GB 的电脑上进行测试，并采用 GPU(NVIDIA GTX 1080Ti)进行加速。为验证本文算法的有效性，我们分别在三个主流的跟踪数据集上做了实验：包含 100 个视频序列(77 个彩色序列)的 OTB2015[13]、包含 128 个视频序列的 TempleColor128[14]数据集和 VOT[15]数据集。

4.1 OTB2015 实验

4.1.1 定性分析

图 4 展示了本文算法与 4 种算法的部分跟踪结果，为说明本文算法在复杂场景下的跟踪性能，我们主要从以下 4 个方面对算法进行定性分析：

1) 目标形变。跟踪过程中的目标形变会使当前帧与模板难以匹配，如视频序列 MotorRolling 和 Diving，在视频序列的前段，所有算法均可跟踪到目标，但在后续帧中，当目标出现旋转时，SiamFC、DSiam、Staple[22]等算法都会跟踪失败，而本文算法通过双模板在线学习目标的表观变化使模板及时适应目标形变，可以在一定程度上解决该问题，仍能较好地跟踪目标。

2) 目标遮挡。目标遮挡是跟踪中常见的问题，如视频序列 Liquor 和 DragonBaby，目标在运动过程中

表 2 参数λ的取值对成功率的影响(OTB2015)

Table 2 Influence of parameter λ values on success rate (OTB2015)

λ	0.70	0.75	0.80	0.85	0.90
Success rate	0.609	0.612	0.610	0.606	0.604

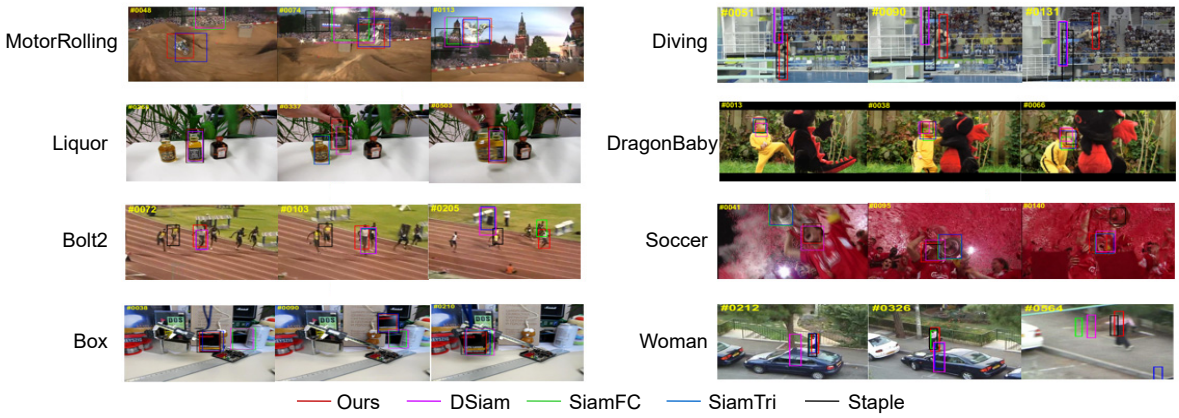


图 4　5 种算法部分跟踪结果对比

Fig. 4　Comparison of partial tracking results of 5 algorithms

会出现部分遮挡，在遮挡情况下，通常的模板更新容易使模型漂移，造成跟踪失败。本文算法通过双模板机制，并结合高置信度更新策略，在发生遮挡时避免对模板进行更新，有效避免了模型漂移。

3) 快速运动。在视频序 Bolt2 和 Soccer 中，目标的快速运动容易导致图像模糊，使目标表观发生变化。在视频 Bolt2 中，Staple[21]算法在第 72 帧就丢失目标，SiamFC[5]和 SiamTri[6]也在第 190 帧丢失目标，本文算法通过双模板在线学习目标的表观变化和背景抑制，在快速运动的情况下能够跟踪到目标。

4) 尺度变化。以视频序列 Box 和 Woman 为例，目标在跟踪过程中均发生了明显的尺度变化。在 Woman 视频 564 帧中，目标尺度明显变大，但大多数算法均跟踪失败，本文算法采用 3 个尺度因子，通过相邻帧尺度自适应策略，在保证速率的同时能够应对目标的尺度变化。

4.1.2 定量分析

在 OTB2015 数据集上，跟踪性能的主要评价指标是成功率和精确度。当预测跟踪框与视频标定跟踪框的重叠率大于一定阈值时，视为跟踪成功，成功率是指跟踪成功的帧数与总帧数的比值。目标中心位置误差由跟踪结果与人工标定的目标中心位置之间的欧氏距离确定，精确度是指目标中心位置误差小于一定阈值的帧数与总帧数的比值。

本文算法与近年来主流的跟踪算法进行了比较，包括基于深度学习的跟踪算法 SiamFC[5]、DSiam[11]、UDT[23]、SiamTri[6]和基于相关滤波的跟踪算法 SRDCF[24]、Staple[22]、DSST[25]。图 5 展示了 9 种算法在 OTB2015 数据集上的成功率曲线和精确度曲线，与基准算法相比，本文算法在 OTB2015 数据集上的精确度提高了 1.6%，成功率提高了 1.1%。

OTB2015 数据集中包含 11 种视频属性，包括尺度变化(SV)、快速运动(FM)、运动模糊(MB)、背景混杂(BC)、遮挡(OCC)、光照变化(IV)、形变(DEF)、低分辨率(LR)、平面内旋转(IPR)、平面外旋转(OPR)、目标超出视野(OV)，表 4 和表 5 分别展示了在 11 种不同属性下算法跟踪的成功率和精度。其中最优结果加粗显示，次优的算法结果由实下划线表示，排名第

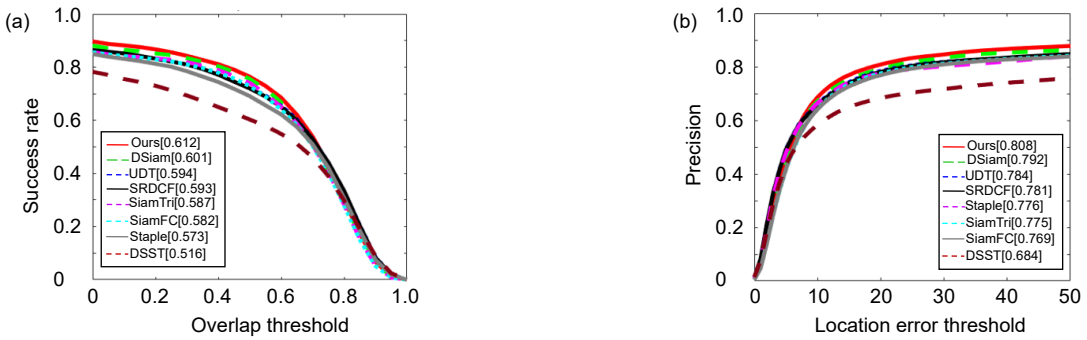


图 5　OTB2015 数据集上不同算法的成功率(a)和精度图(b)

Fig. 5　Success rate (a) and accuracy (b) of different algorithms on OTB2015 data set

表 4 不同属性下算法的跟踪成功率对比结果

Table 4 Comparsion results of tracking success of the algorithm under different attributes

Ours	**0.598**	**0.615**	**0.589**	**0.583**	**0.572**	**0.601**	0.582	0.571	**0.603**	**0.578**	0.618
SiamFC	0.551	0.556	0.579	0.564	0.507	0.570	0.552	0.515	0.545	0.470	0.582
DSiam	0.589	0.592	0.570	0.567	0.553	0.583	0.576	0.566	0.591	0.566	0.612
SiamTri	0.568	0.557	0.573	0.542	0.492	0.585	0.560	0.533	0.573	0.543	**0.627**
UDT	0.565	0.544	0.536	0.552	0.535	0.589	0.562	0.528	0.592	0.460	0.480
Staple	0.525	0.535	0.548	0.560	0.549	0.535	0.589	0.568	0.537	0.476	0.448
DSST	0.482	0.475	0.501	0.457	0.423	0.467	0.540	0.511	0.480	0.386	0.390
SRDCF	0.556	0.547	0.541	0.564	0.539	0.592	**0.596**	**0.578**	0.594	0.460	0.512

表 5 不同属性下算法的跟踪精确度对比结果

Table 5 Comparsion results of tracking accuracy of the algorithm under different attributes

Algorithm	SV	OPR	IPR	OCC	DEF	FM	IV	BC	MB	OV	LR
Ours	**0.796**	**0.816**	**0.781**	**0.772**	**0.737**	**0.777**	0.752	0.749	0.743	0.719	0.862
SiamFC	0.732	0.747	0.742	0.720	0.690	0.732	0.713	0.690	0.701	0.669	**0.875**
DSiam	0.784	0.796	0.770	0.751	0.726	0.754	0.740	0.741	0.731	0.708	0.854
SiamTri	0.752	0.752	0.739	0.714	0.718	0.761	0.713	0.695	0.714	**0.723**	0.859
UDT	0.743	0.756	0.753	0.732	0.703	0.740	0.724	0.701	0.715	0.677	0.852
Staple	0.731	0.725	0.759	0.726	0.732	0.708	0.737	0.722	0.701	0.668	0.682
DSST	0.654	0.650	0.501	0.457	0.543	0.584	0.690	0.681	0.480	0.478	0.581
SRDCF	0.739	0.571	0.742	0.735	0.726	0.758	**0.781**	**0.761**	**0.757**	0.597	0.744

三的算法结果用虚下划线表示。

4.2 TempleColor128 实验

TempleColor128 也是一个极具挑战性的数据集，它包含 128 个彩色视频序列，专门用于评估算法在颜色属性上的跟踪性能。本文算法和 8 种算法进行了对比，包括 DSiam[11]，SiamFC[5]，MEEM[26]，CFNet[27]，UDT[22]，SRDCF[23]，BACF[28]。本文算法在精确度上提高了 1.2%，成功率上提高了 0.9%，如图 6 所示。

4.3 VOT 实验

本文算法在 VOT2015[15]数据集上进行了实验。VOT(visual object tracking)测试平台是目标跟踪领域中主流的测试平台，其测试结果具有较高的权威性。VOT2015 都包含 60 个视频序列，不同于 OTB2015、TempleColor128 等测试平台，它主要从准确率、鲁棒性、期望平均重叠率等方面来评估算法性能。准确率通过计算预测目标框和标签的平均重叠率获取，鲁棒性的计算通过统计跟踪过程中的失败次数得到，基于准确性和鲁棒性，最终计算出期望平均重叠率来反应算法的性能，表 6 展示了本文算法与近几年的主流跟踪算法在 VOT2015 数据集上实验结果，与近几年的主流跟踪算法相比，在期望平均重叠率上有一定的提升。

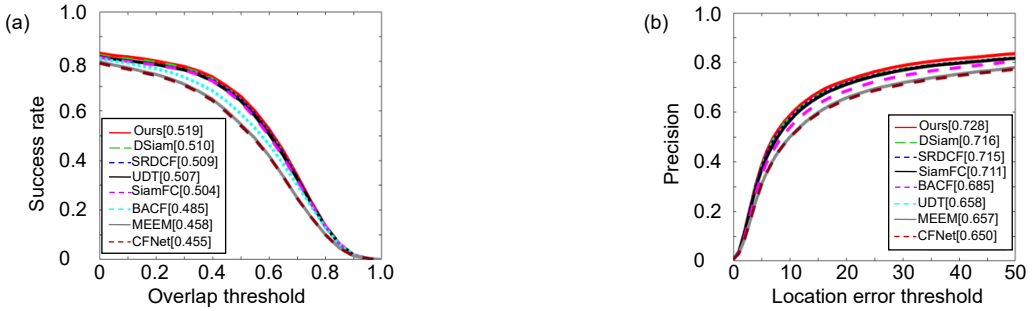


图 6 TempleColor128 数据集上不同算法的成功率(a)和精度图(b)

Fig. 6 Success rate (a) and accuracy (b) of different algorithms on TempleColor128

表 6 VOT2015 数据集上不同算法的精度和鲁棒性对比结果

Table 6 Evaluation on VOT2015 by the means of accuracy and robustness

	Ours	DSiam	HCF	SRDCF	Struck	Staple	SiamFC	LDP
Accuracy	0.65	0.59	0.45	0.56	0.47	**0.53**	0.52	0.51
Robustness	1.03	0.94	0.39	1.24	1.26	1.35	0.88	1.84
EAO	0.296	0.284	0.220	0.288	0.246	0.300	0.274	0.278

4.4 消融实验

本文在基准算法的基础上引入了双模板机制和背景抑制模块。其中双模板机制是在跟踪阶段，将第一帧中的目标当作静态模板，在后续帧中使用高置信度更新策略获取动态模板，使得快速变换模型可以从双模板中学习目标的表观变化；背景抑制模块是根据当前帧的颜色直方图特征计算出搜索区域的目标似然概率图，与深度特征融合，进行抑制背景学习。

为验证本文方法的有效性，在 OTB2015 数据集上分别对两个模块进行了实验。图 7 中，DSiam+T 表示在基准算法上只加入双模板机制，DSiam+B 表示在基准算法上只加入背景机制模块。在 OTB2015 数据集上的实验结果表明，当只在基准算法上引入双模板机制时，精度和成功率分别提高了 1.3%和 0.7%；当只在基准算法上加入背景抑制模块时，精度和成功率分别提高了 0.6%和 0.3%。通过实验，可以有效验证本文所提两个创新点的可行性，对算法的性能均有一定的提升。

4.5 算法跟踪速度

本文算法在跟踪过程中引入双模板学习目标的表观变化，并在背景抑制模块采用颜色直方图特征对跟踪目标的背景进行抑制，对算法跟踪速度有一定的影响。为保证跟踪速率，降低算法在特征提取和相关性计算上的耗时，本文并未采用逐帧更新方式，而是通过引入 APCE 指标，只在该帧响应图的 APCE 指标达到一定阈值时才对模板进行更新，降低了更新频率和特征提取次数，同时在实验中采用 GPU 加速，使本文算法的跟踪速度达到了实时性要求，速度为 29 f/s。如表 7 所示。

5 结论与展望

本文提出的基于在线学习的 Siamese 网络跟踪算法，通过静态模板和动态模板在线学习的目标的表观变化，并依据高置信度更新策略对动态模板进行更新。在背景抑制模块，通过计算搜索区域的颜色直方图特征获取目标似然概率图，与深度特征融合，进一步加强了背景抑制学习。在 OTB2015，TempleColor128 和 VOT 数据集上的实验结果表明，本文算法的测试结果与近几年的主流算法相比均有提高。在下一步工作中，我们尝试将目标表观学习和背景抑制加入其它算法，因为大多数基于孪生网络的跟踪算法仍缺乏模型的在线更新，该模块的引入可能会在一定程度上解决该问题。随着更多网络结构的提出，例如 SiamRPN[7]，SiamDW[29]，SiamRPN++[30]，TADT[31]等，针对特征提

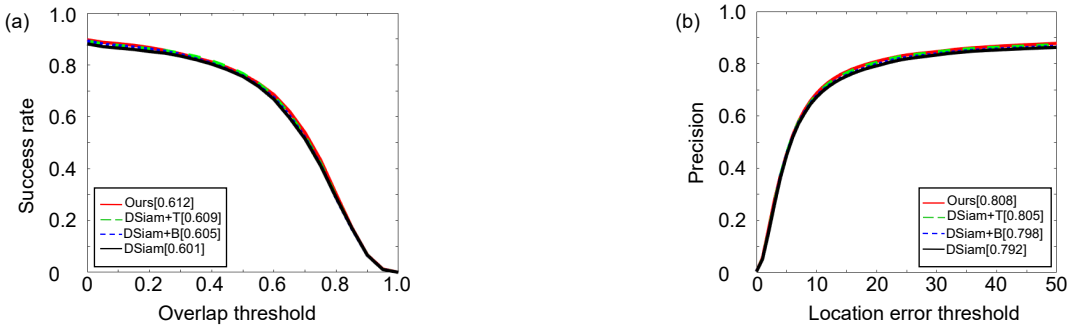


图 7 OTB2015 数据集上加入不同模块算法的成功率(a)和精度图(b)

Fig. 7 Success rate (a) and accuracy (b) of different modules are added into the algorithm on OTB2015 data set

表 7 本文算法与不同算法的跟踪速度对比

Table 7 Comparing our method with different trackers in terms of tracking speed

	Ours	DSiam	SiamFC	SRDCF	MEEM	Struck	TADT	CFNet
Speed	29	45	58	5	10	20	33	41

取，能否引入更加鲁棒的特征提取网络也将是我们下一步工作研究的重点。

参考文献

[1] Hou Z Q, Han C Z. A survey of visual tracking[J]. *Acta Automat Sin*, 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.

[2] Tang X M, Chen Z G, Fu Y. Anti-occlusion and re-tracking of real-time moving target based on kernelized correlation filter[J]. *Opto-Electron Eng*, 2020, **47**(1): 190279.
汤学猛, 陈志国, 傅毅. 基于核滤波器实时运动目标的抗遮挡再跟踪[J]. 光电工程, 2020, **47**(1): 190279.

[3] Lu H C, Li P X, Wang D. Visual object tracking: a survey[J]. *Patt Recog Artif Intell*, 2018, **31**(1): 61–76.
卢湖川, 李佩霞, 王栋. 目标跟踪算法综述[J]. 模式识别与人工智能, 2018, **31**(1): 61–76.

[4] Zhao C M, Chen Z B, Zhang J L. Research on target tracking based on convolutional networks[J]. *Opto-Electron Eng*, 2020, **47**(1): 180668.
赵春梅, 陈忠碧, 张建林. 基于卷积网络的目标跟踪应用研究[J]. 光电工程, 2020, **47**(1): 180668.

[5] Bertinetto L, Valmadre J, Henriques J F, *et al*. Fully-convolutional Siamese networks for object tracking[C]//*European Conference on Computer Vision*, Cham, 2016: 850–865.

[6] Dong X P, Shen J B. Triplet loss in Siamese network for object tracking[C]//*Proceedings of the European Conference on Computer Vision* (*ECCV*), Cham, 2018.

[7] Wang Q, Gao J, Xing J L, *et al*. Dcfnet: Discriminant correlation filters network for visual tracking[Z]. arXiv: 1704.04057v1, 2017.

[8] Li B, Yan J J, Wu W, *et al*. High performance visual tracking with siamese region proposal network[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 8971–8980.

[9] Russakovsky O, Deng J, Su H, *et al*. Imagenet large scale visual recognition challenge[J]. *Int J Comput Vis*, 2015, **115**(3): 211–252.

[10] Real E, Shlens J, Mazzocchi S, *et al*. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 5296–5305.

[11] Guo Q, Feng W, Zhou C, *et al*. Learning dynamic Siamese network for visual object tracking[C]//*Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 2017: 1763–1771.

[12] Kuai Y L, Wen G J, Li D D. Masked and dynamic Siamese network for robust visual tracking[J]. *Inf Sci*, 2019, **503**: 169–182.

[13] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. *IEEE Trans Patt Anal Mach Intellig*, 2015, **37**(9): 1834–1848.

[14] Liang P P, Blasch E, Ling H B. Encoding color information for visual tracking: Algorithms and benchmark[J]. *IEEE Trans Image Process*, 2015, **24**(12): 5630–5644.

[15] Kristan M, Matas J, Leonardis A, *et al*. The visual object tracking vot2015 challenge results[C]//*Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops*, Santiago, Chile, 2015: 1–23.

[16] Wang M M, Liu Y, Huang Z Y. Large margin object tracking with circulant feature maps[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017: 4021–4029.

[17] Hou Z Q, Chen L L, Yu W S, *et al*. Robust visual tracking algorithm based on siamese network with dual templates[J]. *J Electr Inf Technol*, 2019, **41**(9): 2247–2255.
侯志强, 陈立琳, 余旺盛, 等. 基于双模板 Siamese 网络的鲁棒视觉跟踪算法[J]. 电子与信息学报, 2019, **41**(9): 2247–2255.

[18] Possegger H, Mauthner T, Bischof H. In defense of color-based model-free tracking[C]//*Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015: 2113–2120.

[19] Xie Y, Chen Y. Adaptive object tracking based on spatial attention mechanism[J]. *Syst Eng Electr*, 2019, **41**(9): 1945–1954.
谢瑜, 陈莹. 空间注意机制下的自适应目标跟踪[J]. 系统工程与电子技术, 2019, **41**(9): 1945–1954.

[20] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//*NIPS'12*: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, New York, NY, USA, 2012.

[21] Song Y B, Ma C, Gong L J, *et al*. Crest: Convolutional residual learning for visual tracking[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 2017: 2555–2564.

[22] Bertinetto L, Valmadre J, Golodetz S, *et al*. Staple: Complementary learners for real-time tracking[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 1401–1409.

[23] Wang N, Song Y B, Ma C, *et al*. Unsupervised deep tracking[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019: 1308–1317.

[24] Danelljan M, Häger G, Khan F S, *et al*. Learning spatially regularized correlation filters for visual tracking[C]// *Proceedings of the 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015: 4310–4318.

[25] Danelljan M, Häger G, Khan F, *et al*. Accurate scale estimation for robust visual tracking[C]//*British Machine Vision Conference*, Nottingham, 2014.

[26] Zhang J M, Ma S G, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization[C]//*European Conference on Computer Vision*, Cham, 2014: 188–203.

[27] Valmadre J, Bertinetto L, Henriques J, *et al*. End-to-end representation learning for correlation filter based tracking[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017: 2805–2813.

[28] Galoogahi H K, Fagg A, Lucey S. Learning background-aware correlation filters for visual tracking[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 2017: 1135–1143.

[29] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019: 4591–4600.

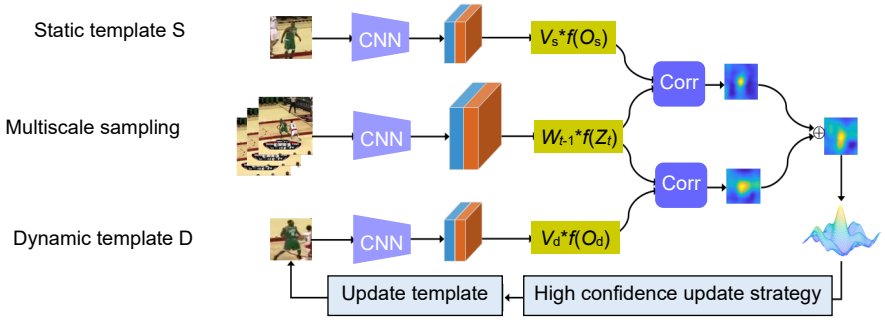
[30] Li B, Wu W, Wang Q, *et al*. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, CA, USA, 2019: 4282–4291.

[31] Li X, Ma C, Wu B Y, *et al*. Target-aware deep tracking[C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019: 1369–1378.

Siamese network visual tracking algorithm based on online learning

Zhang Chengyu[1,2], Hou Zhiqiang[1,2*], Pu Lei[3], Chen Lilin[1,2], Ma Sugang[1,2], Yu Wangsheng[3]

[1]Institute of Computer, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[2]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China;
[3]Information and Navigation Institute, Air Force Engineering University, Xi'an, Shaanxi 710077, China



Visual tracking algorithm based on online learning

Overview: Visual tracking is a fundamental challenging task in computer vision. Tracking predicts a target position in all subsequent frames given the initial frame information. It has been widely used in intelligent surveillance, unmanned driving, military detection, and other fields. In visual tracking, the target is usually faced with scale change, motion blur, target deformation, occlusion. At present, most trackers based on discriminative models include the correlation filters trackers which use hand-crafted features or CNNs and the Siamese network trackers. Visual tracking algorithm based on the Siamese network is an important method in the field of visual tracking in recent years, and it has good performance in tracking speed and accuracy. However, most tracking algorithms based on the Siamese network rely on off-line training model and lack of online update to tracker. Guo et al. proposed the DSiam algorithm, which constructed a dynamic Siamese network structure, including a fast transform learning model, and was able to learn the apparent changes and background suppression of the online target in the tracking phase. But it still has some disadvantages. Firstly, in the tracking stage, the rich information in the history frame is not used. Second, when background suppression, only a Gaussian weight graph is used in the search area, which cannot effectively highlight the target and suppress the background. In order to solve these problems, we propose an online learning-based visual tracking algorithm for Siamese networks. Main tasks as follows:

The algorithm adopts the idea of double template, treats the target in the first frame as a static template, and uses the high confidence update strategy to obtain the dynamic template in the subsequent frame.

In online tracking, the fast transform learning model is used to learn the apparent changes of the target from the double template, and the target likelihood probability map of the search area is calculated according to the color histogram characteristics of the current frame, and the background suppression learning is carried out.

Finally, the response map obtained by the dual templates is weighted and the final prediction result is obtained.

The experimental results on OTB2015, TempleColor128 and VOT datasets show that the test results of this algorithm are improved compared with the mainstream algorithms in recent years, and have better tracking performance in target deformation, similar background interference, fast motion, and other scenarios.

Zhang C Y, Hou Z Q, Pu L, *et al.* Siamese network visual tracking algorithm based on online learning[J]. *Opto-Electron Eng*, 2021, **48**(4): 200140; DOI: 10.12086/oee.2021.200140

Foundation item: National Natural Science Foundation of China (61473309, 61703423)
* E-mail: hou_qz@163.com