Opto-Electronic Engineering
光 电 工 程

Article

2020 年，第 47 卷，第 1 期

DOI: 10.12086/oee.2020.180668

基于卷积网络的目标跟踪应用研究

赵春梅[1,2]，陈忠碧[1*]，张建林[1]

[1] 中国科学院光电技术研究所，四川 成都 610209；
[2] 中国科学院大学，北京 100049

摘要： 本文针对目标跟踪应用，提出了基于 Siamese-FC 跟踪网络的改进卷积网络 Siamese-MF，意在更进一步提升跟踪速度和准确性，满足目标跟踪的工程应用需求。对于跟踪网络，考虑速度和精度的权衡，减少计算量，增加卷积特征的感受野是改进跟踪网络的速度和精度的方向。在卷积网络结构上面进行改进结构创新，改进主要集中为两点：1) 引入特征融合，丰富特征；2) 引入空洞卷积，减少计算量的同时增强感受野。Siamese-MF 算法实现了对于复杂场景目标的实时准确跟踪，在公开数据集 OTB 上测试速度达到平均 76 f/s，跟踪成功率的均值达到 0.44，而跟踪稳定性的均值达到 0.61，实时性、准确性和稳定性均提升，满足目标实时跟踪应用。

关键词： Siamese-MF；特征融合；全卷积；空洞卷积；实时跟踪

中图分类号： TP391.41　　**文献标志码：** A

引用格式： 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

Research on target tracking based on convolutional networks

Zhao Chunmei[1,2], Chen Zhongbi[1*], Zhang Jianlin[1]

[1]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: In this paper, aiming at the application of target tracking, an improved convolutional network Siamese-MF (multi-feature Siamese networks) based on Siamese-FC (fully-convolutional Siamese networks) is proposed to further improve the tracking speed and accuracy to meet the requirements of target tracking in engineering applications. For tracking networks, considering the trade-off between speed and accuracy, reducing computational complexity and increasing the receptive field of convolution feature are the directions to improve the speed and accuracy of tracking networks. There are two main points to improve the structure of convolution network: 1) introducing feature fusion to enrich features; 2) introducing dilated convolution to reduce the amount of computation and enhance the field of perception. Siamese-MF algorithm achieves real-time and accurate tracking of targets in complex scenes. The average speed of testing on OTB of public data sets reaches 76 f/s, the average value of overlap reaches 0.44, and the average value of accuracy reaches 0.61. The real-time, accuracy and stability are improved to meet the

收稿日期：2018-12-19； 收到修改稿日期：2019-03-22
基金项目：重大专项基金(G158207)
作者简介：赵春梅(1993-)，女，硕士，主要从事深度学习、机器学习和目标跟踪的研究。E-mail：841143386@qq.com
通信作者：陈忠碧(1975-)，女，博士，副研究员，主要从事运动目标检测与跟踪的研究。E-mail：ioeyoyo@126.com

requirement in real-time target tracking application.

Keywords: Siamese-MF; feature fusion; full convolution; dilated convolution; real-time tracking

Citation: Zhao C M, Chen Z B, Zhang J L. Research on target tracking based on convolutional networks[J]. *Opto-Electronic Engineering*, 2020, **47**(1): 180668

1 引 言

目标跟踪作为计算机视觉的重要方向之一[1]，应用领域广泛，包括空域目标跟踪、刑侦罪犯监控、交通车辆监控、小区安防监控等[2-4]。目标跟踪在工业应用上面临着场景复杂、目标形态多变、长时跟踪等一系列挑战，如何提取更鲁棒的特征，减少计算量是实时稳定跟踪目标的思考方向。传统的目标跟踪采用人为设定如颜色特征、灰度特征等；用核方法结合直方图特征[5]具有较好的准确性但是计算量大；用 L-K 光流算法[6]只在背景静止以及物体运动速度慢的情况下才能较好地发挥目标跟踪作用；使用均值法结合金字塔特征[7]是基于颜色对目标进行识别，跟踪速度较快，但是当背景存在和目标相似颜色的时候影响跟踪结果，使用场景受限；而核相关滤波(kernel correlation filter，KCF)[8]采用相关特征，实时性好，但是在目标发生尺度变化以及遮挡等情况时会跟丢目标。

手工特征的缺点在于提取特征有限，无法适应普遍场景，泛化能力较差[9-10]，由此引入深度学习来解决特征提取遇到的问题。2012 年卷积网络 AlexNet[11]首次被提出，后期基于卷积网络的结构相继产生，如 VGGNet[12]、Google Incepetion Net[13]、ResNet[14]、DenseNet[15]等。卷积网络往更深层发展，解决了反向传播过程中的梯度消失或梯度弥散相关问题，提取到的语义信息更丰富、更鲁棒，应用在 ImageNet[16]图像分类[17-19]、语义分割[20]、目标检测与识别[21]等方面取得显著成效，但是在目标跟踪上，却因为实时性和数据集较小等受到限制。鉴于此，本文提出基于卷积神经网络的扩展目标跟踪算法 Siamese-MF，通过对卷积层提取的特征进行融合以及引入空洞卷积的措施，达到增强特征表征并且减少计算量的作用，使算法在实时性和准确性上都有提升；在跟踪策略上进行限制，针对尺度变换、遮挡和模糊等状态具有良好的鲁棒性，有着较好的应用价值。

2 研究现状

使用深度学习的目标跟踪网络研究越来越多，总体可分为基于候选目标分类的目标跟踪和基于结构化回归的目标跟踪[22]，基于分类的目标跟踪算法类似于检测，即在搜索区域选取候选框，对候选框进行分类，计算量大，导致速度无法满足实时性要求，而基于结构化回归的算法则是通过概率来判断位置和尺度。

基于分类的目标跟踪算法(如 DLT[23])属于无监督学习算法，是深度学习在目标跟踪的早期应用，但是算法采用自编码器完成重构工作，对分类任务没有很大的贡献，是深度学习在目标跟踪任务上的探索，是基于深度学习的目标跟踪算法的启程。MDNet (multi-domain convolutional neural networks)[24]采用卷积网络提取特征，全连接网络输出背景和目标的分类得分，跟踪精度较高，但在 GPU 上运行速度也只能达到 1 f/s 左右。而 MDNet 的研究团队再次提出 TCNN(CNNs in a tree structure)[25]，算法的核心在于使用了树状 CNN 结构，模型较为复杂，精度得到提升，但速度还有待提高。

基于结构化回归的目标跟踪算法主要有 2015 年提出的 FCNT(fully convolutional networks)[26]，文章中利用大量数据集来预训练卷积网络，用于提取一般特征，然后用第一帧目标信息训练全卷积网络，用于回归目标位置。同年提出的 HCFT(hierarchical convolutional features for visual tracking)[27]结合了深度学习算法的特征提取能力和传统跟踪算法的跟踪速度。2016 年 David 提出的 GOTURN(generic object tracking using regression networks)[28]算法，采用了一种离线训练的方式，能够实现 100 f/s 的实时跟踪，但只适用于短时跟踪。而在 2016 年提出的孪生网络 Siamese-FC (fully-convolutional siamese networks)[29]，算法采用相似性来实现目标跟踪，在精度上面表现良好，速度达到 58 f/s，但有待提升卷积层特征提取能力。

本文针对扩展目标，提出了基于 Siamese-FC 的改进网络的 Siamese-MF。Siamese-MF 的卷积层为 AlexNet[11]的前五层卷积层，用于提取目标特征，同时通过空洞卷积融合第一层、第三层以及第五层卷积层的特征作为提取到的特征。卷积层的浅层特征提取边缘和位置信息，高层特征提取语义特征。Siamese-MF 结合低中高层的特征融合得到更鲁棒的特征，同时引

入空洞卷积减少计算量，增加感受野。Siamese-MF 采用离线训练，在线跟踪，通过特征融合提高跟踪准确性，空洞卷积提高速度，多尺度克服跟踪过程中的尺度变化，同时采用模板积累来适应长时跟踪。

3 Siamese-MF 跟踪网络

Siamese-FC 跟踪网络在实时性上表现较好，精度有待提高。提高精度需要提升特征鲁棒性，获取更丰富的语义特征，同时需要考虑获取特征的计算量。Siamese-MF 则是基于 Siamese-FC 改进的扩展目标跟踪网络，在提升精度的同时保证了实时性。

3.1 基础网络 Siamese-FC

Siamese-FC[29] 的跟踪结构很简单，由 Conv1~Conv5 五个卷积层提取目标和搜索区域的特征，然后通过目标特征和搜索区域特征进行全卷积得到得分图，得分最高点则为搜索区域中目标所在位置。

3.2 改进网络 Siamese-MF

本文提出的 Siamese-MF 网络的前馈网络通过训练得到，在训练过程中使用 AlexNet[11]的 5 个卷积层 Conv1~Conv5，同时添加空洞卷积层[30]Skip1 和 Skip2 分别用于 Conv1 和 Conv3 的输出和 Conv5 的特征融合。空洞卷积的作用是在不经过下采样损失信息的情况下，增加特征感受野，保证卷积后的特征包含较大范围的信息。与普通卷积相比，空洞卷积的卷积核在普通卷积核的基础上补零，也就是稀疏化的普通卷积核，这样计算量减少，信息增加，特征尺度减小。对于卷积层的选择，考虑因素有计算量和特征充分性。低层卷积层表示边缘信息和位置信息，所提取图像当中的线特征和边缘特征，属于底层信息。高层卷积层所提取图像当中的语义信息，属于高层特征。在目标检测和目标识别等操作中一般卷积层较深，能够提取足够的特征用于分类；而对于卷积层越深的网络，计算量越大，实时性则越差。本文的目标跟踪课题，由于只需要提取目标特征，不用于分类，即对于目标的语义特征要求不高。本文选择五层卷积，对 Conv1、Conv3 和 Conv5 的输出进行特征融合，获取更丰富的目标特征。

对于跟踪过程，将目标模板和搜索区域通过卷积层提取特征，再通过全卷积层进行交叉相关分析，得到目标在搜索区域的得分图，最大值所在即为目标所在位置。改进总结为以下三点：

1) 使用特征融合获取更全面的特征。将 Conv1、Conv3 和 Conv5 的输出进行特征融合，获取更丰富的目标特征，提高跟踪精度。

2) 引入空洞卷积。空洞卷积在增加感受野的同时减少计算量，提升跟踪速度和精度。

3) 设置模板更新规则，适应于长时跟踪。

3.2.1 Siamese-MF 网络结构

Siamese-MF 的前馈网络结构如图 1 所示，通过训练得到卷积层参数。网络的输入为预处理后的图片，包括目标模板和搜索区域，搜索区域为上一帧目标所在位置的 2×2 倍区域，通过对原始图片进行候选框裁剪以及尺寸变换后得到尺寸为 127×127×3 的模板和 255×255×3 的搜索区域，经过卷积层进行特征提取，最后经过全卷积得到目标和搜索区域的相关性图。

Siamese-MF 网络的前馈网络每一层的操作参数以及操作结果如表 1 所示，包括目标和搜索区域的输入、卷积尺度、步长以及输出。卷积层提取目标以及搜索区域的特征。在连接层 Skip1 和 Skip2 中加入空洞卷积[30]，将 Conv1 和 Conv3 的输出和 Conv5 的输出匹配，增加感受野的同时减少计算量。

表 1 中涉及的符号较多，此处举例说明，“Input：3@127×127”是指输入通道数为 3，输入图像大小为 127×127；“Filter_size:96@11×11”是指卷积核数目为 96，卷积核大小为 11×11；“Filter_size:32@3×3+3>7×7”是指

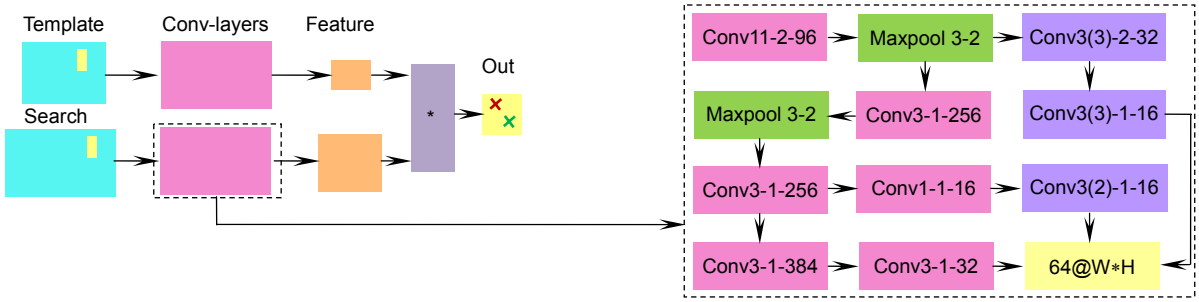


图 1 Siamese-MF 前馈网络

Fig. 1 Feedforward network of Siamese-MF

表 1 Siamese-MF 网络操作以及结果

Table 1 Operation and results of Siamese-MF network

Operation	Input		Filter_size	Stride	Out	
Conv1	3@127×127	3@255×255	96@11×11	2	96@59×59	96×123×123
Maxpooling	96@59×59	96×123×123	96@3×3	2	96@29×29	96@61×61
Skip1	96@29×29	96@61×61	32@3×3+3>7×7	2	32@12×12	32@28×28
	32@12×12	32@28×28	16@3×3+3>7×7	1	16@6×6	16×22×22
Conv2	96@29×29	96@61×61	256@5×5	1	256@25×25	256×57×57
Maxpooling	256@25×25	256×57×57	256@3×3	2	256@12×12	256×28×28
Conv3	256@12×12	256×28×28	384@3×3	1	384@10×10	384×26×26
Skip2	384@10×10	384×26×26	16@1×1	1	16@10×10	16×26×26
	16@10×10	16×26×26	16@3×3+2>5×5	1	16@6×6	16×22×22
Conv4	384@10×10	384@26×26	384@3×3	1	384@8×8	384×24×24
Conv5	384@8×8	384×24×24	32@3×3	1	32@6×6	32×22×22

空洞卷积的参数，卷积核数目为 32，卷积核大小为 3×3，空洞为 3，得到实际卷积核尺寸为 7×7；“Output：96@59×59”是指输出通道数为 96，输入图像大小为 59×59。

3.2.2 Siamese-MF 算法

Siamese-MF 网络的作用流程为：将目标和搜索区域通过相同的卷积网络，得到目标特征和搜索区域的特征，对目标特征层和搜索区域特征层进行全卷积，得到目标在搜索区域的相关性图，相关性最大的位置即为目标在搜索区域中的位置。根据视频中扩展目标的运动方位设定搜索区域为目标框的 2×2 倍大小。同时深层的卷积网络导致位置信息丢失，这是在跟踪领域不愿意看到的，所以本文在卷积层中使用了特征融合，结合了浅层的位置信息和深层的语义信息。

本文的训练是在 ILSVRC2015 数据集上进行，前馈网络中的卷积层采用 AlexNet 的 Conv1~Conv5 层。在跟踪视频序列中目标一般不会太大，故目标输入尺寸设定为 127×127，而二倍于目标模板的搜索区域的输入尺寸设定为 255×255。通过卷积层后分别得到 6×6 和 22×22 的特征，进行全卷积之后得到 17×17 的相关图。

为了节省参数的调试时间，本文网络的训练参数初始值采用 Siamese-FC 的训练参数，经过参数调试，最终确定使用随机梯度下降法(SGD)进行训练，Momentum 为 0.9，Weight Decay 为 0.0005，Learning Rate 为 0.0001。

在训练过程中，主要是获取卷积层的参数。参数的获取是通过得分图与标签得分的误差进行反向传播，从而修改卷积层的权值。记 y 为正负样本的真实标签，将正样本标签取值为 1，负样本标签取值为 0。同时设置一个距离参数，距离为以搜索区域的目标中心位置为圆心的一个半径，搜索区域中大于该值设标签 y 为 0，小于该值 y 设为 1：

$$y=\begin{cases}1, & d<R\\ 0, & d>R\end{cases} 。 \tag{1}$$

以 f 为训练中得分图的置信结果，那么 $l(y,f)$表示单张图片的逻辑损失函数：

$$l(y,f)=\log(1+\exp(-y.f)) 。 \tag{2}$$

对每一组训练图片的损失(loss)，给出如下式：

$$loss=\sum\log(1+\exp(-y.f)) 。 \tag{3}$$

通过 SGD 最小化误差进行优化从而得到网络参数为

$$\arg\min_{(W)}(loss) 。 \tag{4}$$

训练集来自于 ILSVRC2015 数据集，训练次数为 50 次。对于每一个训练视频选取 16 张图片进行训练，每一张图片都有对应的标签，用于求训练误差和损失。

算法 1 给出了 Siamese-MF 算法前馈网络训练过程，采用伪代码的方式给出网络结构、用于训练的数据、损失计算参数、卷积层需要更新的权值以及更新速率、循环次数、网络的输出结果等。

算法 1 Siamese-MF 训练过程算法流程

```
if loop_time<50:
input data_set =' ILSVRC2015' with16 samples ;
compute loss and update {w1,…,w7} with learning_rate =0.0001, then SGD with momentum=0.9 and weight_dacay=0.0005
else:
    out pretrained Conv1~Conv5 filters{w1,…,w5}, Skip1~Skip2 filters{w6,w7 };
```

3.3 跟踪

3.3.1 多尺度跟踪策略

对于已经训练好的网络，固定网络参数不变，作为前馈网络用于扩展目标跟踪。跟踪过程中，对于输入的视频序列，截取上一帧的目标当作模板，并且以上一帧目标位置中心为标准，在当前帧以该中心位置为中心，截取目标尺寸三个尺度的 2×2 倍区域作为搜索区域。跟踪时将多个搜索区域特征与模板特征进行全卷积，得到三个相关图。选取最大相关值的尺度 $s_{\max}$ 为当前目标尺度，以最大相关值尺度的最大相关值所在位置 $p_{\max}$ 为当前帧目标位置。

$$s_{\max}=\max(\max_{f_i}(f)) \quad , \tag{5}$$

$$p_{\max}=\max(f(s_{\max})) \quad 。 \tag{6}$$

3.3.2 长时跟踪策略

基于相关跟踪算法存在一个缺陷；如果出现目标消失的情况，那么提取到的当前帧模板就为空；如果直接使用第一帧模板作为当前帧跟踪模板，由于时间的漂移和目标形态变换，模板特征与当前帧目标特征会有所差异。即是说：长时跟踪存在一定挑战。本文采用模板累积的方式在时间序列上对模板特征进行累积，设置权值 w 为当前帧模板(T_{new})的权值，而 $(1-w)$ 为前一帧模板(T_{last})的权值，从而模板具有一定的时序效应，适用于目标消失或者长时跟踪的情况，累积式：

$$T_{\text{new}}=w\cdot T_{\text{new}}+(1-w)T_{\text{last}} \quad 。 \tag{7}$$

图 2 给出了跟踪过程的算法流程，包括 Siamese-MF 的图片预处理、特征提取、交叉相关、坐标回归，输出目标响应最大的位置。

4 实 验

硬件实验环境：Intel Core i7-6700 CPU@3.40 GHz×8，GeForce GTX 1080 GPU。

软件实验环境：Linux Ubuntu 16.04，python 3.5，Pytorch3.0。

本文提出的 Siamese-MF 算法需要验证两个方面。第一，在 Siamese-FC 上面性能的提升，为此将对公开数据集 OTB2015 进行测试，作为对比实验操作；第二应用价值，将对 ILSVRC2015 数据集中的飞机测试集进行测试，验证算法在飞机目标跟踪上面的性能表现。其中，测试视频序列中包含不同扩展目标的尺度变换、旋转、运动方向改变等干扰因素，以及隐形、遮挡、光照变化等多个复杂情况，有利于验证 Siamese-MF 算法的实用价值。

4.1 评价参数

为了显示 Siamese-MF 算法的良好性能，将测试集同时经过 Siamese-MF 和 Siamese-FC 算法进行测试。本文使用与 Siamese-FC 评价指标相同的三个指标[31]：

1) 跟踪成功率(Overlap)

用表示 S_{Overlap}，其定义：

$$S_{\text{Overlap}}=\frac{R\bigcap G}{R\bigcup G} \quad , \tag{8}$$

其中：R 表示跟踪结果，G 表示标识的真实位置，当 $S_{\text{Overlap}}>0.34$ 可认为跟踪成功。

2) 跟踪速度

跟踪速度表示每秒钟跟踪的帧数(frame per second，FPS，用 v_{FPS} 表示)，12 f/s 是连贯图片的最低标准，20 f/s 是 RPG 游戏运行的最低标准。定义式：

$$v_{\text{FPS}}=N_{\text{frame}}/L_{\text{time}} \quad , \tag{9}$$

其中：N_{frame} 表示视频中的帧数，L_{time} 表示跟踪视频时长[31]。

3) 跟踪中心误差

跟踪中心误差(用 E_{CLE} 表示)，定义式：

$$E_{\text{CLE}}=\sqrt{(x_{\text{R}}-x_{\text{G}})^2+(y_{\text{R}}-y_{\text{G}})^2} \quad , \tag{10}$$

其中：$(x_{\text{R}}, y_{\text{R}})$表示跟踪的中心坐标，$(x_{\text{G}}, y_{\text{G}})$表示标识的中心坐标。精度(Accuracy)表示测试集中 CLE 小于某一阈值的比例。

4.2 实验结果

本文将给出在公开数据集 OTB2015 综合测试结果和在 ILSVRC2015 数据集中飞机测试集的测试结果细节。

4.2.1 OTB2015

对于跟踪算法，现在学术界一般采用公开的数据

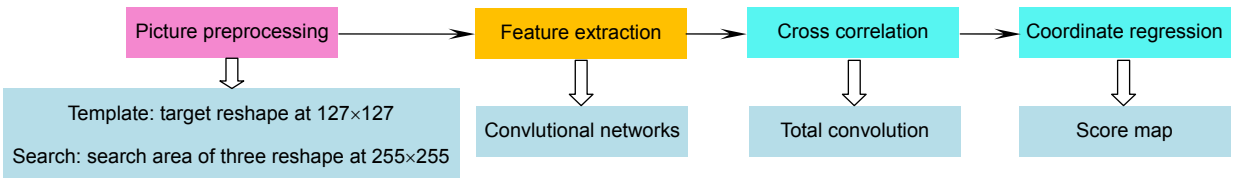


图 2 Siamese-MF 跟踪流程

Fig. 2 Tracking process of Siamese-MF

集进行测试，便于在同一评价条件下对比不同算法的性能优劣程度。在公开数据集 OTB2015 上对 Siamese-MF 算法进行测试，对比改进算法和原始算法在评价指标上的性能提升，结果如表 2。

表 2 Siamese-MF 与 Siamese-FC 在 OTB2015 对比测试结果

Table 2 Test results of Siamese-MF and Siamese-FC on OTB2015

Model	Overlap	Accuracy	*v*/(f/s)
Siamese-MF	0.44	0.61	76
Siamese-FC	0.27	0.52	58

表 2 中跟踪成功率(Overlap)改进后提高 0.17，跟踪稳定性(Accuracy)改进后提高 0.09，跟踪速度提升了 18 f/s。由此可得出，在卷积层上面的改进提升了跟踪算法的速度、精度以及稳定性。算法的改进效果明显。图 3 为 Siamese-MF 算法在部分数据集上的跟踪情况，分别为视频中的第 1 帧、第 50 帧、第 100 帧和第 200 帧。

4.2.2 ILSVRC2015 飞机测试集

在飞机测试集上面进行测试，是为了验证在实际应用中算法的性能表现，以及观察面临各种跟踪挑战时的跟踪效果。

这里采用 4.1 中定义的跟踪成功率(Overlap)、跟踪速度、跟踪稳定性即精度(Accuracy)作为定量评价标准，测试集为 ILSVRC2015 检测数据集的飞机测试集，具有各种飞行环境的 19 个视频，如图 4 为测试视频的定量评价指标折线图分析，横坐标表示视频序列，纵坐标表示评价指标。从图中分析可知，对于同一个跟踪视频 Siamese-MF 比 Siamese-FC 的跟踪效果更好，在准确性，实时性以及稳定性上面都有一定程度的提升。

通过柱状图分析可以直观对比 Siamese-MF 算法和 Siamese-FC 算法在指定评价参数上跟踪效果，表 3 为 ILSVRC2015 检测数据集的飞机测试视频序列在两种算法上的评价参数值。从表中可以看出，在跟踪精度上，Siamese-MF 的平均跟踪准确率比 Siamese-FC 高 6 个百分点；在跟踪稳定性上，Siamese-MF 比 Siamese-FC 高 10 个百分点；而在跟踪速度上，Siamese-MF 达到 40 f/s，基本满足实时跟踪。

5 总 结

本文提出的算法 Siamese-MF 是在 Siamese-FC 的卷积层以及跟踪策略上作出的改进，并且在公开数据集 OTB2015 以及 ILSVRC2015 检测数据集的飞机测试集上面进行测试。在多个环境下的测试集上的实验证明，该方法在扩展目标跟踪应用中具有较好的鲁棒性，基本满足实时性要求，并且有较高的准确性；同时在对飞机目标的跟踪应用上表现更好。本算法在卷积层中

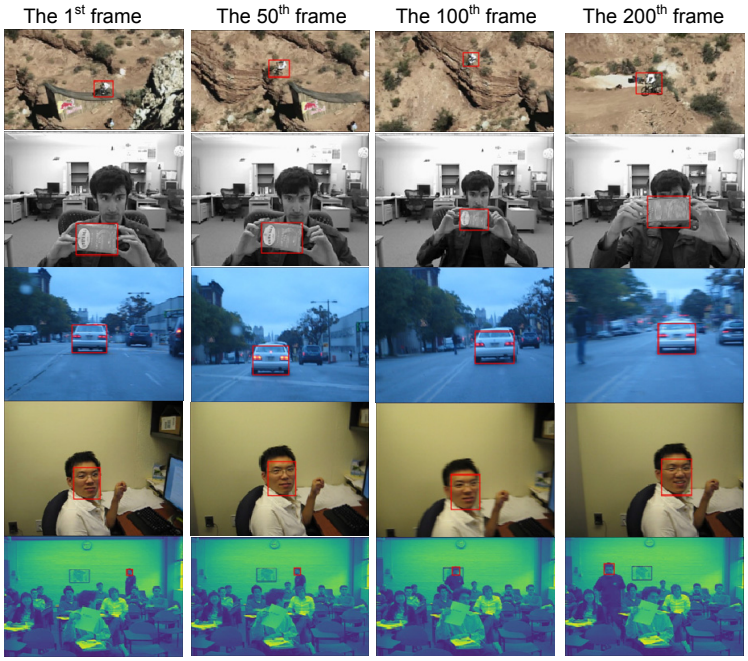


图 3 Siamese-MF 在 OTB2015 部分的跟踪结果

Fig. 3 Tracking results of Siamese-MF on OTB2015

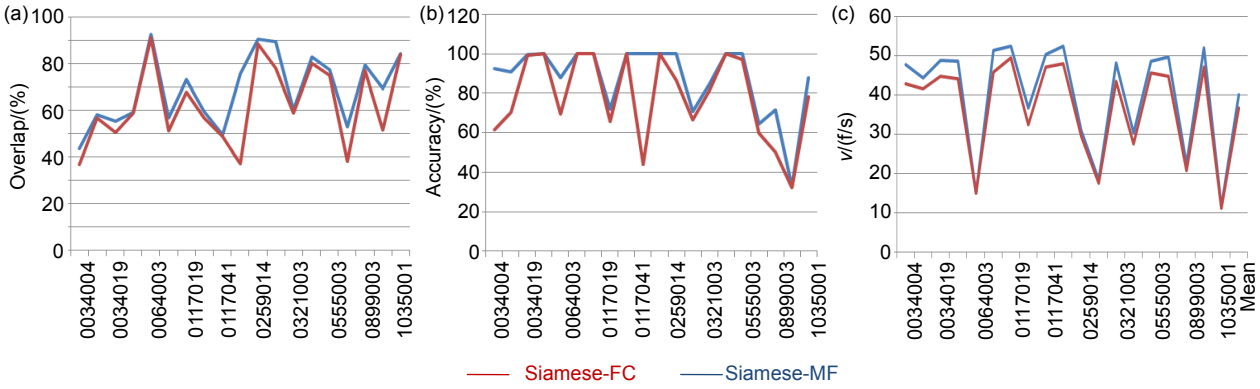


图 4 定性评价指标分析。(a) 跟踪成功率；(b) 跟踪稳定性；(c) 跟踪速度

Fig. 4 Qualitative evaluation index analysis. (a) Overlap; (b) Accuracy; (c) Velocity

表 3 Siamese-MF 与 Siamese-FC 的定量分析结果

Table 3 Quantitative analysis results of Siamese-MF and Siamese-FC

Videos	Overlap/(%)		Accuracy/(%)		*v*/(f/s)	
	Siam-MF	Siam-FC	Siam-MF	Siam-FC	Siam-MF	Siam-FC
0034004	43.7	36.9	92.4	61.6	47.8	42.8
0034014	58	56.8	90.7	70.1	44.4	41.7
0034019	55.4	50.6	99.6	99.3	48.9	44.8
0034023	59.1	58.7	100	100	48.6	44.1
0064003	92.4	91.2	88	69.6	15.1	15
0117004	56.8	51.2	100	100	51.4	45.9
0117019	73.4	67.6	100	100	52.4	49.5
0117024	59.4	56.7	71.8	65.6	36.8	32.4
0117041	49.5	48.8	100	100	50.4	47.1
0259004	75.8	37	100	43.9	52.4	48
0259014	90.6	88.3	100	100	31	29.5
0259019	89.5	78.1	100	86.7	18.4	17.5
0321003	60.2	58.9	70.6	66.7	48.2	43.5
0473003	82.9	80.1	84.4	81	30.3	27.5
0555003	77.3	74.9	100	100	48.7	45.6
743004	52.9	38	100	97.3	49.6	44.7
0899003	79.5	76.9	64.5	59.9	22.5	20.8
1000004	69.5	51.7	71.4	50	52.1	47.2
1035001	84.3	83.8	32.6	32.1	11.6	11.2
Mean	69	62.4	87.7	78.1	40	36.8

加入了特征融合和空洞卷积，在获取更丰富特征的同时减少计算量，获取更鲁棒的特征，适应变化的环境。对跟踪策略进行改进，加入模板在时间序列的累积，从而长时跟踪表现较好。在测试过程中对于尺度变化、遮挡、隐形、干扰等有良好的表现，速度基本满足实时要求，以飞机目标测试网络，可用于跟踪应用。本文所提出的方法还有进一步的优化和提升空间，如考虑利用频域的特征来进行相关性分析，那么计算量将大大减少，在保证精度的情况下进一步提升跟踪速度。

参考文献

[1] Yilmaz A, Javed O, Shah M. Object tracking: a survey[J]. *ACM Computing Surveys*, 2006, **38**(4): 13.

[2] Sivanantham S, Paul N N, Iyer R S. Object tracking algorithm

implementation for security applications[J]. *Far East Journal of Electronics and Communications*, 2016, **16**(1): 1–13.

[3] Kwak S, Cho M, Laptev I, *et al*. Unsupervised object discovery and tracking in video collections[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 3173–3181.

[4] Luo H B, Xu L Y, Hui B, *et al*. Status and prospect of target tracking based on deep learning[J]. *Infrared and Laser Engineering*, 2017, **46**(5): 502002.
罗海波, 许凌云, 惠斌, 等. 基于深度学习的目标跟踪方法研究现状与展望[J]. 红外与激光工程, 2017, **46**(5): 502002.

[5] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, **25**(5): 564–575.

[6] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//*Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981: 674–679.

[7] Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]//*Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 1822–1829.

[8] 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.

[9] Fan X S, Xu Z Y, Zhang J L. Dim small target tracking based on improved particle filter[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 170569.
樊香所, 徐智勇, 张建林. 改进粒子滤波的弱小目标跟踪[J]. 光电工程, 2018, **45**(8): 170569.

[10] Xi Y D, Yu Y, Ding Y Y, *et al*. An optoelectronic system for fast search of low slow small target in the air[J]. *Opto-Electronic Engineering*, 2018, **45**(4): 170654.
奚玉鼎, 于涌, 丁媛媛, 等. 一种快速搜索空中低慢小目标的光电系统[J]. 光电工程, 2018, **45**(4): 170654.

[11] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012: 1097–1105.

[12] Karen S Y, Andrew Z M. Very Deep Convolutional Networks for Large-scale Image Recognition[Z]. arXiv: 1409.1556[cs:CV], 2015.

[13] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[14] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2016.

[15] Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2017.

[16] Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge[J]. *International Journal of Computer Vision*, 2015, **115**(3): 211–252.

[17] Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: delving deep into convolutional nets[Z]. arXiv: 1405.3531[cs:CV], 2014.

[18] Shelhamer E, Long G, Darrell T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 640–651.

[19] 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*, 2014: 580–587.

[20] Li H. An overview of target tracking algorithm based on deep learning[J]. *Heilongjiang Science and technology information*, 2017(17): 49.
李贺. 基于深度学习的目标跟踪算法研究综述[J]. 黑龙江科技信息, 2017(17): 49.

[21] Wang N Y, Yeung D Y. Learning a Deep Compact Image Representation for Visual Tracking[C]//*NIPS*. Curran Associates Inc. 2013: 809–817.

[22] Nam H, Baek M, Han B. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking[Z]. arXiv: 1608.07242v1[cs:CV], 2016.

[23] Wang L J, Ouyang W L, Wang X G, *et al*. Visual tracking with fully convolutional networks[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision* (*ICCV*), 2015: 3119–3127.

[24] Ma C, Huang J B, Yang X K, *et al*. Hierarchical convolutional features for visual tracking[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision* (*ICCV*), 2015.

[25] Heid D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks[C]//*Proceedings of the 14th European Conference on Computer Vision*, 2016: 749–765.

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

[27] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[Z]. arXiv: 1511.07122[cs:CV], 2016.

[28] Wang H Y, Yang Y T, Zhang Z, *et al*. Deep-learning-aided multi-pedestrian tracking algorithm[J]. *Journal of Image and Graphics*, 2017, **22**(3): 349–357.
王慧燕, 杨宇涛, 张政, 等. 深度学习辅助的多行人跟踪算法[J]. 中国图象图形学报, 2017, **22**(3): 349–357.

[29] Wang X D. The influence of visual angle on the playability of games [J]. *Henan Science and Technology*, 2014(7): 12
王晓冬. 视觉角度对游戏可玩性的影响[J]. 河南科技, 2014(7): 12.

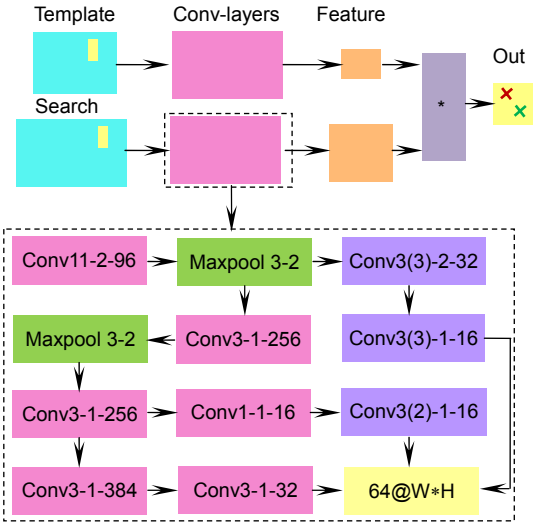
[30] Horikoshi K, Misawa K, Lang R. 20-fps motion capture of phase-controlled wave-packets for adaptive quantum control[C]//*Proceedings of the 15th International Conference on Ultrafast Phenomena XV*, 2006: 175–177.

[31] Zhao C M, Chen Z B, Zhang J L. Application of aircraft target tracking based on deep learning[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 180261.
赵春梅, 陈忠碧, 张建林. 基于深度学习的飞机目标跟踪应用研究[J]. 光电工程, 2019, **46**(9): 180261.

Research on target tracking based on convolutional networks

Zhao Chunmei[1,2], Chen Zhongbi[1*], Zhang Jianlin[1]

[1]Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan 610209, China;
[2]University of Chinese Academy of Sciences, Beijing 100049, China



Feedforward network of Siamese-MF

Overview: Deep learning has achieved good results in image classification, semantic segmentation, target detection and target recognition. However, it is still restricted by small sample training sets on object tracking. Object tracking is one of the most important researches in the field of computer vision, and has a wide range of applications. The challenge of object tracking lies in the complex states such as the target rotation, multi target, blur target, complex background, size change, target occlusion, fast moving and so on. Aiming at target tracking, this paper proposes an improved convolution network Siamese-MF (multi-feature Siamese networks) based on Siamese-FC (fully-convolutional Siamese networks). For tracking networks, considering the balance between speed and accuracy, reducing computational complexity and increasing the receptive field of convolution feature are the directions to improve the speed and accuracy of tracking networks. The improvement of the classical convolution network structure is mainly focused on two points: 1) introducing feature fusion to enrich features; 2) introducing dilated convolution to reduce computational complexity and enhance the receptive field. The improved convolution layer acts as feature extraction layer, and calculates the correlation between the target and the search area through the full convolution layer, so as to get the location of the tracking target according to the correlation graph. Siamese-MF algorithm achieves real-time and accurate tracking of targets in complex scenes. The average speed test on OTB2015 reaches 76 f/s, the mean value of overlap reaches 0.44, and the mean value of precision reaches 0.61, which meets the requirement in real-time tracking application of targets. For target tracking in this paper, the Siamese-MF networks are trained by using 5 convolutional layers of Conv1~Conv5 of AlexNet and 2 connected layers Skip1~Skip2 to extract the feature of target. In the tracking process, the trained networks are used as feed-forward networks, and the maximum score of outputs is regarded as the target location, while template updating is done in time series. Also the result of tracking is adaptive to scale transformation.

Citation: Zhao C M, Chen Z B, Zhang J L. Research on target tracking based on convolutional networks[J]. *Opto-Electronic Engineering*, 2020, **47**(1): 180668

Supported by Major Special Fund (G158207)
* E-mail: ioeyoyo@126.com