基于深度谱卷积神经网络的高效视觉目标跟踪算法

郭 强[1],芦晓红[1],谢英红[2],孙 鹏[1]

(1. 中国刑事警察学院图书馆,辽宁 沈阳 110035;
2. 沈阳大学 信息工程学院, 辽宁 沈阳 110044)

摘 要: 提出了一种基于深度频谱卷积神经网络的视觉目标跟踪算法。该算法在深度模型训练阶段采用谱池化替代深度卷积神经网络中的最大池化过程,用贝叶斯分类器替代 softmax 损失层计算最大分类值,并将其整合到深度神经网络跟踪框架中,通过新网络计算输入正负样本的概率分布预测目标位置。该算法充分利用谱池化在频域下降维到任意维度且计算高效的优点,克服了最大池化采样造成大量空间信息丢失的不足,提升了计算速度。在权威多场景视频标准测试库上对所提算法进行验证,结果验证了该算法兼顾了效率和跟踪精度,有效提高跟踪器的性能,在相同测试条件下,文中算法性能优于同类对比算法。

关键词: 视觉跟踪; 深度学习; 卷积神经网络; 谱池化

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3788/IRLA201847.0626005

Efficient visual target tracking algorithm based on deep spectral convolutional neural networks

Guo Qiang[1], Lu Xiaohong[1], Xie Yinghong[2], Sun Peng[1]

(1. Library of National Police University of China, Shenyang 110035, China;
2. School of Information Science and Engineering, Shenyang University, Shenyang 110044, China)

Abstract: The visual target tracking algorithm based on deep learning spectrum convolutional neural networks was presented. The spectral pooling was adopted instead of max pooling in the deep convolutional neural network, then the softmax loss layer was replaced with Bayesian theorem to compute maximum classifier score, and integrated it into the deep neural network tracking framework. The location of the target can be obtained by calculating the probability distribution of the input samples. The advantages of feature dimension reduction at random with spectral pooling and computation efficiency was taken to avoid much spatial information lost, which also helped to improve the computation speed. Compared with the original algorithm and other state-of-the-art methods, the proposed tracking method shows excellent performances on test baseline dataset.

Key words: visual tracking; deep learning; convolutional neural networks; spectral pooling

收稿日期: 2018-01-10; **修订日期:** 2018-02-20

基金项目: 国家自然科学基金(61603415, 61602322, 61503274);辽宁省教育厅科学研究一般项目(L2015558, W2015393)

作者简介: 郭强(1982-),男,讲师,博士,主要从事图像智能处理、机器人智能导航等方面的研究。Email: royinchina@163.com

0 引 言

视觉目标跟踪是计算机视觉领域的关键技术之一,在场景监控、运动捕捉、机器人自主导航、人机交互等方面有着广泛的应用[1-2]。随着深度学习技术的兴起,2013 年以来,许多学者采用深度卷积神经网络方法设计鲁棒的视觉跟踪算法[3]。卷积神经网络具有分类正确率高、集成特征提取、鲁棒性强等优点,逐渐在性能上超越传统方法,使得其成为跟踪方向新的研究热点。参考文献[4]首次将深度模型运用在单目标视觉跟踪算法中,提出了离线预训练结合在线微调的思路,解决了跟踪中训练样本不足的问题,但其离线预训练目标为图片重构,这与在线跟踪区分目标和背景的目标不一致。此外,与传统人工特征相比,其 4 层深度网络模型对特征刻画能力并无明显优势。参考文献[5]利用卷积神经网络的 AlexNet 结构作为获取特征的网络模型,对网络的输入和输出大小进行了限定,引入了空间金字塔采样,提高了定位的精度。但离线预训练使用的是大量无关联图片,未使用更贴合跟踪实质的时序关联数据。参考文献[6]利用 ImageNet 上预训练出的卷积神经网络提取目标特征,并对 VGG-16 的不同层特征图进行了分析,结合 ensemble 思路构建跟踪框架结构,但更新策略易将不正确的目标表观更新到模型中。参考文献[7]提出一种由粗粒度到细粒度提取深度特征的跟踪算法,利用相关滤波器确定最终的跟踪框,跟踪结果显示出深度特征结合相关滤波器的巨大优势。但是用同一个 CNN 难以完成所有训练序列中前景和背景区分的任务。针对此问题,参考文献[8]直接用跟踪视频预训练 CNN 获得总的目标表示能力的方法,提出了创新的多域训练方法和训练数据交叉运用的思路。但 MDNet 虽然网络结构较小,速度仍较慢。近年来离散傅里叶变换以其快速卷积运算的高效性,在深度学习领域逐渐受到重视。参考文献[9]研究成果表明针对任意滤波尺寸,利用频域离散傅里叶变换计算卷积操作要远比空间域计算方法高效。

为此,文中提出一种基于优化的深度频谱卷积神经网络的视觉目标跟踪算法。首先,针对预训练的深度卷积神经网络结构,利用谱池化保留目标的空间分辨率有效学习判别性特征,用频域上的有效系数进行特征维度缩减,缩减计算时间。之后利用输出层的贝叶斯分类器进行分类获得跟踪结果,克服了跟踪时直接训练样本不足的问题。

1 基于深度谱卷积谱网络跟踪算法

图 1 所示为所提出跟踪算法的网络结构,该网络的目标是通过深度谱卷积神经训练判别式模型实现目标外观和背景的二分类。

1.1 经典卷积神经网络结构

卷积神经网络是深度学习最成功的模型之一,

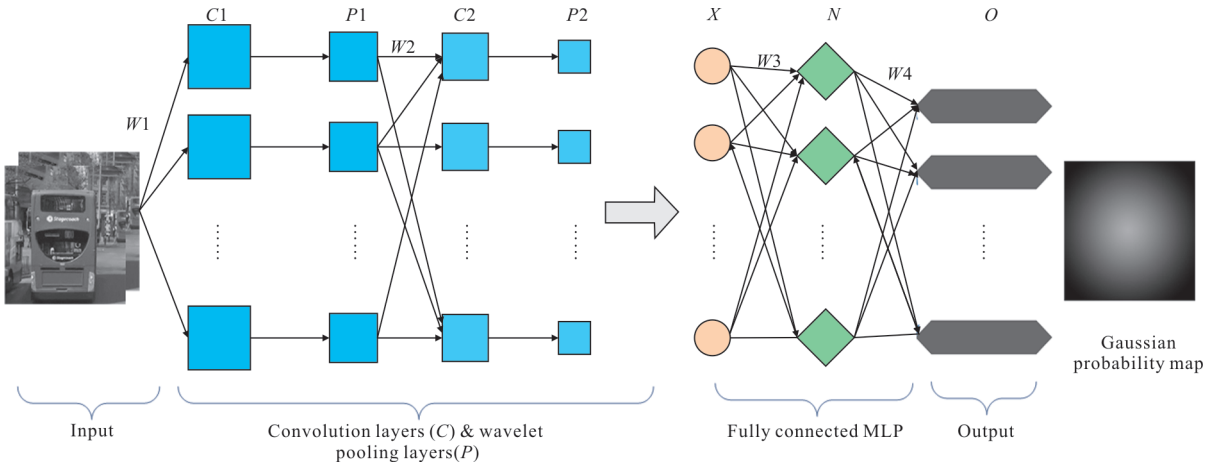


图 1 深度谱卷积神经网络结构图

Fig.1 Structure diagram of deep spectral CNN

是一种多层前馈神经网络，能够从数据中学习并提取特征。通常由输入、隐含层特征提取部分、全连接层、输出层构成。特征提取部分由卷积层 C、非线性变换和下采样交替操作进行而构成，输入图像通过若干个可训练的滤波器组即卷积核进行非线性卷积，卷积后在每一层产生特征映射图，然后利用非线性变换对特征进行筛选，利用池化方法进行下采样操作从而得到分辨率降低的图像，在一定程度上增加网络对位移、缩放、扭曲的鲁棒性。通常卷积层计算如下：

$$x_j^l=f\left(\sum_{i\in M_j}x_i^{l-1}*w_{ij}^l+b_j^l\right) \tag{1}$$

式中：$f(\cdot)$为非线性激活函数；M_j 为特征图或输入层的感受野；x 为特征图的输出；$*$ 是卷积操作；w 为卷积核进行局部连接和权值共享；l 为层数；b 为偏置。

最后由降维的特征图构成的特征向量通过 softmax 层，进行最后的分类和归一化。

1.2 深度谱卷积网络

上述池化算法的选择决定了子采样提取到特征的有效性，特征提取的有效程度决定了网络的学习性能。然而目前网络结构中的最大池化或平均池化虽然增加了网络的鲁棒性，但是不可避免造成信息的大量损失，如最大池化中的最大值仅包含非常局部化信息，并不能表示整体窗口。文中在下采样中引入谱池化[10]，其低通滤波操作在相同维度下可以保留更多信息，而且并没有其他池化方法中的大幅度维度缩减，其实现通过矩阵截取可以减少计算量[11]。表 1 为谱池化算法。图 2 是谱池化后效果示意图，即使分辨率下降较多依然具有较强辨别性。

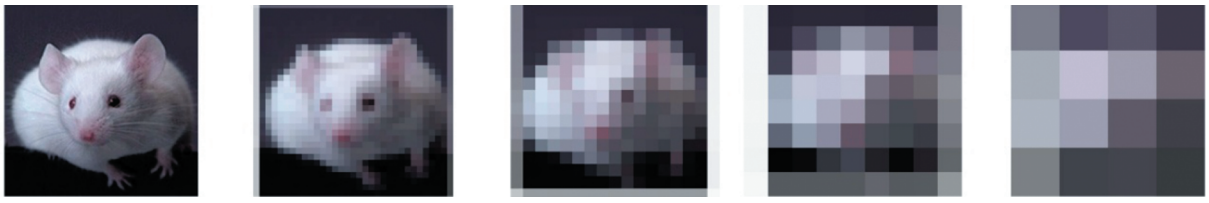
表 1 卷积神经网络前向传播中的谱池化

Tab.1 Spectral pooling in feed-forward of CNNS

Input: feature map, dimension reduction size
1. Discrete Fourier Transform
2. crop spectrum by DFT
3. treat corner case
4. Inverse Discrete Fourier Transform
Output: sub-sampling map

(a) 谱池化

(a) Spectral pooling



(b) 最大池化

(b) Max pooling

图 2 谱池化和最大池化效果示意图

Fig.2 Diagram of spectrum pooling and max pooling effect

如图 1 所示，引入谱池化后的网络首先通过输入训练样本，经过卷积层、谱池化层、全连接层及一组朴素贝叶斯分类器，根据输出的高斯分布概率图得到跟踪结果。参照参考文献[12]的网络结构，文中优化后的参数化结构：

$$\left(C_{3\times3}^{96+32m}\rightarrow SP_{\downarrow\lfloor\gamma H_m\rfloor\times\lfloor\gamma H_m\rfloor}\right)_{m=1}^{M}\rightarrow FC^{1\ 024}\rightarrow FC^{512}\rightarrow\text{Bayesian} \tag{2}$$

式中：C_S^F 为卷积层 C 中 F 个滤波模板，尺寸是 S；H_m 为 m 层特征图的高度；γ 为特征图维度缩减因子；

$SP_{\downarrow d}$为谱池化后输出维度 d 的下采样特征图结果；FC 为全连接层；Bayesian 为分类器层。

网络结构预训练需要进行初始化，在目标跟踪的训练数据不足的情况下，文中使用辅助的大量非跟踪训练数据进行初步预训练，学习所有可识别物体特征的通用特征，之后在实际跟踪时，利用当前跟踪目标的有限样本信息对预训练模型仅全连接层进行微调，从而学习特定目标的特征，使模型对当前跟踪目标有更强的分类性能。

1.3 提出的跟踪器模型及在线微调网络

图 1 的全连接层输出可视为特征向量，即每个图像样本可以表示为：$S=[s_1,s_2,\cdots,s_N]^{\mathrm{T}}$。

如给定正负样本的概率分布，目标的位置可以由朴素贝叶斯分类器进行后验判别获得：

$$c(s)=\frac{p(s|y=1)p(y=1)}{\sum_{y=0,1}p(s|y)p(y)}=\sigma(h_k(s)) \tag{3}$$

$$h_k(s)=\log\left(\frac{\prod_{k=1}^{K}p(g_k(s)|y=1)p(y=1)}{\prod_{k=1}^{K}p(g_k(s)|y=0)p(y=0)}\right)=\sum_{k=1}^{K}\varnothing_k(s) \tag{4}$$

$$\varnothing_k(s)=\log\left(\frac{p(g_k(s)|y=1)}{p(g_k(s)|y=0)}\right) \tag{5}$$

$p(g_k(s)|y=1)$和 $p(g_k(s)|y=0)$服从高斯分布，由上式可以获得正负样本的概率分布。当获得下一帧目标位置后，需要对模型进行更新以适应外观变化，包括分类器模型参数更新和网络权重更新。分类器模型参数即正样本的均值和方差，更新方式见参考文献[13]。网络权重更新通过计算样本的梯度值 $\partial\varnothing_k/\partial s$。

2 实验结果和分析

2.1 实验说明

为了评估文中算法的有效性，文中进行了两组实验，分别进行了定量和定性分析。实验中所采用的测试标准库是由 Wu 等[14]提出的国际上权威的跟踪算法评测库，该库搜集了 100 个公开的测试序列包括光照变化、遮挡、旋转、尺度变化、复杂背景等多种跟踪场景，以及 31 个公开代码的跟踪算法。实验从中选取了 MDNET 跟踪[8]、SINT 跟踪[15]、HDT[16]跟踪、FCNT[6]跟踪、KCF 跟踪[17]、CNT[18]、CNN-SVM[19]经典方法进行了比较。准确率和成功率是跟踪算法性能评估的常用标准，文中将曲线下面积作为成功率计算方法，定义为 Score=$A\cap B/A\cup B$，其中 A 是目标真实位置处的矩形框，B 是跟踪器预测结果的矩形框位置，∩和∪表示交集和并集。

参考文献[14]提出新的跟踪性能评估指标，根据叠加区域曲线阈值面积(AUC)获取更客观的成功率指标。而准确率(Precision)依然采用标准方式，即目标中心位置欧式距离空间上的差值。高性能的跟踪器应该具有更强的鲁棒性即高成功率和准确率。

2.2 实验结果

(1) 定量分析

表 1 列出了测试库 31 种算法中具有代表性的 5 种算法与文中算法进行了对比，主要测试数据基于最新的 OTB100[14]，分别利用 AUC 和 Precision 指标评估性能。由表 2 可见，文中基于谱卷积神经网络的跟踪算法 (Ours) 优于列出的其他几种深度学习算法以及经典的基于核相关滤波 KCF 算法，通过高性能 GPU 及稀疏的更新策略，文中的算法跟踪速度可以达到实时。

表 2 不同跟踪算法的跟踪性能比较

Tab.2 Comparison of the performance among different tracking algorithm

Testing sequences	Ours	Deep-SRDCF	CF2	HDT	CNN-SVM	KCF
AUC-OTB100	0.677	0.635	0.562	0.654	0.554	0.477
Precision-OTB100	0.902	0.851	0.837	0.848	0.814	0.693
CNN algorithm	Yes	Yes	Yes	Yes	Yes	No

为了验证跟踪算法对测试序列初始化位置的鲁棒性，定量分析部分采用空间鲁棒性成功率即目标初始位置在不同标定位置进行评测。

图 3 是空间鲁棒性成功率和空间鲁棒性准确率曲线，图例是各算法基于 TB50[14]测试数据集的的跟踪成功率及准确率均值，如图 3 所示的成功率曲线及均值，文中所提算法对初始位置的干扰具有较强鲁棒性，由成功率图可见，当跟踪目标位置与真实位置矩形框重叠区域超过 70%时，文中所提算法与当前跟踪性能最佳的 MDNet 算法性能相近，

甚至占优。而准确率曲线可以看出在 5 个像素位置误差内，文中跟踪算法更精准。

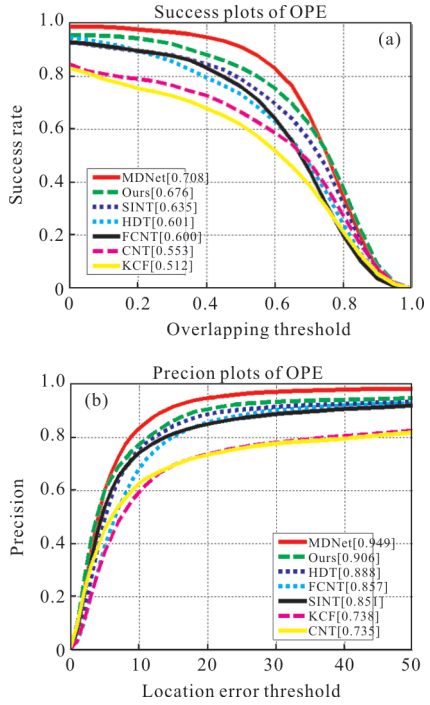


图 3 基于 TB50 测试库的空间鲁棒性成功率和准确率

Fig.3 Success and precision plots of OPE over the TB50 test data base

(2) 定性分析

图 4 是基于视频测试序列的各种算法跟踪结果，Couple 序列是测试尺度变化和非平面旋转时行人目标在复杂背景下的跟踪效果。在测试序列的开始阶段，各跟踪器均可以根据初始帧特征学习准确跟踪目标，但是随着背景干扰及尺度和面外旋转发生，利用正负样本进行特征学习的文中算法仍可以准确锁定目标位置。在 Freeman4 序列中，发生了目标遮挡及旋转变化，方框 KCF 跟踪算法因为初始化矩阵不能自适应改变，会有漂移现象发生。而其他几种深度学习算法的下采样层采用最大池化操作丢失了较多的空间信息，导致跟踪目标丢失。在 CarDark 序列中，目标发生了较明显的尺度和光照及背景变化，其他几种算法如 FCNT 等是用网络中每个卷积层进行目标位置的预测，而文中算法是将目标位置及尺度作为整体进行处理，因此算法跟踪更加准确。MotorRolling 序列中，目标摩托车手在复杂背景下具有快速移动模糊及面内旋转等测试属性，文中算法跟踪效果不如 MDNet 等三种算法，因为快速移动造成的运动目标模糊与谱池化操作叠加影响样本轮廓特征的学习，从而影响跟踪性能。Jogging-2 的跟踪任务相对简单，除了 CNT 算法外均跟踪成功，因为 CNT 构造的两层卷积神经网络虽然速度优秀，但是因为没有预训练及池化操作，与深度学习算法有很大的区别，效果并不突出。Skating 最大的难点是剧烈的光照变化以及大幅度旋转，几种深度学习算法因为卷积神经网络对几何变换、形变、光照具有一定程度的不变性，所以总体表现良好，但是在弱光下文中算法仍需改进。

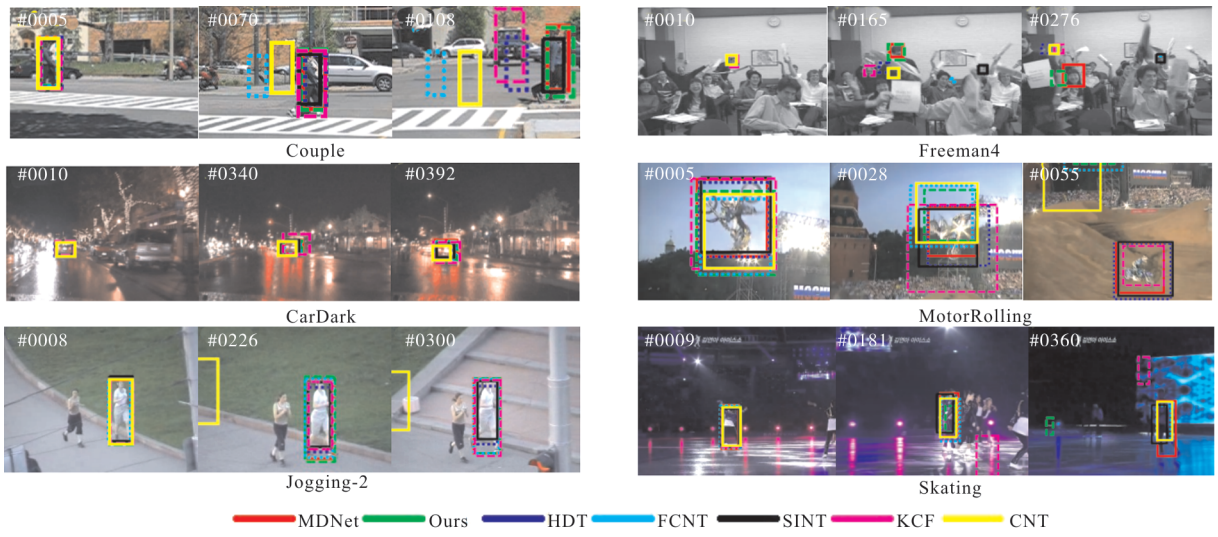


图 4 7 种算法的跟踪效果比较

Fig.4 Tracking results of the sequences with the 7 algorithm

3 结 论

文中提出一种高效的基于深度学习的跟踪算法，结合了判别式模型处理特征分类的优点，利用谱池化进行卷积神经网络的改进，保留了有效空域信息的同时又有效降低特征到任意预设维度，同时将贝叶斯分类器整合到深度学习网络结构输出层中，从而选出正负样本中概率分布响应值最高的点估计为目标位置。文中采用傅里叶变换将特征由时域变换到频域计算滤波操作，提高了算法的计算性能。对最新视频测试库进行算法的性能评估，与其他几种典型深度学习代表算法相比，文中算法对部分遮挡、形变等复杂环境具有更好的鲁棒性。下一步研究工作计划用小波基取代傅里叶基进行池化操作，以减少特征局部信息的丢失和计算效能。

参考文献：

[1] Liu Zhi, Huang Jiangtao, Feng Xin. Action recognition model construction based on multi-scale deep convolution neural network [J]. ***Optics and Precision Engineering***, 2017, 25(3): 799-805. (in Chinese)
刘智, 黄江涛, 冯欣. 构建多尺度深度卷积神经网络行为识别模型[J]. 光学 精密工程, 2017, 25(3): 799-805.

[2] Pei Xiaomin, Fan Huijie, Tang Yandong. Action recognition method of spatio-temporal feature fusion deep learning network [J]. ***Infrared and Laser Engineering***, 2018, 47(2): 0203007. (in Chinese)
裴晓敏, 范慧杰, 唐延东. 时空特征融合深度学习网络人体行为识别方法[J]. 红外与激光工程, 2018, 47(2): 0203007.

[3] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)
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): 0502002.

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

[5] Wang N, Li S, Gupta A, et al. Transferring rich feature hierarchies for robust visual tracking [EB/OL]. 2015-04-23. https://arxiv.org/abs/1501. 04587, 2015.

[6] Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Chile: IEEE Computer Society, 2016: 3119-3127.

[7] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual Tracking [C]//Proceedings of the IEEE International Conference on Computer Vision. Chile: IEEE Computer Society, 2016: 3074-3082.

[8] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [EB/OL]. 2016-01-06. https://arxiv.org/pdf/1510.07945, 2015.

[9] Vasilache N, Johnson J, Mathieu M, et al. Fast convolutional nets with fbfft: A GPU Performance Evaluation [EB/OL]. 2015-04-10. https://arxiv.org/pdf/ 1412.7580, 2014.

[10] Rippel O, Snoek J, Adams R P. Spectral representations for convolutional neural networks [C]//International Conference on Neural Information Processing Systems. Canada: MIT Press, 2015: 2449-2457.

[11] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks [EB/OL]. 2017-10-19. https://arxiv.org/ pdf/1512.07108, 2015.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012: 1097-1105.

[13] Boris B, Yang M H, Belongie S.Visual tracking with online multiple instance learning [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 983-990.

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

[15] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1420-1429.

[16] Qi Y, Zhang S, Qin L, et al. Hedged deep tracking [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4303-4311.

[17] Henriques J F, Rui C, Martins P, et al. High-speed tracking with kernelized correlation filters [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2014, 37 (3): 583-561.

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

[19] Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network [J]. ***Computer Science***, 2015: 597-606.