



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象学报 中国图形学报

2022
04
VOL.27

ISSN1006-8961
CN11-3758/TB



口罩人脸姿态分类 P1125



第27卷第4期（总第312期）
2022年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



残差密集结构的东巴画渐进式重建(第1084页)



结合双模板融合与李生网络的鲁棒视觉目标跟踪(第1191页)



注意力机制的曲面沉浸式投影系统补偿(第1238页)

综述

中国图像工程: 2021

章毓晋 1009

人脸伪造及检测技术综述

曹申豪, 刘晓辉, 毛秀青, 邹勤 1023

数据集论文

MTMS300: 面向显著物体检测的多目标多尺度基准数据集

李楚为, 张志龙, 李树新 1039

图像处理和编码

利用条件生成对抗网络的光场图像重聚焦

谢柠宇, 丁宇阳, 李明悦, 刘渊, 律睿悠, 晏涛 1056

二阶变分图像恢复模型的重启动快速ADMM方法

宋田田, 潘振宽, 魏伟波, 李青 1066

残差密集结构的东巴画渐进式重建

蒋梦洁, 钱文华, 徐丹, 吴昊, 柳春宇 1084

图像分析和识别

融合自注意力机制的生成对抗网络跨视角步态识别

张红颖, 包雯静 1097

着装场景下双分支网络的人体姿态估计

吕中正, 刘骊, 付晓东, 刘利军, 黄青松 1110

嵌入双尺度分离式卷积块注意力模块的口罩人脸姿态分类

陈森椒, 刘文波, 张弓 1125

联合损失优化下的高相似度奶山羊身份识别

尚诚, 王美丽, 宁纪锋, 李群辉, 姜雨, 王小龙 1137

对抗一致性约束的无监督域自适应绝缘子检测

李梅玉, 李仕林, 赵明, 方正云, 张亚飞, 余正涛 1148

注意力机制改进轻量SSD模型的海面小目标检测

贾可心, 马正华, 朱蓉, 李永刚 1161

注意力引导网络的显著性目标检测

何伟, 潘晨 1176

结合双模板融合与李生网络的鲁棒视觉目标跟踪

陈志良, 石繁槐 1191

流形正则化约束的图像语义分割

肖振久, 宗佳旭, 兰海, 魏宪, 唐晓亮 1204

空洞可分离卷积和注意力机制的实时语义分割

王因, 侯志强, 蒲磊, 马素刚, 程环环 1216

自适应多任务学习的自动艺术分析

杨冰, 向学勤, 孔万增, 施妍, 姚金良 1226

虚拟现实与增强现实

注意力机制的曲面沉浸式投影系统补偿

雷清桦, 杨婷, 程鹏 1238

遥感图像处理

遥感影像空间分治快速匹配

卫春阳, 乔彦友 1251

Chinagraph 2020

结合门循环单元和生成对抗网络的图像文字去除

王超群, 全卫泽, 侯诗玉, 张晓鹏, 严冬明 1264

面向人体骨骼运动数据优化的双自编码器网络

李书杰, 朱海生, 王磊, 刘晓平 1277

采用蒸馏训练的时空图卷积动作识别融合模型

杨清山, 穆太江 1290

面向时变体数据的特征可视化方法

刘力 1302

ChinaVR 2020

带法向约束的隐式T样条曲线重构

任浩杰, 寿华好, 莫佳慧, 张航 1314

观看经度联合加权全景图显著性检测算法

孙耀, 陈纯毅, 胡小娟, 李凌, 邢琦玮 1322

CONTENTS

JOURNAL OF IMAGE AND GRAPHICS



Gradual model reconstruction of Dongba painting based on residual dense structure(P1084)



Double template fusion based siamese network for robust visual object tracking(P1191)



An attention mechanism based inter-reflection compensation network for immersive projection system(P1238)

Review

- Image engineering in China: 2021
Zhang Yujin 1009
- A review of human face forgery and forgery-detection technologies
Cao Shenhao, Liu Xiaohui, Mao Xiuqing, Zou Qin 1023

Dataset

- MTMS300: a multiple-targets and multiple-scales benchmark dataset for salient object detection
Li Chuwei, Zhang Zhilong, Li Shuxin 1039

Image Processing and Coding

- Light field image re-focusing based on conditional generative adversarial networks leverage
Xie Ningyu, Ding Yuyang, Li Mingyue, Liu Yuan, Lyu Ruimin, Yan Tao 1056
- Restart fast ADMM methods for second-order variational models of image restoration
Song Tiantian, Pan Zhenkuan, Wei Weibo, Li Qing 1066
- Gradual model reconstruction of Dongba painting based on residual dense structure
Jiang Mengjie, Qian Wenhua, Xu Dan, Wu Hao, Liu Chunyu 1084

Image Analysis and Recognition

- The cross-view gait recognition analysis based on generative adversarial networks derived of self-attention mechanism
Zhang Hongying, Bao Wenjing 1097
- Dual branch network for human pose estimation in dressing scene
Lyu Zhongzheng, Liu Li, Fu Xiaodong, Liu Lijun, Huang Qingsong 1110
- A embedded dual-scale separable CBAM model for masked human face poses classification
Chen Senqiu, Liu Wenbo, Zhang Gong 1125
- Joint loss optimization based high similarity identification for milch goats
Shang Cheng, Wang Meili, Ning Jifeng, Li Qunhui, Jiang Yu, Wang Xiaolong 1137
- Unsupervised domain adaptation insulator detection based on adversarial consistency constraints
Li Meiyu, Li Shilin, Zhao Ming, Fang Zhengyun, Zhang Yafei, Yu Zhengtao 1148
- Attention-mechanism-based light single shot multiBox detector modelling improvement for small object detection on the sea surface
Jia Kexin, Ma Zhenghua, Zhu Rong, Li Yonggang 1161
- The salient object detection based on attention-guided network
He Wei, Pan Chen 1176
- Double template fusion based siamese network for robust visual object tracking
Chen Zhiliang, Shi Fanhuai 1191
- Image semantic segmentation based on manifold regularization constraint
Xiao Zhenjiu, Zong Jiaxu, Lan Hai, Wei Xian, Tang Xiaoliang 1204
- Real-time semantic segmentation analysis based on cavity separable convolution and attention mechanism
Wang Nan, Hou Zhiqiang, Pu Lei, Ma Sugang, Cheng Huanhuan 1216
- Automatic art analysis based on adaptive multi-task learning
Yang Bing, Xiang Xueqin, Kong Wanzeng, Shi Yan, Yao Jinliang 1226

Virtual Reality and Augmented Reality

- An attention mechanism based inter-reflection compensation network for immersive projection system
Lei Qinghua, Yang Ting, Cheng Peng 1238

Remote Sensing Image Processing

- Spatial divide and conquer based remote sensing image quick matching
Wei Chunyang, Qiao Yanyou 1251

Chinagraph 2020

- Gate recurrent unit and generative adversarial networks for scene text removal
Wang Chaoqun, Quan Weize, Hou Shiyu, Zhang Xiaopeng, Yan Dongming 1264
- Dual auto-encoder network for human skeleton motion data optimization
Li Shujie, Zhu Haisheng, Wang Lei, Liu Xiaoping 1277
- Action recognition using ensembling of different distillation-trained spatial-temporal graph convolution models
Yang Qingshan, Mu Taijiang 1290
- A feature visualization method for time-varying volume data
Liu Li 1302

ChinaVR 2020

- Implicit T-spline curve reconstruction with normal constraint
Ren Haojie, Shou Huahao, Mo Jiahui, Zhang Hang 1314
- Saliency detection algorithm of panoramic images using joint weighting with observer's attention longitude
Sun Yao, Chen Chunyi, Hu Xiaojuan, Li Ling, Xing Qiwei 1322

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2022)04-1191-13

论文引用格式: Chen Z L and Shi F H. 2022. Double template fusion based siamese network for robust visual object tracking. Journal of Image and Graphics, 27(04): 1191-1203 (陈志良, 石繁槐. 2022. 结合双模板融合与孪生网络的鲁棒视觉目标跟踪. 中国图象图形学报, 27(04): 1191-1203) [DOI: 10.11834/jig.200660]

结合双模板融合与孪生网络的鲁棒视觉目标跟踪

陈志良, 石繁槐*

同济大学电子与信息工程学院, 上海 201804

摘要: 目的 视觉目标跟踪算法主要包括基于相关滤波和基于孪生网络两大类。前者虽然精度较高但运行速度较慢, 无法满足实时要求。后者在速度和精度方面取得了出色的跟踪性能, 然而, 绝大多数基于孪生网络的目标跟踪算法仍然使用单一固定的模板, 导致算法难以有效处理目标遮挡、外观变化和相似干扰物等情形。针对当前孪生网络跟踪算法的不足, 提出了一种高效、鲁棒的双模板融合目标跟踪方法 (siamese tracker with double template fusion, Siam-DTF)。方法 使用第 1 帧的标注框作为初始模板, 然后通过外观模板分支借助外观模板搜索模块在跟踪过程中为目标获取合适、高质量的外观模板, 最后通过双模板融合模块, 进行响应图融合和特征融合。融合模块结合了初始模板和外观模板各自的优点, 提升了算法的鲁棒性。结果 实验在 3 个主流的目标跟踪公开数据集上与最新的 9 种方法进行比较, 在 OTB2015 (object tracking benchmark 2015) 数据集中, 本文方法的 AUC (area under curve) 得分和精准度分别为 0.701 和 0.918, 相比于性能第 2 的 SiamRPN++ (siamese region proposal network++) 算法分别提高了 0.6% 和 1.3%; 在 VOT2016 (visual object tracking 2016) 数据集中, 本文方法取得了最高的期望平均重叠 (expected average overlap, EAO) 和最少的失败次数, 分别为 0.477 和 0.172, 而且 EAO 得分比基准算法 SiamRPN++ 提高了 1.6%, 比性能第 2 的 SiamMask_E 算法提高了 1.1%; 在 VOT2018 数据集中, 本文方法的期望平均重叠和精确度分别为 0.403 和 0.608, 在所有算法中分别排在第 2 位和第 1 位。本文方法的平均运行速度达到 47 帧/s, 显著超出跟踪问题实时性标准要求。结论 本文提出的双模板融合目标跟踪方法有效克服了当前基于孪生网络的目标跟踪算法的不足, 在保证算法速度的同时有效提高了跟踪的精确度和鲁棒性, 适用于工程部署与应用。

关键词: 视觉目标跟踪 (VOT); 孪生网络; 特征融合; 双模板机制; 深度学习

Double template fusion based siamese network for robust visual object tracking

Chen Zhiliang, Shi Fanhuai*

College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China

Abstract: **Objective** Visual object tracking (VOT) analysis has challenged for computer vision research. Current trackers can be roughly segmented into two categories like correlation filter trackers and Siamese network based trackers. Correlation filter trackers train a circular correlation based regressor analysis in the Fourier domain. Siamese network based trackers have improved the speed and accuracy issue of deep features. A Siamese network consists of two branches which implicitly

收稿日期: 2020-11-13; 修回日期: 2021-02-19; 预印本日期: 2021-02-26

* 通信作者: 石繁槐 fshshi@tongji.edu.cn

基金项目: 上海市科技兴农重点攻关项目 (沪农科创字 (2018) 第 3-6 号. 2020-02-08-00-07-F01480)

Supported by: Key Research Projects for Developing Agriculture through Science and Technology in Shanghai ((2018) 3-6. 2020-02-08-00-07-F01480)

encodes the original patches to another space and then fuses them with an identified tensor to generate a single output. However, most Siamese network based trackers utilize the single fixed template to resolve occlusion, appearance change and distractors problems. We illustrate an efficient and robust Siamese network based tracker via double template fusion, referred as Siamese tracker with double template fusion (Siam-DTF). The demonstrated Siam-DTF has a double template mechanism in related to qualified robustness. **Method** Siam-DTF consists of three emerging branches like initial template z , appearance template z_a and search area x . First, we facilitate the appearance template search module (ATSM) which fully utilizes the information of historical frames to efficiently obtain the appropriate and high-quality appearance template when the initial template is not consistent with the current frame. The appearance template, which is flexible and adaptive to the appearance changes of the object, can represent the object well when facing hard tracking challenges. We choose the frame with the highest confidence in the historical frames to crop the appearance template. To filter out low-quality template, we drop the appearance template if its predicted box has a lower intersection-of-union or its confidence score is lower than that of the initial template. In order to balance the accuracy and speed of our tracker, we use a sparse update strategy on the appearance template. In terms of theoretical analysis and experimental validations, we clarify that the confidence score change of tracker reflects the tracking quality more. When the max confidence of current frame is lower than average confidence of the historical N frames with a certain margin m , we conduct the ATSM to update the appearance template. Next, our fusion module illustration achieves more robust results based on these two templates. The initial template and the appearance template branch are integrated in terms of fusion of score maps and fusion of features. **Result** The nine tailored trackers model including the correlation filter trackers and Siamese network based trackers demonstrated on three public tracking datasets in the context of object tracking benchmark 2015 (OTB2015), VOT2016 and VOT2018. In OTB2015, the quantitative evaluation metrics contained area under curve (AUC) and precision. The proposed Siam-DTF is capable to rank 1 st both in terms of AUC and precision. Compared with the baseline tracker Siamese region proposal network ++ (SiamRPN ++), our Siam-DTF improves 0.6% in AUC and 1.3% in precision. Since we unveiling the power of deep feature of historical frames, Siam-DTF precision is prior to correlation filter tracker efficient convolution operators (ECO) by 0.8%. In VOT2016 and VOT2018, the quantitative evaluation metrics contained accuracy (average overlap while tracking successfully) and robustness (failure times). The overall performance is evaluated using expected average overlap (EAO) which takes account of both accuracy and robustness. In VOT2016, Siam-DTF achieves the qualified EAO score 0.477 and the least failure rate 0.172. For EAO score, our method outperforms the baseline tracker SiamRPN ++ and the second best tracker SiamMask_E by 1.6% and 1.1%, respectively. Also our method decreases the failure rate from 0.200 to 0.172 compared to SiamRPN ++, indicating that our Siam-DTF tracker robustness is well. In VOT2018, Siam-DTF obtains the good result with accuracy of 0.608. Siam-DTF also obtains the second good EAO score 0.403. As for tracking speed, Siam-DTF tracker not only achieves a substantial improvement, but also running efficiently at 47 frame per second (FPS). In summary, it is concluded that all these consistent results show the strong generalization ability of our tracker Siam-DTF.

Conclusion We propose an efficient and robust Siamese tracker with double template fusion, referred as Siam-DTF. Siam-DTF fully utilizes the information of historical frames to obtain the appearance template with good adaptability. All 3 benchmarks analysis demonstrate the effectiveness based on and Siam-DTF consistent results.

Key words: visual object tracking; siamese network; fusion of feature; double-template mechanism; deep learning

0 引言

视觉目标跟踪 (visual object tracking, VOT) 在机器人 (Held 等, 2016)、视觉监控 (Xing 等, 2010)、医学影像分析 (Liu 等, 2012)、人机交互 (Zhang 和 Vela, 2015) 和自动驾驶 (Lee 和 Hwang, 2015) 等许多计算机视觉领域都有广泛应用。给定一个视频序

列, 目标跟踪的核心任务是对每一帧图像鲁棒地估计目标的运动状态 (位置、尺寸等), 但是对于计算机来说, 实现准确的目标跟踪非常困难, 主要挑战包括旋转、遮挡、形变、尺度变换和背景干扰等。

基于判别相关滤波器 (discriminative correlation filter, DCF) 的目标跟踪方法在精度和鲁棒性方面较传统的跟踪算法有了很大提高。Bolme 等人 (2010) 通过最小化傅里叶域下的期望输出与实际输出之间

的平方误差之和,提出了 MOSSE (minimum output sum of squared error) 跟踪算法,是最早利用相关滤波器研究跟踪的工作。基于 MOSSE, Henriques 等人 (2012) 提出核循环结构 (circulant structure with kernel, CSK) 跟踪算法,利用目标外观的循环结构构建大量样本,并且用核正则化最小二乘法训练相关滤波器,实现了更精确的目标定位。Danelljan 等人 (2015) 提出空间正则化判别相关滤波器 (spatially regularized discriminative correlation filters, SRDCF) 算法,解决了训练样本周期性假设引起的边界效应,在学习阶段引入空间正则化分量,并且根据空间位置对相关滤波器系数进行惩罚。为了避免通过穷举尺度空间搜索估计目标大小, DSST (discriminative scale space tracking) 跟踪算法 (Danelljan 等, 2017b) 通过学习相关滤波器分别估计平移和尺度,提出了一种新的尺度自适应跟踪方法。ECO (efficient convolution operators) 算法 (Danelljan 等, 2017a) 重新优化了 DCF 框架的核心公式,引入因式分解的卷积操作和生成样本空间模型,大幅减少了参数量,解决了 DCF 框架过拟合和计算复杂性过高的问题,提高了 ECO 算法的速度和精度。

近年来,基于深度学习技术,特别是基于孪生网络的目标跟踪算法 (Zhu 等, 2018; Li 等, 2019a) 引起了广泛关注。基于孪生网络的目标跟踪算法将视觉目标跟踪问题转化为学习模板图像与候选搜索区域图像之间的通用相似度得分图,在主流的目标跟踪公开数据集中取得了不错的性能,运行速度也满足实时性要求。然而,绝大多数基于孪生网络的跟踪算法都有一个很大的局限性,即当跟踪目标被遮挡或出现较大外观变化时,跟踪性能通常会显著下降。针对现有基于孪生网络的跟踪算法存在的问题,本文提出一种高效、鲁棒的基于孪生网络的双模板融合目标跟踪算法 (siamese tracker with double template fusion, Siam-DTF), 改善当前孪生网络跟踪算法在目标遮挡和较大外观变化情况下跟踪性能下降问题,在 OTB2015 (object tracking benchmark 2015) (Wu 等, 2015)、VOT2016 (visual object tracking 2016) (Kristan 等, 2016) 和 VOT2018 (Kristan 等, 2018) 3 个大型目标跟踪数据集上均获得了最先进的跟踪结果,运行速度超过实时,可达 47 帧/s。本文工作的主要贡献有: 1) 对当前基于孪生网络的跟踪算法在目标遮挡和较大外观变化情况下跟踪效果

下降的原因进行分析,发现其核心原因是单一固定的初始模板。2) 为了克服单一模板的不足,提出一种新颖有效的双模板机制,提高了基于孪生网络的跟踪算法的性能。3) 提出一个外观模板搜索模块 (appearance template search module, ATSM), 在跟踪过程中为目标对象获取合适、高质量的外观模板,同时设计了有效的融合策略,从而获得更加鲁棒的跟踪结果。

1 相关工作

近年来,视觉目标跟踪得到了广泛研究,提出了许多提高跟踪性能的方法 (Smeulders 等, 2014; Henriques 等, 2015; Li 等, 2018a; Bhat 等, 2019; Wang 等, 2018, 2019; Danelljan 等, 2016)。

1.1 基于孪生网络的视觉目标跟踪

基于孪生网络的跟踪算法在精确性和效率上取得了不错的平衡。与传统的基于相关滤波的跟踪算法不同,基于孪生网络的目标跟踪算法将视觉跟踪描述为一个互相关问题,可以更好地利用深度网络支持端到端学习的优点。典型的孪生网络由两个具有共享参数的分支组成,即学习目标特征表示的模板分支和代表当前搜索区域的搜索分支。模板通常是从视频帧序列中的第 1 帧的标注框中得到,记为 z , 后续每一帧的搜索区域记为 x 。孪生网络以 z 和 x 两个分支作为输入,用一个离线训练好的共享权重的主干网络 φ 提取两个分支的特征,主干网络的参数为 θ 。对模板分支和搜索分支的特征进行卷积运算,可以得到当前帧的跟踪响应图,响应图上的数值代表了目标出现在每个位置的得分大小。响应图计算为

$$f_{\theta}(z, x) = \varphi_{\theta}(z) * \varphi_{\theta}(x) + b \quad (1)$$

式中, b 表示模拟相似度偏移的偏差项。式 (1) 相当于模板 z 在图像 x 上的执行穷尽搜索,得到每个位置的相似度得分。

一般来说,孪生网络通过从训练视频中搜集大量的图像对 (z, x) 和相应的真实标签 y 进行离线训练。训练过程中,主干网络的参数 θ 得到不断优化。为了使响应图 $f_{\theta}(z, x)$ 中的最大值与目标位置相匹配,通常在训练集中最小化逻辑损失 l , 即

$$\arg \min_{\theta} (y, f_{\theta}(z, x)) \quad (2)$$

在上述理论基础上,为了提高跟踪性能,对基于孪生网络的跟踪算法提出了许多改进,通常使用经典的相对较浅的 AlexNet (Krizhevsky 等, 2017) 作为框架的主干网络。SiameseFC (fully-convolutional siamese networks) 跟踪算法 (Bertinetto 等, 2016) 以图像对作为输入, 一个图像对包括一幅模板图像和一幅候选搜索区域图像, 然而, SiameseFC 算法无法准确估计目标尺度, 只能通过多尺度搜索处理尺度变化。SiamRPN (siamese region proposal network) 跟踪算法 (Li 等, 2018a) 通过在孪生网络之后引入区域候选网络 (region proposal network, RPN) 解决这一问题。RPN 网络最初用于目标检测 (Ren 等, 2017; Liu 等, 2016), 可以进行联合分类和回归, 并且支持端到端离线训练。DaSiamRPN (distractor-aware siamese region proposal network) 跟踪算法 (Zhu 等, 2018) 增加了一个干扰物感知模块, 并在在线跟踪过程中显式地抑制干扰物的得分。此外, Li 等人 (2019a) 提出 SiamRPN++ 跟踪算法, 采用一种有效的采样策略打破了空间不变性的限制, 将最初放置在中心的目标移动不同的像素, 使训练样本中目标的空间分布更加均衡, 消除了中心偏差的影响, 充分利用深层网络的表征能力, 首次在更深层的 ResNet50 (deep residual networks) (He 等, 2016) 主干网络上成功训练出高效的孪生网络。HASiam (hierarchical attention siamese network) 算法 (Shen 等, 2020) 构造了一个层次化的注意力孪生网络, 提高了跟踪性能。上述基于孪生网络的跟踪算法在速度和精度方面都取得了优异的性能。

1.2 特征融合

深层神经网络中不同层次的特征具有不同的代表能力, 浅层特征包含更多细节的空间信息, 高层特征包含更多抽象的语义线索。许多跟踪算法已经证明, 使用特征融合可以提高视觉跟踪算法的鲁棒性。C-RPN (siamese cascaded region proposal networks) 算法 (Fan 和 Ling, 2019) 使用一种多阶段的跟踪框架, 通过级联一系列 RPN, 将高层特征融合到低级 RPN 中。SiamRPN++ 算法 (Li 等, 2019a) 从 ResNet50 最后 3 个残差块中提取多级特征, 然后对 RPN 输出采用加权求和, 这种特征的分层聚合获得了可观的收益。

2 基于孪生网络的双模板融合目标跟踪方法

2.1 孪生网络跟踪算法性能下降的原因分析

随着深度学习和大规模视觉目标跟踪数据集的发展, 跟踪性能有了很大提高。虽然当前基于孪生网络的跟踪算法已经可以应付许多复杂的挑战, 但仍有一些困难的情况会导致跟踪失败。

为了进一步提高基于孪生网络的跟踪算法的鲁棒性, 本文分析了当前孪生网络跟踪框架的内在机制, 发现性能下降的原因在于使用了单一固定的初始模板。在当前的孪生网络跟踪框架下, 算法通常有初始模板 z 和搜索区域 x 两个输入分支。式 (1) 表明, 孪生网络跟踪相当于在给定模板 z 的前提下在当前帧 x 中找到与模板 z 最相似的位置。图 1 展示了当前孪生网络跟踪框架下单一固定模板的限制, 图 1 各子图中左上角数字表示图像处于序列的第几帧, 绿框为真实标注。在整个跟踪过程中, 飞行的鸟先后经历了云层遮挡和较大的外观变化。而现有的基于孪生网络的跟踪算法只使用第 1 帧的标注框作为模板, 在面对上述复杂情况时的跟踪性能自然会不可避免地下降。单一固定的初始模板无法很好地应对目标出现的各种复杂变化, 同时也限制了模型对于跟踪历史帧的深层特征的充分利用。

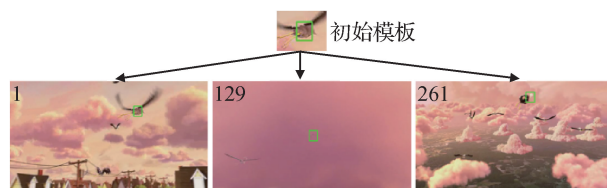


图 1 单一固定模板的局限性

Fig. 1 Limitation of the single fixed template

2.2 本文算法框架

为了克服上述限制, 提高基于孪生网络的跟踪算法的鲁棒性, 本文提出了一种新的双模板机制。除了使用第 1 帧的标注框作为初始模板外, 增加 1 个外观模板分支, 通过外观模板搜索模块, 在跟踪过程中为目标获取合适、高质量的外观模板。与现有基于孪生网络的跟踪算法一样, 初始模板从第 1 帧的标注框获取并且在整个跟踪过程中固定不变。而

外观模板则是根据历史帧的跟踪结果得到的,可以灵活调整,能适应目标物体的复杂外观变化。在双模板机制下,初始模板与外观模板相结合,二者起到互补的作用。最后,通过精心设计的融合模块可以充分利用双模板,提高了跟踪算法的鲁棒性。

如图2所示,本文提出的 Siam-DTF 算法框架由初始模板 z 、外观模板 z_a 和搜索区域 x 共3个分支组成, $\varphi(x)$ 表示特征提取。虚线区域是当前主流的基于孪生网络的跟踪算法框架,只使用单一固定的初始模板。响应图由式(1)计算得到。图2底部是本文添加的外观模板分支,通过外观模板搜索模块

可以获得与当前搜索区域 x 最对应的外观模板 z_a ,然后通过计算得到更精确的响应图,具体计算为

$$f_{\theta}(z_a, x) = \varphi_{\theta}(z_a) * \varphi_{\theta}(x) + b \quad (3)$$

式中, φ_{θ} 和 b 的含义与式(1)相同。

值得注意的是,本文提出的双模板机制同时利用了长期和短期记忆信息。初始模板 z 作为目标对象的真实标注,包含了对目标的长期记忆,而根据历史帧获得的外观模板 z_a 则包含更多的短期记忆。在整个在线跟踪过程中,本文算法融合了初始模板 z 和外观模板 z_a 的输出,优化了最终的跟踪结果,大幅提高了基于孪生网络的跟踪算法的鲁棒性。

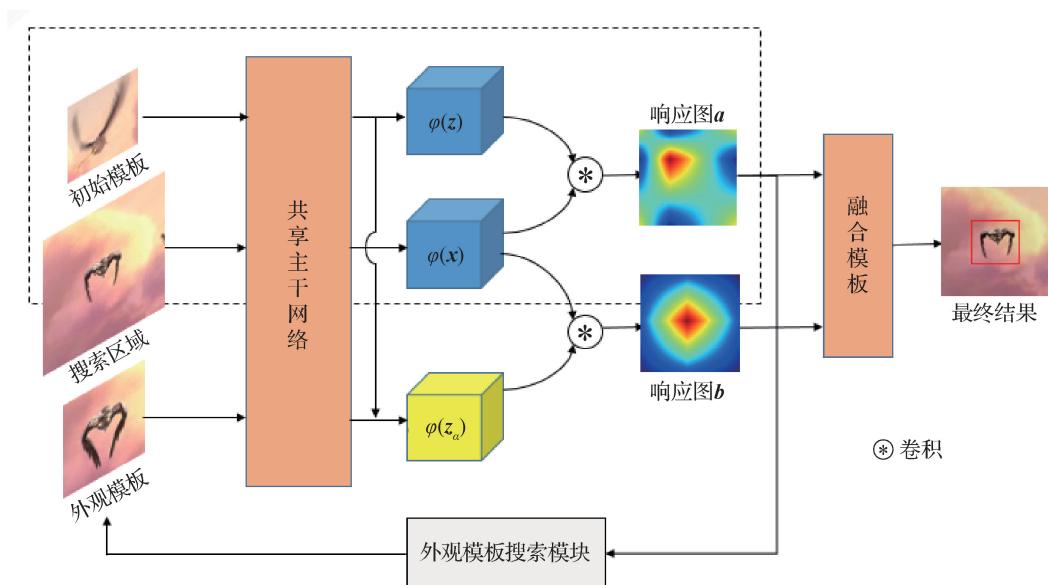


图2 本文提出的 Siam-DTF 算法框架

Fig. 2 Illustration of our proposed Siam-DTF framework

2.3 外观模板搜索模块

模板的选择是基于孪生网络的跟踪算法中的一个关键部分。好的模板可以产生准确稳定的结果,而不合适的模板则会严重降低跟踪算法的性能。本文提出的外观模板搜索模块的目的在于从历史帧中获得与当前搜索区域 x 最对应的外观模板 z_a 。

跟踪响应图的置信度得分表示模板与搜索区域在各位置的相似程度。考虑到相邻帧之间的图像内容差异很小,本文提出的外观模板搜索模块从具有最高置信度得分的历史帧中裁剪出一个新的最近模板 z' ,表示为

$$z' = \text{crop}(\arg\max_{\theta} f_{\theta}^c(z, x)) \quad (4)$$

式中, $\text{crop}(\cdot)$ 是裁剪操作,以该帧跟踪结果中心裁剪出新模板, $f_{\theta}^c(z, x)$ 是跟踪响应图的置信度得分。

图3展示了外观模板搜索模块(ATSM)的处理过程。

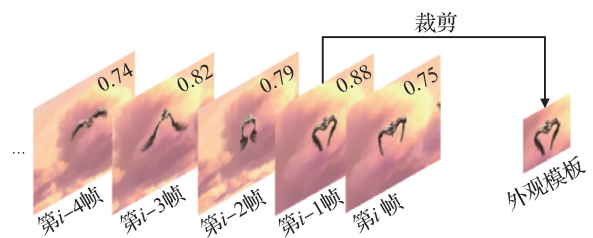


图3 外观模板搜索模块

Fig. 3 Illustration of the proposed appearance template search module

在大部分的遮挡、模糊或相似物干扰场景下,本文都能找到这样合适的 z' 作为外观模板。但是在少部分场景下,利用置信度得分裁剪出的图像块 z' 本身可能与当前目标图像外观不太一致,或者有模糊、

位置误差等,此时低质量的 z' 会对跟踪结果造成较大偏差,导致后续帧出现跟踪漂移。为了过滤掉这些低质量的模板,本文在外观模板上添加了置信度得分和回归框交并比(intersection-of-union, IoU)两个约束项,大幅提高了外观模板的质量。对于低质量的模板本文直接丢弃,采用稳妥可靠的初始模板,从而防止对跟踪结果造成负面影响。最终的外观模板 z_a 表示为

$$z_a = \begin{cases} z' & \text{IoU}(f_{\theta}^{\text{reg}}(z, x), f_{\theta}^{\text{reg}}(z', x)) \geq u, \\ & f_{\theta}^c(z', x) - f_{\theta}^c(z, x) \geq v \\ 0 & \text{其他} \end{cases} \quad (5)$$

式中, $f_{\theta}^{\text{reg}}(z, x)$ 表示使用模板 z 预测的回归框, $\text{IoU}(box_1, box_2)$ 表示 box_1 和 box_2 的交并比, u, v 是预先设定的阈值参数,用以避免回归框的漂移。

理论上可以在每一帧都获取新的外观模板,但实验分析和常识表明没有必要每一帧都激活双模板机制。图4显示了SiamRPN++跟踪算法在OTB2015数据集basketball序列上的置信度得分变化。可以看到,在序列前面部分的绝大多数帧中,跟踪算法的置信度得分非常高(大多高于0.97)。通过分析每一帧图像的内容,可以发现跟踪置信度得分的急剧下降通常意味着严峻的挑战。例如,图4中的点A和点B分别出现了遮挡、较大外观变化和干扰物。总之,跟踪算法的置信度得分变化很大程度上反映了当前帧的跟踪质量。

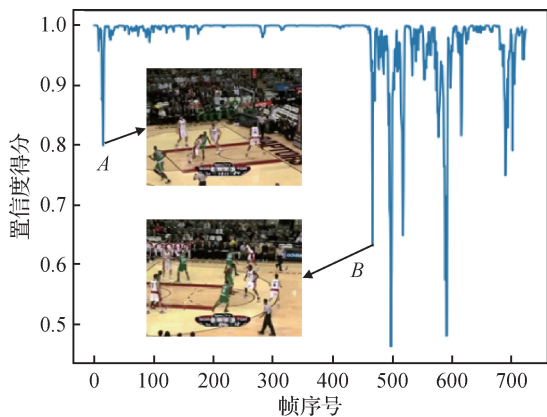


图4 SiamRPN++跟踪算法在OTB2015数据集basketball序列上的置信度得分变化

Fig. 4 The confidence change of SiamRPN++ tracker on sequence basketball in OTB2015 dataset

在上述分析启发下,本文引入一种自适应策略

确定是否执行外观模板搜索模块获得外观模板。具体地说,在当前帧的最大置信度得分以一定阈值 m 低于历史 N 帧的平均置信度得分时,本文算法开始启动外观模板搜索模块并通过式(5)获取外观模板。换句话说,这个稀疏更新机制保证本文算法在初始模板与当前帧不匹配时获取适当的外观模板。

2.4 融合模块

为了充分发挥双模板机制的优势,在这两个分支后面设计了一个融合模块,分别使用响应图融合和特征融合两种不同的融合方法组合初始模板和外观模板分支。

1) 响应图融合。最直接、简单和有效的方法是融合初始模板分支和外观模板分支的输出响应图。具体方法为根据式(1)和式(3)分别获得两个分支对应的响应图,然后直接在响应图上采用加权和,即

$$s = \alpha \cdot f_{\theta}(z, x) + (1 - \alpha) \cdot f_{\theta}(z_a, x) \quad (6)$$

式中, s 是最终融合后的响应图, $\alpha \in [0, 1]$ 表示响应图融合时的权重系数。通常,参数 α 表示初始模板与当前搜索区域的相似程度。 α 越小,初始模板与当前搜索区域的相似程度越低。如果 $\alpha = 1$,则式(6)退化为原来的式(1),意味着只使用初始模板。

响应图融合策略兼顾了初始模板和外观模板各自的优点,有效提高了视觉跟踪算法的鲁棒性。

2) 特征融合。已经有许多研究表明(Li等, 2019a; Fan和Ling, 2019),使用特征融合可以提高视觉跟踪算法的鲁棒性。浅层特征包含更多细节的空间信息,而高层特征包含更多抽象的语义线索。由于这两个分支具有相同的空间分辨率,可以先融合初始模板和外观模板的特征,然后利用融合后的模板特征和搜索区域特征计算最终的响应图。具体为

$$\begin{cases} \varphi_{\theta}(z, z_a) = \beta \cdot \varphi_{\theta}(z) + (1 - \beta) \cdot \varphi_{\theta}(z_a) \\ s = \varphi_{\theta}(z, z_a) * \varphi_{\theta}(x) + b \end{cases} \quad (7)$$

式中, s 是最终的响应图, $\beta \in [0, 1]$ 表示特征融合时的权重系数。 β 越大,则特征融合时初始模板所占的权重越大,反之外观模板占的权重越大。

对以上两种融合方法均进行了测试,实验结果表明,直接使用响应图融合即可获得明显的性能提升,在响应图融合的基础上添加特征融合后性能可以进一步得到提升。提出的Siam-DTF跟踪算法的

整体流程如下:

输入: 视频序列 $f_1, f_2, \dots, f_L$; 初始目标状态 s_1 。

输出: 目标在后续第 i 帧中的状态 s_i 。

初始化: 设置初始模板 z 并计算其特征 $\varphi_\theta(z)$; 记第 i 帧 f_i 的响应图最大得分为 f_i^c 。

对于 $2 \sim L$ 之间的每一个 i , 执行:

1) 根据前一帧的跟踪结果 s_{i-1} 获取当前帧的搜索区域 x_i 。

2) 开始跟踪: 使用式(1)计算由初始模板 z 生成的响应图 $f_\theta(z, x)$ 。

3) 使用外观模板搜索模块 ATSM。如果 $f_i^c < \frac{1}{N} \sum_{t=i-N}^{i-1} f_t^c - m$, 则通过式(4)和式(5)计算外观模板 z_a ; 使用式(1)计算由外观模板 z_a 生成的响应图 $f_\theta(z_a, x)$ 。

4) 使用融合模块, 通过式(6)和式(7)计算当前帧的跟踪结果 s_i 。

3 实验与分析

3.1 模型训练与评估

本文提出的 Siam-DTF 算法使用共享权值的主干网络提取各分支的特征, 可以省去额外的离线训练阶段, 直接使用已有的孪生网络模型。因此, 本文提出的 Siam-DTF 方法可以很容易地应用在任何基于孪生网络的跟踪算法上。在后续实验中, 本文将该方法应用到 SiamRPN++ 跟踪算法上。

为了验证 Siam-DTF 跟踪方法的性能, 本文在 OTB2015 (Wu 等, 2015)、VOT2016 (Kristan 等, 2016) 和 VOT2018 (Kristan 等, 2018) 3 个大型跟踪数据集上进行实验。跟踪算法遵循与 SiamRPN++ (Li 等, 2019a) 相同的配置过程, 并且在所有实验中使用统一、公开的评估标准。

3.2 实验细节

本文提出的 Siam-DTF 方法基于 Python3.7 和 PyTorch 0.4.1 实现, 硬件 GPU 为单张 GeForce GTX 2080 Ti 显卡。初始模板的输入尺寸为 127×127 像素, 搜索区域尺寸为 255×255 像素。外观模板搜索模块 (ATSM) 和融合模块的参数 N, m, u, v 和 α 根据经验分别设置为 7, 0.15, 0.6, 0.15 和 0.75。值得注意的是, 对于同一数据集中的所有视频序列, 所

有的超参数设置都固定不变。

本文提出的 Siam-DTF 方法高效, 主要有两个原因。首先, 只有在当前帧响应图的置信度得分满足一定条件时, 外观模板搜索模块 (ATSM) 才会激活, 在大部分简单场景下不会启动。因此这种稀疏触发的策略只带来很小的计算量, 很好地在速度和精度之间保持了平衡。其次, 大部分额外增加的计算量来自于外观模板特征 $\varphi_\theta(z_a)$ 的计算。由于 z_a 是从历史帧 x_k 中裁剪出来的, 历史帧 x_k 的特征 $\varphi_\theta(x_k)$ 已经计算过, 因此可以隐式地对 $\varphi_\theta(x_k)$ 进行线性变换获得外观模板的特征 $\varphi_\theta(z_a)$ 。基于以上两点原因, Siam-DTF 跟踪方法的性能在有效提升的同时, 能够保持以 47 帧/s 的速度高效运行。

3.3 与最新方法的比较

为了进一步验证 Siam-DTF 跟踪方法的性能, 与当前主流、先进的跟踪算法进行了对比。本文选取了一些最新的基于孪生网络的跟踪算法, 以及其他主流的基于判别相关滤波器 (DCF) 的跟踪算法进行比较。对比实验在 OTB2015 (Wu 等, 2015)、VOT2016 (Kristan 等, 2016) 和 VOT2018 (Kristan 等, 2018) 3 个大型公开跟踪数据集上完成。

3.3.1 在 OTB2015 数据集上的对比实验

OTB2015 数据集由 100 个具有挑战性的视频序列组成, 是一个广泛用于跟踪算法性能评估的公开数据集。跟踪算法的评估有两个指标: AUC (area under curve) 和精准度 (precision)。AUC 是成功率曲线下的面积, 成功率曲线显示了当阈值从 0 ~ 1 变化时成功帧的比率, 其中成功的帧指的是预测框和真实目标框的交并比大于给定的阈值。精准度曲线则代表了预测框中心和真实目标框中心距离在 20 像素范围内的帧数所占的百分比。实验中, 将本文方法与 SiamRPN++ (Li 等, 2019a)、SiamDW (deeper and wider Siamese networks) (Zhang 和 Peng, 2019)、C-RPN (Fan 和 Ling, 2019)、GradNet (gradient-guided network) (Li 等, 2019b)、ATOM (accurate tracking by overlap maximization) (Danelljan 等, 2019)、DaSiamRPN (Zhu 等, 2018)、Siam-BM (better match in Siamese network) (He 等, 2018)、ECO (Danelljan 等, 2017a) 和 ECO-HC (efficient convolution operators with hand-crafted feature) (Danelljan 等, 2017a) 等有代表性的跟踪算法进行比较。各算法在 OTB2015 数据集上的成功率曲线和精准度

曲线分别如图 5 和图 6 所示,具体指标的比较结果如表 1 所示。在所有比较的算法中,本文提出的 Siam-DTF 方法在 AUC 和精准度两个指标均排名第 1。与基础的跟踪算法 SiamRPN ++ (Li 等,2019a) 相比,Siam-DTF 方法在 AUC 指标上提高了 0.6%,精准度提高了 1.3%。这得益于本文提出的双模板机制,可以有效提高算法在面对遮挡、外观变化和相似干扰物时的性能。由于充分利用了跟踪历史帧的深度特征,Siam-DTF 的精准度比现有性能最好的相关滤波跟踪算法 ECO(Danelljan 等,2017a) 高 0.8%。

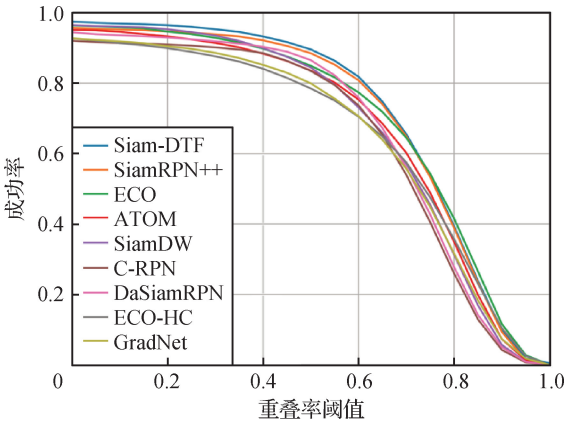


图 5 不同方法在 OTB2015 数据集上的成功率曲线
Fig. 5 Curves of rate of success by different methods on OTB2015 dataset

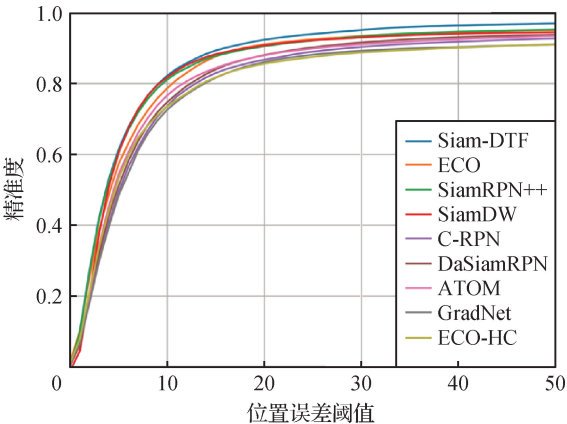


图 6 不同方法在 OTB2015 数据集上的精准度曲线
Fig. 6 Curves of precision by different methods on OTB2015 dataset

3.3.2 在 VOT2016 数据集上的对比实验

VOT2016 数据集由 60 个具有挑战性的视频序列组成,是近年来流行的目标跟踪数据集之一。VOT2016 数据集根据精确度和鲁棒性对算法的性能进行评估。精确度(accuracy)指成功跟踪时的平

表 1 不同方法在 OTB2015 数据集上的性能比较

Table 1 Comparison of performance among different methods on OTB2015 dataset

算法	AUC	精确度
SiamRPN ++	0.695	0.905
SiamDW	0.666	0.900
C-RPN	0.663	0.882
GradNet	0.639	0.861
ATOM	0.667	0.879
DaSiamRPN	0.658	0.880
Siam-BM	0.662	0.864
ECO	0.691	0.910
ECO-HC	0.643	0.856
Siam-DTF(本文)	0.701	0.918

注:加粗字体表示各列最优结果。

均重叠,鲁棒性使用失败(failure) 次数表示。整体性能评估通过期望平均重叠(expected average overlap, EAO) 衡量,该指标同时考虑了精确度和鲁棒性。实验中,将本文方法与 SiamRPN ++ (Li 等,2019a)、SiamDW (Zhang 和 Peng, 2019)、C-RPN (Fan 和 Ling, 2019)、SiamMask_E (Chen 和 Tsotsos, 2019)、ASRCF (adaptive spatially-regularized correlation filters) (Dai 等, 2019)、DaSiamRPN (Zhu 等, 2018)、SiamRPN (Li 等, 2018a)、ECO (Danelljan 等, 2017a) 和 ECO-HC (Danelljan 等, 2017a) 等 VOT 挑战赛排名靠前的跟踪算法进行比较,结果如表 2 所示。可以看

表 2 不同方法在 VOT2016 数据集上的性能比较

Table 2 Comparison of performance among different methods on VOT2016 dataset

算法	EAO	精确度	失败次数
SiamRPN ++	0.461	0.643	0.200
SiamDW	0.370	0.580	0.240
C-RPN	0.363	0.594	0.950
SiamMask_E	0.466	0.677	0.224
ASRCF	0.391	0.563	0.187
DaSiamRPN	0.411	0.610	0.220
SiamRPN	0.344	0.560	1.080
ECO	0.374	0.540	0.720
ECO-HC	0.322	0.530	1.080
Siam-DTF(本文)	0.477	0.640	0.172

注:加粗字体表示各列最优结果。

出, Siam-DTF 方法的 EAO 得分最高, 达到了 0.477, 失败次数最少, 仅为 0.172。在 EAO 得分上, 本文方法比基础算法 SiamRPN++ 提高了 1.6%, 比得分第 2 高的 SiamMask_E 算法提高了 1.1%。虽然相比于 SiamMask_E, 本文方法的精确度得分相对较低, 但是值得注意的是, SiamMask_E 算法是额外利用了分割标注信息预测旋转矩形框, 对非刚体目标会得到相对更精确的预测框。与 SiamRPN++ 算法相比, 本文方法的失败次数从 0.200 降低到 0.172, 充分表明了 Siam-DTF 方法在面临复杂环境挑战时具有更高的鲁棒性。

3.3.3 在 VOT2018 数据集上的对比实验

VOT2018 数据集由 60 个具有挑战性的视频序列组成, 数据集标注和评估标准与 VOT2016 数据集相同。实验中, 将本文方法与 SiamRPN++ (Li 等, 2019a)、SiamDW (Zhang 和 Peng, 2019)、C-RPN (Fan 和 Ling, 2019)、ATOM (Danelljan 等, 2019)、LADCF (learning adaptive discriminative correlation filters) (Xu 等, 2019)、DaSiamRPN (Zhu 等, 2018)、SiamRPN (Li 等, 2018a) 和 ECO (Danelljan 等, 2017a) 等 8 种最先进的跟踪算法进行比较, 具体指标的比较结果如表 3 所示。

表 3 不同方法在 VOT2018 数据集上的性能比较

Table 3 Comparisons of performance among different methods on VOT2018 dataset

算法	EAO	精确度	失败次数
SiamRPN++	0.413	0.602	0.239
SiamDW	0.370	0.580	0.240
C-RPN	0.363	0.594	0.950
ATOM	0.401	0.59	0.204
LADCF	0.389	0.503	0.159
DaSiamRPN	0.326	0.569	0.337
SiamRPN	0.244	0.490	0.460
ECO	0.28	0.484	0.276
Siam-DTF (本文)	0.403	0.608	0.258

注: 加粗字体表示各列最优结果。

从表 3 可以看出, 与 SiamRPN++ 和 ATOM 相比, 本文方法 Siam-DTF 表现出了具有竞争力的跟踪性能。此外, 在所有跟踪算法中, Siam-DTF 方法在

精确度得分上获得了最好的结果, 达到 0.608。精确度得分的显著提升进一步表明本文方法能有效辅助对跟踪目标的定位。

图 7 进一步展示了本文算法与 SiamRPN++、ECO 和 Siam-DTF 等跟踪算法在 3 个挑战性的视频序列 bird1、jumping 和 ironman 上的定性对比结果。

从图 7(a) 可以看出, 在跟踪目标遮挡时, SiamRPN++ 和 ECO 算法均出现了目标丢失, 当遮挡物消失后也无法恢复, 而本文方法则在遮挡刚开始发生以及遮挡消失后均可以跟踪目标。从图 7(b) 可以看出, 本文方法在面对目标出现运动模糊和较大外观变化时依然具有良好的跟踪效果, SiamRPN++ 算法则出现了跟踪漂移。从图 7(c) 可以看出, 本文方法在光线差、目标快速运动以及遮挡等复杂环境下的跟踪结果依然良好, 鲁棒性相比于 SiamRPN++ 和 ECO 算法更佳。通过在具有挑战性的视频序列上的定性对比表明, 本文提出的 Siam-DTF 方法在遮挡、模糊和目标出现较大外观变化的情况下仍然可以准确定位目标, 而 SiamRPN++ 和 ECO 算法均出现了跟踪目标丢失的情况。本文提出的双模板机制可以大幅提高孪生网络跟踪算法的鲁棒性, 除了定位更准确以外, 本文方法比其他跟踪算法预测的边界框也更加精确 (如图 6(c))。

3.4 消融实验

为验证外观模板搜索模块及置信度约束和 IoU 约束的作用, 使用 SiamRPN++ 和本文方法进行消融实验, 结果如表 4 所示。可以看出, SiamRPN++ 方法直接将 z' 作为外观模板, 没有取得预期的收益, 反而因为低质量的模板略微降低了算法的综合性能。而本文方法对外观模板添加了置信度约束和 IoU 约束后, 过滤掉了低质量的外观模板, 因此 Siam-DTF_b 和 Siam-DTF_c 分别取得了一定的性能提升。值得注意的是, IoU 约束相比于置信度约束带来的性能提升更明显, AUC 提升分别为 +0.4% 和 +0.2%, 精确度提升分别为 +1.0% 和 +0.4%, 这是因为 IoU 对回归框的位置和尺度约束更强。本文方法最终采用式(5)方案, 同时对外观模板添加了置信度约束和 IoU 约束, 可以有效提取出与当前帧目标外观一致的外观模板, 剔除低质量的模板, 因此获得了更大的性能提升。

3.5 超参数设置

本文通过理论分析和实验验证, 总结出超参数



图 7 在 3 个挑战性的视频序列上的跟踪结果比较
Fig. 7 Qualitative results of trackers on 3 challenging video sequences
((a) bird1 ; (b) jumping ; (c) ironman)

表 4 外观模板搜索模块消融实验
Table 4 Ablation experiment on appearance template search module

方法	外观模板及置信度约束和 IoU 约束	AUC	精度度
SiamRPN ++	未使用	69. 5	90. 5
Siam-DTF _a	外观模板	69. 4 (- 0. 1)	90. 5 (+ 0. 0)
Siam-DTF _b	外观模板 + 置信度约束	69. 7 (+ 0. 2)	90. 9 (+ 0. 4)
Siam-DTF _c	外观模板 + IoU 约束	69. 9 (+ 0. 4)	91. 5 (+ 1. 0)
Siam-DTF _{final}	外观模板 + 置信度约束 + IoU 约束	70. 1 (+ 0. 6)	91. 8 (+ 1. 3)

注:加粗字体表示各列最优结果;括号内数字表示相对于 SiamRPN ++ 的指标变化。

的建议取值区间如表 5 所示。参照表 5 进行超参数设置时,本文双模板机制可以得到更好的发挥,算法可以取得稳定、明显的性能提升。

4 结 论

本文分析了当前基于孪生网络的跟踪算法在面对目标出现遮挡及较大外观变化时的不足,提出了一个高效、鲁棒的双模板融合孪生网络跟踪算法

表 5 超参数建议取值区间
Table 5 Recommended value range for hyper-parameters

超参数	建议取值区间
N	[3 , 8]
m	[0. 1 , 0. 2]
u	[0. 6 , 0. 65]
v	[0. 1 , 0. 2]
α	[0. 7 , 0. 8]

Siam-DTF, 可以充分利用历史帧的信息, 获得灵活的、能适应目标外观变化的外观模板。这种新颖的双模板机制打破了传统基于孪生网络的跟踪算法中单一固定模板的局限性, 提高了算法在面对遮挡、外观变化和相似干扰物时的跟踪性能。本文提出的外观模板搜索模块能在初始模板与当前帧不匹配时有效地获得合适、高质量的外观模板, 最后使用的融合模块也充分结合了这两种模板的优点。在 3 个主流的目标跟踪公开数据集上的实验证明了本文方法的有效性, 在 3 个数据集上 Siam-DTF 均取得了优秀的跟踪结果。

本文提出的外观模板搜索模块取得了有效的性能提升, 但是从历史帧中裁剪出一个外观模板的做法相对简单, 双模板机制可以进一步扩充为多模板机制。后续工作将通过聚类的方式建立一个外观模板库, 每个外观模板即为对目标在不同状态下的典型表达, 通过融合多个外观模板和初始模板的结果, 进一步提升算法的鲁棒性。

参考文献 (References)

- Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional Siamese networks for object tracking//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56].
- Bhat G, Danelljan M, Van Gool L and Timofte R. 2019. Learning discriminative model prediction for tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South); IEEE: 6181-6190 [DOI: 10.1109/ICCV.2019.00628].
- Bolme D S, Beveridge J R, Draper B A and Lui Y M. 2010. Visual object tracking using adaptive correlation filters//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA; IEEE: 2544-2550 [DOI: 10.1109/CVPR.2010.5539960].
- Chen B X and Tsotsos J K. 2019. Fast visual object tracking with rotated bounding boxes [EB/OL]. [2020-09-30]. <https://arxiv.org/pdf/1907.03892.pdf>
- Dai K N, Wang D, Lu H C, Sun C and Li J H. 2019. Visual tracking via adaptive spatially-regularized correlation filters//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA; IEEE: 4665-4674 [DOI: 10.1109/CVPR.2019.00480].
- Danelljan M, Bhat G, Khan F S and Felsberg M. 2019. ATOM: accurate tracking by overlap maximization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA; IEEE: 4655-4664 [DOI: 10.1109/CVPR.2019.00479].
- Danelljan M, Bhat G, Khan F S and Felsberg M. 2017a. ECO: efficient convolution operators for tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA; IEEE: 6931-6939 [DOI: 10.1109/CVPR.2017.733].
- Danelljan M, Häger G, Khan F S and Felsberg M. 2017b. Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(8): 1561-1575 [DOI: 10.1109/TPAMI.2016.2609928].
- Danelljan M, Häger G, Khan F S and Felsberg M. 2015. Learning spatially regularized correlation filters for visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile; IEEE: 4310-4318 [DOI: 10.1109/ICCV.2015.490].
- Danelljan M, Robinson A, Khan F S and Felsberg M. 2016. Beyond correlation filters: learning continuous convolution operators for visual tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 472-488 [DOI: 10.1007/978-3-319-46454-1_29].
- Fan H and Ling H B. 2019. Siamese cascaded region proposal networks for real-time visual tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA; IEEE: 7944-7953 [DOI: 10.1109/CVPR.2019.00814].
- He A F, Luo C, Tian X M and Zeng W J. 2018. Towards a better match in Siamese network based visual object tracker//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany; Springer: 132-147 [DOI: 10.1007/978-3-030-11009-3_7].
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA; IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90].
- Held D, Guillory D, Rebsamen B, Thrun S and Savarese S. 2016. A probabilistic framework for real-time 3D Segmentation using spatial, temporal, and semantic cues//Proceedings of 2016 Conference on Robotics: Science and Systems. Cambridge, USA; MIT [DOI: 10.15607/RSS.2016.XII.024].
- Henriques J F, Caseiro R, Martins P and Batista J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy; Springer: 702-715 [DOI: 10.1007/978-3-642-33765-9_50].
- Henriques J F, Caseiro R, Martins P and Batista J. 2015. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3): 583-596 [DOI:

10.1109/TPAMI.2014.2345390]

- Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Čehovin L, Vojř T, Häger G, Lukežič A, Fernández G, Gupta A, Petrosino A, Memarmoghdam A, Garcia-Martin A, Montero A S, Vedaldi A, Robinson A, Ma A J, Varfolomeiev A, Alatan A, Erdem A, Ghanem B, Liu B, Han B, Martinez B, Chang C M, Xu C S, Sun C, Kim D, Chen D P, Du D W, Mishra D, Yeung D Y, Gundogdu E, Erdem E, Khan F, Porikli F, Zhao F, Bunyak F, Battistone F, Zhu G, Roffo G, Subrahmanyam G R K S, Bastos G, Seetharaman G, Medeiros H, Li H D, Qi H G, Bischof H, Possegger H, Lu H C, Lee H, Nam H, Chang H J, Drummond I, Valmadre J, Jeong J C, Cho J I, Lee J Y, Zhu J K, Feng J Y, Gao J, Choi J Y, Xiao J J, Kim J W, Jeong J, Henriques J F, Lang J, Choi J, Martinez J M, Xing J L, Gao J Y, Palaniappan K, Lebeda K, Gao K, Mikolajczyk K, Qin L, Wang L J, Wen L Y, Bertinetto L, Rapuru M K, Poostchi M, Maresca M, Danelljan M, Mueller M, Zhang M D, Arens M, Valstar M, Tang M, Baek M, Khan M H, Wang N Y, Fan N N, Al-Shakarji N, Miksik O, Akin O, Moallem P, Senna P, Torr P H S, Yuen P C, Huang Q M, Martin-Nieto R, Pelapur R, Bowden R, Laganière R, Stolkin R, Walsh R, Krah S B, Li S K, Zhang S P, Yao S Z, Hadfield S, Melzi S, Lyu S W, Li S Y, Becker S, Stuart Golodetz S, Kakanuru S, Choi S, Hu T, Mauthner T, Zhang T Z, Pridmore T, Santopietro V, Hu W M, Li W B, Hübner W, Lan X Y, Wang X M, Li X, Li Y, Demiris Y, Wang Y F, Qi Y K, Yuan Z J, Cai Z X, Xu Z, He Z Y and Chi Z Z. 2016. The visual object tracking vot 2016 challenge results//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 777-823 [DOI: 10.1007/978-3-319-48881-3_54]
- Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Zajc LČ, Vojř D T, Bhat G, Lukežič A, Eldesokey A, Fernández G, García-Martín Á, Iglesias-Arias Á, Alatan A A, González-García A, Petrosino A, Memarmoghdam A, Vedaldi A, Muhič A, He A F, Smeulders A, Perera A G, Li B, Chen B Y, Kim C, Xu C S, Xiong C Z, Tian C, Luo C, Sun C, Hao C, Kim D, Mishra D, Chen D M, Wang D, Wee D, Gavves E, Gundogdu E, Velasco-Salido E, Khan F S, Yang F, Zhao F, Li F, Battistone F, De Ath G, Subrahmanyam G R K S, Bastos G, Ling H B, Galoogahi H K, Lee H, Li H J, Zhao H J, Fan H, Zhang H G, Possegger H, Li H Q, Lu H C, Zhi H, Li H Y, Lee H, Chang H J, Drummond I, Valmadre J, Martin J S, Chahl J, Choi J Y, Li J, Wang J Q, Qi J Q, Sung J, Johnander J, Henriques J, Choi J, Van De weijer J, Herranz J R, Martínez J M, Kittler J, Zhuang J F, Gao J Y, Grm K, Zhang L C, Wang L J, Yang L X, Rout L, Si L, Bertinetto L, Chu L T, Che M Q, Maresca M E, Danelljan M, Yang M H, Abdelpakey M, Shehata M, Kang M, Lee N, Wang N, Miksik O, Moallem P, Vicente-Moñivár P, Senna P, Li P X, Torr P, Raju P M, Qian R H, Wang Q, Zhou Q, Guo Q, Martín-Nieto R, Gorthi R K, Tao R, Bowden R, Everson R, Wang R L, Yun S, Choi S, Vivas S, Bai S, Huang S P, Wu S H, Hadfield S, Wang S W, Golodetz S, Ming T, Xu T Y, Zhang T Z, Fischer T, Santopietro V, Štruc V, Wei W, Zuo W M, Feng W, Wu W, Zou W, Hu W M, Zhou W G, Zeng W J, Zhang X F, Wu X H, Wu X J, Tian X M, Li Y, Lu Y, Law Y W, Wu Y, Demiris Y, Yang Y C, Jiao Y F, Li Y H, Zhang Y H, Sun Y X, Zhang Z, Zhu Z, Feng Z H, Wang Z H and He Z Q. 2018. The sixth visual object tracking vot2018 challenge results//Proceedings of 2018 European Conference on Computer Vision. Munich, Germany; Springer: 3-53 [DOI: 10.1007/978-3-030-11009-3_1]
- Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]
- Lee K H and Hwang J N. 2015. On-road pedestrian tracking across multiple driving recorders. IEEE Transactions on Multimedia, 17(9): 1429-1438 [DOI: 10.1109/TMM.2015.2455418]
- Li B, Wu W, Wang Q, Zhang F Y, Xing J L and Yan J J. 2019a. SiamRPN ++: evolution of Siamese visual tracking with very deep networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA; IEEE: 4277-4286 [DOI: 10.1109/CVPR.2019.00441]
- Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018a. High performance visual tracking with Siamese region proposal network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA; IEEE: 8971-8980 [DOI: 10.1109/CVPR.2018.00935]
- Li P X, Chen B Y, Ouyang W L, Wang D, Yang X Y and Lu H C. 2019b. GradNet: gradient-guided network for visual object tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South); IEEE: 6161-6170 [DOI: 10.1109/ICCV.2019.00626]
- Li P X, Wang D, Wang L J and Lu H C. 2018b. Deep visual tracking: review and experimental comparison. Pattern Recognition, 76: 323-338. [DOI: 10.1016/j.patcog.2017.11.007]
- Liu L W, Xing J L, Ai H Z and Ruan X. 2012. Hand posture recognition using finger geometric feature//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba, Japan; IEEE: 565-568
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands; Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Shen J B, Tang X, Dong X P and Shao L. 2020. Visual object tracking by hierarchical attention Siamese network. IEEE Transactions on

- Cybernetics, 50 (7): 3068-3080 [DOI: 10.1109/TCYB. 2019. 2936503]
- Smeulders A W M, Chu D M, Cucchiara R, Calderara S, Dehghan A and Shah M. 2014. Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1442-1468 [DOI: 10.1109/TPAMI.2013.230]
- Wang Q, Teng Z, Xing J L, Gao J, Hu W M and Maybank S. 2018. Learning attentions: residual attentional Siamese network for high performance online visual tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4854- 4863 [DOI: 10.1109/CVPR. 2018. 00510]
- Wang Q, Zhang L, Bertinetto L, Hu W M and Torr P H S. 2019. Fast online object tracking and segmentation: a unifying approach//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 1328-1338 [DOI: 10.1109/CVPR.2019.00142]
- Wu Y, Lim J and Yang M H. 2015. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1834-1848 [DOI: 10.1109/TPAMI.2014.2388226]
- Xing J L, Ai H Z and Lao S H. 2010. Multiple human tracking based on multi-view upper-body detection and discriminative learning//Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE: 1698-1701 [DOI: 10.1109/ICPR. 2010.420]
- Xu T Y, Feng Z H, Wu X J and Kittler J. 2019. Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. IEEE Transactions on Image Processing, 28(11): 5596-5609 [DOI: 10.1109/TIP.2019.2919201]
- Zhang G C and Vela P A. 2015. Good features to track for visual SLAM//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1373-1382 [DOI: 10.1109/CVPR.2015.7298743]
- Zhang Z P and Peng H W. 2019. Deeper and wider Siamese networks for real-time visual tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 4586-4595 [DOI: 10.1109/CVPR. 2019. 00472]
- Zhu Z, Wang Q, Li B, Wu W, Yan J J and Hu W M. 2018. Distractor-aware Siamese networks for visual object tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 103-119 [DOI: 10.1007/978-3-030-01240-3_7]

作者简介



陈志良, 1995 年生, 男, 硕士研究生, 主要研究方向为计算机视觉与模式识别。

E-mail: zlchen@tongji.edu.cn



石繁槐, 通信作者, 男, 副教授, 主要研究方向为计算机视觉与模式识别。

E-mail: fhshi@tongji.edu.cn