计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):1−9 [doi: 10.15888/j.cnki.csa.006720]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的目标视频跟踪算法综述①

陈　旭, 孟朝晖

(河海大学 计算机与信息学院, 南京 211100)
通讯作者: 陈　旭, E-mail: chenxu19930614@163.com

摘　要: 深度学习理论在计算机视觉中的应用日趋广泛, 在目标分类、检测领域取得了令人瞩目的成果, 但是深度学习理论在目标跟踪领域的早期应用中, 由于存在跟踪时只有目标为正样本, 缺乏数据支持, 对位置信息依赖程度高等问题, 因而应用效果并不理想, 传统方法仍占据主流地位.近年来, 随着技术的不断发展, 深度学习在目标跟踪方向取得了长足的进步.本文首先介绍了目标跟踪技术的基本概念和主要方法, 然后针对深度学习在目标跟踪领域的发展现状, 从基于深度特征的目标跟踪和基于深度网络的目标跟踪两方面重点阐述了深度学习在该领域的应用方法, 并对近期较为流行的基于孪生网络的目标跟踪进行了详细介绍.最后对近年来深度学习在目标跟踪领域取得的成果, 以及未来的发展方向作了总结和展望.

关键词: 计算机视觉; 目标跟踪; 深度学习; 深度特征; 孪生网络

引用格式: 陈旭,孟朝晖.基于深度学习的目标视频跟踪算法综述.计算机系统应用,2019,28(1):1–9. http://www.c-s-a.org.cn/1003-3254/6720.html

Survey on Video Object Tracking Algorithms Based on Deep Learning

CHEN Xu, MENG Zhao-Hui

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Deep learning has achieved remarkable results in target detection and classification when applied to computer vision. But in the field of object tracking, the target is only considered as a positive sample. Being lack of data support and more dependent on the location information, deep learning did not achieve remarkable effect in the object tracking field, while the traditional methods still occupy the main position. However, with the development of technology, deep learning has progressed greatly in the direction of object tracking in recent years. This paper introduces the basic concept and the main methods of target tracking technology. Combined with the development of deep learning in recent years in the field of target tracking, the emphasis is on the basic approach of target tracking technology with tracking by deep feature and tracking based on deep network and introduces the recently popular target tracking based on Siamese network in detail. At the end, the achievements of deep learning in the field of target tracking in recent years and future development of object tracking are summarized and prospected.

Key words: computer vision; object tracking; deep learning; deep features; Siamese networks

随着技术的发展, 基于视频的跟踪技术在日常生活中得到了广泛的运用, 如在辅助驾驶系统 (ADSD)[1], 机器人视觉[2,3], 人机交互[4,5]、智能监控[6,7]等领域取得了较好的成果. 在计算机视觉领域, 基于视频的目标跟踪技术一直都是研究的要点和难点, 它基本的流程是通过在视频初始帧给定的目标检测框得到所要跟踪的目标, 然后通过一系列视觉方法得到目标的特征, 并在接下来的视频帧中成功定位到该目标, 从而得到目标

① 收稿时间: 2018-06-27; 修改时间: 2018-07-20; 采用时间: 2018-08-08; csa 在线出版时间: 2018-12-07

运动的速度、轨迹和方向等信息, 进一步应用在各个领域上去[8].

不同于目标检测, 由于基于视频的目标跟踪中常会遇到场景复杂, 种类繁多并且多变、影响参数过多, 以及由物体本身运动特性产生的遮挡、形变、位置急剧变化等情况, 因此, 如何找到一种行之有效的方法, 使其能够具有足够的鲁棒性处理以上可能存在的各种复杂情况成为了当下亟待解决的问题[9–11].

随着深度学习理论的不断发展, 基于深度学习对于目标强大的特征抽象能力以及对目标运动过程的拟合能力, 人们开始将深度学习应用在基于视频的目标跟踪领域上来. 本文从现有的基于深度学习的跟踪算法以及未来的发展趋势角度对基于深度学习的目标跟踪算法进行分析和展望.

1 相关技术

从目标跟踪的基本流程来看, 如图 1 所示, 主要分为 3 个部分: 外观建模部分, 搜索策略部分, 模型更新部分.

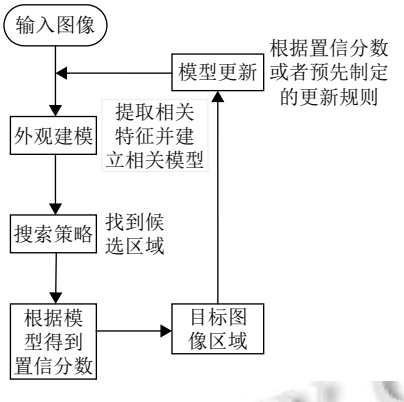


图 1 目标跟踪基本流程

外观建模阶段主要包含两方面的工作, 一是通过特征提取对物体外观进行抽象描述. 在特征提取阶段, 所提取的特征不仅需要对运动目标最具代表性特征进行完备描述, 基于实时性考虑, 通常也需要在计算速度上具备一定优势. 所提取的特征可以是单一的颜色[12,13]、纹理、灰度、几何特征, 也可以是多种特征的组合变换[14,15], 针对不同的跟踪目标在不同场景下的应用可以灵活采用不同的特征对物体的外观进行建模. 二是通过外观模型判断当前的搜索区域与目标的匹配程度, 计算置信分数从而确定候选区域是目标区域的可能性, 最终达到找到目标的效果.

目标跟踪的搜索策略阶段就是在跟踪的过程中, 根据当前物体所在位置, 找出帧与帧之间可能存在的位置关系, 从而在下一帧中得到物体可能所在的候选位置. 基本原理就是通过构建运动模型建立约束, 得到一组目标位置的可能候选区域. 常见的运动模型有滑动窗口、粒子滤波[16]、卡尔曼滤波[17]等.

模型更新阶段. 基于目标在跟踪过程中可能发生的变化, 跟踪问题在跟踪过程中需要一个在线更新机制实时更新目标的外观模型用以防止可能发生的漂移现象. 常用的方法主要分为逐帧更新、等间隔更新及启发式更新等方法.

目前的跟踪算法可以分为生成式模型和判别式模型两种. 生成式方法是从通过提取的相关特征中学习目标的外观模板, 在搜索区域寻找匹配程度最高的区域作为目标的方法. 其常用的方法主要有高斯混合模型[18], 贝叶斯网络[19], 马尔科夫模型等.

判别式方法则把跟踪问题转化为一个背景与前景的二分类问题, 通过提取的相关特征训练一个分类器, 在目标区域实现前景与背景的区分, 这种方法也被称之为 track by detection[20]. 经典的方法有 struck[21], TLD[22]方法等. 而在实际的跟踪过程中, 由于判别式方法同时兼顾了前景与背景的信息, 所达到的效果往往比生成式方法更为优秀.

在近几年的跟踪研究中, 基于速度与性能的双重考虑, 相关滤波方法 (Correlation Filter) 占据了一席之地. 相关滤波方法通过极高的效率解决岭回归问题完成对目标的判断. 2011 年, Bolme 等人[23]提出了最小输出均方误差和 (Minimum Output Sum of Squared Error filter, MOSSE) 滤波方法, 基于信号中相关性原理, 提取图像的灰度特征, 运用最小均方误差的原理找到使得目标能够得到最大响应的滤波器. Henriques 等人[24]提出了 KCF (Kernelized Correlation Filters) 滤波方法, 其利用循环矩阵的原理, 将相关滤波器的求解过程运用快速傅里叶变换转换到了频域, 同时提出了解决多特征融合的方法, 加入了 HOG (Histograms of Oriented Gradients) 特征[25]实现了跟踪效果的极大进步. Danelljan 等人[26]提出自适应颜色属性的 CN (Color Name) 方法, 将输入特征变为 11 种颜色特征, 然后将特征维数利用降维的思想转变成 2 维再训练相关滤波器. Danelljan 等人[27]提出 DSST (Discriminative Scale Space Tracking) 方法, 通过分别训练位置滤波器和尺度

滤波器得到目标位置的响应. Danelljan 等人[28]提出 SRDCF (Spatially Regularized Discriminative Correlation Filters) 方法, 利用高斯分布提出新的惩罚项, 并直接指定了尺度的分配, 在一定程度上解决了目标跟踪中针对不同尺度的情况.

随着深度学习在计算机视觉领域的不断发展, 在目标检测识别领域取得了令人瞩目的成果[29,30], 人们纷纷开始把眼光投向了一直是计算机视觉领域难点的目标跟踪方向[31,32]. 2013 年以来, 深度学习开始用于目标跟踪, 并且为这些问题提供了一些解决思路. 借助一些深度神经网络模型, 如自动编码机[33], 卷积神经网络[34]、循环神经网络[35]、孪生网络[36]等, 获取了一定的成果.

但是由于深度学习需要使用大量训练数据训练相应模型, 而在目标跟踪中所能使用的正负样本往往只有第一帧所给出的目标与背景, 训练样本数量严重不足, 而且在卷积网络中, 层次越深的网络对物体的抽象表达能力越强, 但在跟踪中, 过于深层次的网络在卷积池化的过程中逐渐丢失了物体的位置信息, 这对主要任务是完成对物体位置确定的跟踪来说往往不能达到预期的效果, 如何在保持一定语义信息的基础上充分保留空间信息也是深度学习在目标跟踪上的一个应用难点.

将深度学习应用在目标跟踪领域, 通常有两种思路. 为了体现深度特征对于目标物体强大的特征表示能力[37], 可以将深度特征代替传统的手工特征放入相关滤波器中以加强语义信息提高跟踪精度. 为了体现深度网络强大的拟合能力, 使用一个或者多个网络结构的组合实现目标跟踪也取得了先进的效果. 因此, 基于深度学习的目标跟踪通常分为基于深度特征的目标跟踪以及基于深度网络的目标跟踪.

2 目标跟踪

2.1 基于深度特征的目标跟踪

随着深度网络的发展, 其在目标分类方向的应用也越发的成熟. 最近几年的 VOT (Visual Object Tracking) 竞赛中, 相关滤波+深度特征的方法取得了先进的表现. 传统滤波方法中, 往往采用简单的灰度特征, Hog 特征, 亦或是相关的光流特征[38,39]和颜色特征. 这些手工制作的传统特征往往具有很丰富的目标信息, 但是无法提取高级的语义信息, 而且需要很强的先验信息, 往往针对特定的场景具有很强的适应性, 但在环境复杂的跟踪过程中, 物体的快速形变、遮挡等因素将导致这些传统的人工特征变化过为剧烈, 导致跟踪的失败. 随着深度学习的发展, 其强大的学习能力以及优秀的特征表达能力在计算机视觉的其他领域如在检测与识别中展现了巨大的潜力, 人们由此意识到了深度特征对于目标所具有的强大表示能力. 因此, 人们开始将目标放在了鲁棒性更强的卷积神经网络 (CNN) 特征上来.

Ma 等人[40]提出了 HCF (Hierarchical Convolutional Features for visual tracking) 方法, 利用已知的图像位置, 根据其对应的 Conv3-4, Conv4-4, Conv5-4 特征, 训练三个不同的相关滤波器. 在下一帧中与同样在相关区域范围内的相关层的特征滤波得到响应的位置分数. 通过三层位置的最大响应点做逐层精细的位置预测, 并以最终最底层的带有最多空间位置信息的预测结果作为输出.

Qi 等人[41]利用集成学习的思想, 提出了 HDT (Hedged Deep Tracking) 方法, 将许多个追踪器结合在一起获得一个更强的追踪器. 在追踪过程中, 基于上一帧图像的目标位置裁剪后, 利用 VGG16[42–44]提取出 6 个深度特征, 并使用这些深度特征训练独立的相关滤波器计算各自的响应. 利用每个弱跟踪器初步估计其目标位置, 通过自适用权重算法, 将所有的弱跟踪器集成一个强跟踪器. 在在线跟踪过程中, 通过自适应性算法对所有弱跟踪器计算加权平均损失, 由此计算每个弱跟踪的权重, 最小化弱跟踪器的累计误差.

Danelljan 等人[45]提出了连续卷积跟踪算子 (Continuous Convolution Operators for visual Tracking, CCOT). 基于特征融合改进的基础上, 通过在连续的分辨率序列中学习, 创建时域连续的的相关滤波器, 可以将不同分辨率的特征图作为滤波器的输入, 使得传统特征和深度特征能够深度结合, 最后, 融合多个响应, 得到目标的估计位置.

针对 CCOT 处理高维特征需要训练多个滤波器, 为了提高时间效率和空间效率, Danelljan 等人[46]提出了高效卷积算子 (Efficient Convolution Operators for tracking, ECO). ECO 构造一组更小的滤波器, 有效的利用矩阵分解操作, 降低了模型的大小, 防止过高的维度导致的效率低下和过拟合问题. 同时 ECO 使用高斯混合模型表示不同的目标外观, 使得训练集具有多样性, 防止对连续数帧样本的过拟合. 同时改变模板更新的策略, 降低更新频率, 提高效率.

2.2 基于深度网络的目标跟踪

深度特征加上相关滤波的方法在速度和精度上在一定程度上都有所保证, 但是考虑到深度网络强大的拟合能力与适应能力, 人们开始了对于完整的深度网络在目标跟踪上的研究. 但是在目标跟踪领域, 深度学习仍然面对着很大的挑战. 基于深度网络的目标跟踪不得不面对两个至关重要的难题. (1) 目标跟踪的样本数量严重不足, 在目标跟踪中, 往往只有初始帧的目标框, 这对需要基于大量数据才能达到优异效果的深度学习来说是一个很严重挑战. (2) 在注重实时性的目标跟踪领域, 具有比较理想效果的网络往往需要较大的计算量, 即便在 GPU 环境下也很难做到实时的效果. 针对这些问题, 一些基于深度网络的目标跟踪方法陆续被提出.

在基于深度学习的目标跟踪中, 在处理单目标跟踪问题的时候, 不需要太大的网络, 在卷积网络中, 一般底层包含更多的空间信息, 而高层包含更多的语义信息, 与目标检测不同, 在网络结构越深的情况下, 语义信息越来越抽象, 但同时会其所包含的空间信息也将被稀释, 不利于获取目标跟踪中最需要用到的物体的空间信息. 同样, 基于深度网络的目标跟踪本质上仍然是判别式模型, 所以目标跟踪只需要区分相应的两个类别, 即前景跟后景, 所以并不需要太大的网络. 而在整个视频中, 所需要追踪的目标也同样较小, 其输入尺寸自然也小, 综上并基于网络速度考虑, 目标跟踪所使用的深度网络不会太深太大.

基于迁移学习的思想, 采用 VGG-M 的部分结构, Nam 等人[47]提出多域卷积网络 (Multi-Domain convolutional neural Networks, MDNet). 如图 2 所示, 整个网络只运用 3 层卷积和两层全连接最后用以实现前景和背景的二分类任务. 跟踪的整体思路是运用不同视频通过目标检测网络提取所有运动物体的特征, 将用于目标分类问题的特征转移到跟踪领域上来, 通过 K 个视频交替训练, 每段视频训练得到独立的第六层全连接层 (fc6 层) 并且获得前面的共享层. 当测试一个新视频时, 随机初始化一个 fc6, 通过第一帧的正负样本得到新的网络权重. 在跟踪时, 只需要由跟踪过程中产生的正负样本在线微调 fc4–fc6 即可实现目标的跟踪. 基于深度网络的目标跟踪在速度上很难达到实时的效果, 但在准确率却达到了 state-of-the-art 的效果, 夺得了 VOT15 的冠军.

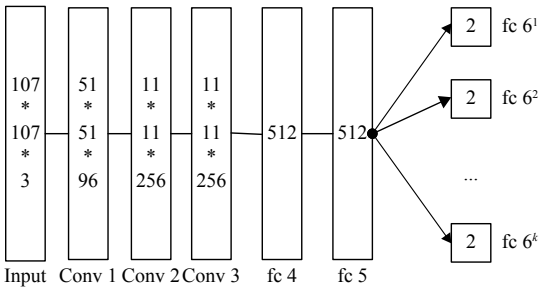
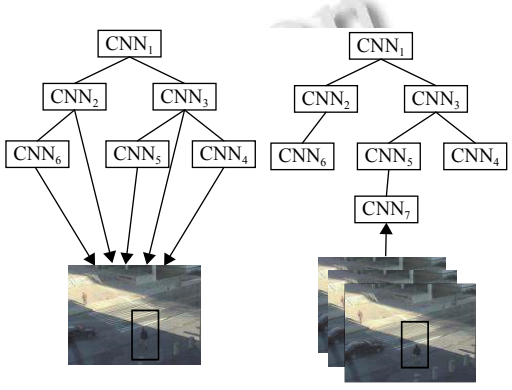


图 2 多域卷积网络 (MDNet)

Wang 等人[48]针对跟踪中无法使用大量训练样本的情况, 将 CNN 的训练过程看成是一个集成学习的过程, 将能够输出对应特征图的每个卷积核当做基准学习器. 通过 VGGnet 得到预训练的特征图, 经过共有两层卷积层结构的 CNN-A 网络, 每一个通道的卷积层卷积核都将被看做一个基学习器, 通过相应的损失函数, 得到热度图. 从这些基学习器中选出训练误差最小的放入集合中, 其余放入备选集合中. 当新的训练样本到来时, 使用随机梯度下降方法对所有的分类器进行更新, 当训练误差高于给定阈值时, 便从候选集合找出训练误差最小的基分类器. 由于 CNN 每个通道都参与了训练过程, 最终训出来的网络具有更好的泛化性能.

Nam 等人[49]针对模型应该随着目标的变化而稳定变化的出发点, 提出了树结构 CNN, 如图 3 所示.



(a) 状态估计 (b) 模型更新

图 3 树形结构的 CNN (T-CNN)

同时考虑了多个卷积神经网络, 用其线性加权组合来确定目标的位置. 为减少运算, 使用一个 CNN 时, 共享前三层卷积层, 只需要保存全连接层即可. 同时, 每十次跟踪, 便创建一个新的 CNN 节点, 同时从原先的集合中选择使新节点可靠性最高的节点作为其父节点, 并对其全连接层微调, 并删除最老的节点. 同样由于计算量过于庞大, 很难达到实时效果, 但在识别精度

上达到了 VOT2016 竞赛最先进的表现.

Wang 等人[50]针对 CNN 的特性, 不仅仅是把 CNN 特征当成一个黑盒的特征提取器, 通过对不同层的 CNN 特征进行了大量的分析实验, 使用 VGG16 网络中第十个卷积层和第 13 个卷积层分别送入两个只具有简单网络结构的 SNet 和 GNet, 前者对于前景后景的区分较为敏感, 后者对区分类别信息更于敏感. 两个网络分别输出相应的前景热图, 最后通过对干扰项的判定决定采用哪种热图, 从而输出目标的位置.

Han 等人[51]基于 bagging 集成[52]的在线跟踪思路, 提出 BranchOut 方法. 网络工作类似于 MDNet, 但是与 MDNet 不同的是, BranchOut 由三个卷积层和多个全连接层的分支组成. 网络中每个独立的分支 (Branch) 有着不同数量的全连接层, 数量一般为 1 或者 2, 用来保证目标的可抽象能力. 同时在跟踪过程中, 根据伯努利分布选择相应随机 Branch 子集来训练网络. 在没有预训练的情况下 BranchOut 也达到了 state-of-art 的效果.

2.3 算法小结

针对以上深度学习方法, 将上文提到的部分算法在 OTB2015[53]上进行相关性能的分析, 如表 1 所示. 其中, Precision 指标表示实际目标中心位置与通过跟踪算法得到的目标中心的偏差, Success 指标表示通过跟踪算法得到的目标区域与实际目标局域的重叠比.

表 1 深度学习跟踪算法表现

算法名称	Precision	Success
MDNet	0.909	0.678
ECO	0.899	0.680
CCOT	0.898	0.871
DeepSRDCF[54]	0.851	0.635
HDT	0.848	0.564
SRDCF	0.789	0.598
DSST	0.680	0.513
KCF	0.696	0.477

不难看出, 无论是基于深度特征还是直接使用深度网络的目标跟踪过程, 所利用的都是通过具有大量参数的网络所训练出的对物体特征的强大的表示能力, 对比相应的传统手工特征, 在跟踪精度上具有绝对的领先优势, 但是相应的考虑到深度学习所涉及的计算量, 对于绝大多数采用深度学习方法的目标跟踪算法其实时速率往往达不到 10 fps, 这对于只有达到实时效果才能产生实用价值的目标跟踪领域来说, 还需要一段时间的发展才能达到理想的效果.

3 基于孪生网络的目标跟踪发展

对于采用深度学习的目标跟踪方法而言, 由于计算参数量过于庞大, 虽然在精度上有着无可比拟的优势, 但是在实际应用以及相关实践上却难以发挥功效. 而对于深度学习的方法来说, 所耗费的大部分计算时间都集中在在线更新时所需要计算的反向传播的过程上, 而离线训练避免微调只计算前向传播的过程所花费的时间在跟踪领域是完全可以接受的, 因此同样基于相关性思想的孪生全卷积网络也成为了近几年研究的热点方向. 针对相关滤波处理模型快速变化能力差, 同时其采用循环矩阵所造成的边界效应难以解决, 基于深度网络思想, 人们提出了交叉相关的思想, 使用卷积操作来代替滑动窗口检测.

Bertinetto 等人[55]基于孪生全卷积网络提出了 SiameseFC (Fully-Convolutional Siamese networks for object tracking) 的网络结构用于目标跟踪, 如图 4 所示, 将一对图像通过具有同样结构的 2 个 CNN 网络得到特征的降维映射, 然后通过卷积实现相关性的计算, 得到目标位置的响应. 在最后的响应图上插值回原图大小实现目标的定位. 同时, SiameseFC 也采用了跟踪问题中常采用的尺度自适应方法, 在 3 个尺度上达到了 86 fps, 在 5 个尺度上也达到了 58 fps 的效果. 由于采用了相关性计算的方法, 采用离线训练的方式, 避免在线微调, 使得网络速度达到了令人满意的效果.

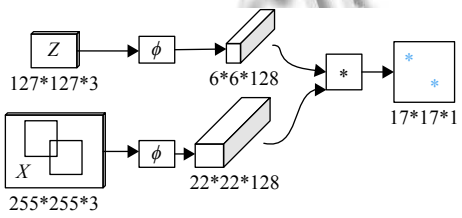


图 4 孪生全卷积网络 SiameseFC

基于相关滤波的方法近年来表现上佳, 考虑到基于 CNN 的相关滤波方法无法实现相应的端到端的快速学习, 基于 CNN 的随机梯度下降等方法效率低速度慢, 而基于傅里叶变换的相关滤波方法只需要使用快速傅里叶变换以及较少的参数运算便可以得到岭回归问题的解. 基于这种原因, Valmadre 等人[56]将孪生全卷积网络与相关滤波方法相结合, 以求达到性能上的进步. 基于 SiameseFC 的结构添加了 CF 层, 通过将卷积的计算转换到频域内计算, 使得相关滤波器的损失能够以反向传播的方式参与到网络的训练过程, 也就是将 CF 解释为可微的 CNN 层. 跟踪时使用 CF 高效的

在线学习效率每帧生成一个模板, 性能相较 SiameseFC 有了一定的进步.

Wang 等人[57]使用 CNN 特征代替 DCF (Discriminative Correlation Filters)[24]中的 HOG 特征, 使用 CNN 来学习最适合 DCF 使用的特征, 结合 CCOT 的特征插值方法, DSST 的尺度估计方法, 将 DCF 学习的过程融入反向传播, 可以在通道数较少的情况下得到鲁棒的跟踪效果.

由于 SiameseFC 使用固定的网络深度提取目标特征, 对于容易判断的目标相对于使用底层特征或相应的手工特征来说计算量过于庞大. 为了在提高计算效率并且保证精度, Huang 等人[58]提出了 EAST (EArly-Stopping Tracker) 跟踪算法. 针对不同情况灵活使用不同特征. 基本框架使用 SiameseFC 的互相关, 特征级联的方式为从灰度特征到 HOG 特征再到第一层卷积特征一直到第五层卷积特征. 首先使用速度较快的特征检测目标, 得出置信度. 根据阈值大小, 当简单特征无法正确判别时, 通过 Q-Learing[59]的方法加入更加深层次的的网络特征获取目标的高级语义特征解决当目标外观在一定程度上发生形变遮挡等干扰情况.

针对 SiameseFC 网络的面对目标显著变化效果差, 为提升网络的泛化效果, 基于网络的深层特征表示更高级的语义含义, He 等人[60]提出了具有语义分支和外观分支的 SA-Siam 网络. 如图 5 所示, 其中外观分支 (A-Net) 基本结构与 SiameseFC 相同, 获得目标位置的响应. 语义分支 (S-Net) 选择改变步长的 AlexNet[61]的 Conv-4 及 Conv-5 特征, 连接多层特征, 通过第一帧确定的注意力权重选择多通道特征中对该目标表达影响程度高的通道, 再与搜索区域卷积得到响应热图, 在跟踪时使用加权求和方法得到最后的目标响应位置.

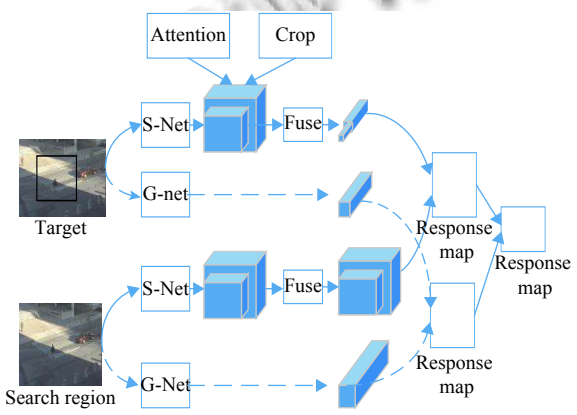


图 5 双重孪生网络 (SA-Siam)

同样基于提升 SiameseFC 的泛化能力, Wang 等人[62]提出了 RASNet (Residual Attentional Siamese Network for high performance online visual tracking). 基于注意力机制, RASNet 采用一般注意力权重, 通道注意力权重, 残差注意力权重, 将 3 种注意力权重融合后再与目标搜索区域进行互相关, 得到目标位置映射.

4 结论与展望

随着深度学习技术在计算机视觉领域的不断发展, 深度学习以其强大的模型学习能力取得了越来越显著的效果. 但是, 由于深度学习对数据依赖性强以及目标跟踪中数据量不足的特点, 深度学习在目标跟踪领域还有很长的一段路要走. 但是, 从近几年的 VOT 竞赛结果中可以看出, 随着跟踪技术的不断发展, 基于数据驱动的学习方式在跟踪领域也必将占据着越来越重要的作用. 本文对于目前目标跟踪领域存在的问题以及对于未来发展方向的展望, 总结如下:

(1) 相关滤波方法仍是现今应用能力最强的方法. 深度特征+相关滤波方法在性能上近年来取得了显著的效果. 如何最大限度利用深度特征对于目标强大的特征表达能力并运用实时性较高的滤波算法是今年来滤波类算法提高精度的重点方向之一.

(2) 将用来分类的网络迁移到跟踪领域达到 tracking-by-detection 在目前来说还远远没有达到深度学习在其他如目标检测、分割等领域的应用成就, 随着现代科学技术的发展, 在注重实时性的目标跟踪领域发展还需要一定时间, 但不难看出, 基于深度网络的目标跟踪在跟踪精度上具有较大优势, 目标跟踪的发展应是速度与精度共同提升的过程, 因此, 进一步强化和改进深度网络在目标跟踪上的应用对于目标跟踪的发展意义重大.

(3) 对于如何实现端到端的跟踪效果还需要不断优化. 近年来基于 Siamese 网络架构利用相关性思想的跟踪方法迅速发展, 在一定程度上解决了神经网络计算量大、效率低、速度慢的问题, 给人们带来了解决相关问题的新思路, 该方法正逐渐成为近年来目标跟踪的热门方向之一, 对于目标跟踪技术的发展有着十分积极的作用.

参考文献

1 Rios-Cabrera R, Tuytelaars T, van Gool L. Efficient multi-

camera vehicle detection, tracking, and identification in a tunnel surveillance application. Computer Vision and Image Understanding, 2012, 116(6): 742–753. [doi: 10.1016/j.cviu.2012.02.006]

2 Desouza GN, Kak AC. Vision for mobile robot navigation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 237–267. [doi: 10.1109/34.982903]

3 Bonin-Font F, Ortiz A, Oliver G. Visual navigation for mobile robots: A survey. Journal of Intelligent and Robotic Systems, 2008, 53(3): 263–296. [doi: 10.1007/s10846-008-9235-4]

4 Erol A, Bebis G, Nicolescu M, *et al*. Vision-based hand pose estimation: A review. Computer Vision and Image Understanding, 2007, 108(1-2): 52–73. [doi: 10.1016/j.cviu.2006.10.012]

5 Mitra S, Acharya T. Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(3): 311–324. [doi: 10.1109/TSMCC.2007.893280]

6 Haritaoglu I, Harwood D, Davis LS. W4: Real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 809–830. [doi: 10.1109/34.868683]

7 Siebel NT, Maybank SJ. The advisor visual surveillance system. Proceedings of ECCV 2004 Workshop "Applications of Computer Vision ". Prague, Czech Republic. 2004. 103–111.

8 Li X, Hu WM, Shen CH, *et al*. A survey of appearance models in visual object tracking. ACM Transactions on Intelligent Systems and Technology, 2013, 4(4): 58.

9 Wu Y, Lim J, Yang MH. Online object tracking: A benchmark. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 2411–2418.

10 Guan H, Xue XY, An ZY. Advances on application of deep learning for video object tracking. Acta Automatica Sinica, 2016, 42(6): 834–847.

11 Guan H, Xue XY, An ZY. Video object tracking via visual prior and context information. Journal of Chinese Computer Systems, 2016, 37(9): 2074–2078.

12 Pérez P, Hue C, Vermaak J, *et al*. Color-based probabilistic tracking. Proceedings of the 7th European Conference on Computer Vision Copenhagen. Denmark. 2002. 661–675.

13 Possegger H, Mauthner T, Bischof H. In defense of color-based model-free tracking. Proceeding of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 2113–2120.

14 Li Y, Zhu JK. A scale adaptive kernel correlation filter tracker with feature integration. In: Agapito L, Bronstein MM, Rother C, eds. European Conference on Computer Vision. Cham: Springer. 2015. 254–265.

15 Bertinetto L, Valmadre J, Golodetz S, *et al*. Staple: Complementary learners for real-time tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1401–1409. [doi: 10.1109/CVPR.2016.156]

16 Chang C, Ansari R. Kernel particle filter for visual tracking. IEEE Signal Processing Letters, 2005, 12(3): 242–245. [doi: 10.1109/LSP.2004.842254]

17 Zhang ZT, Zhang JS. A new real-time eye tracking based on nonlinear unscented Kalman filter for monitoring driver fatigue. Journal of Control Theory and Applications, 2010, 8(2): 181–188. [doi: 10.1007/s11768-010-8043-0]

18 Wang HZ, Suter D, Schindler K, *et al*. Adaptive object tracking based on an effective appearance filter. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(9): 1661–1667. [doi: 10.1109/TPAMI.2007.1112]

19 Murphy KP. Dynamic bayesian networks: Representation, inference and learning[Ph.D. thesis]. Berkeley: University of California, 2002.

20 Yang HX, Shao L, Zheng F, *et al*. Recent advances and trends in visual tracking: A review. Neurocomputing, 2011, 74(18): 3823–3831. [doi: 10.1016/j.neucom.2011.07.024]

21 Hare S, Golodetz S, Saffari A, *et al*. Struck: Structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2096–2109. [doi: 10.1109/TPAMI.2015.2509974]

22 Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(7): 1409–1422. [doi: 10.1109/TPAMI.2011.239]

23 Bolme DS, Beveridge JR, Draper BA, *et al*. Visual object tracking using adaptive correlation filters. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 2544–2550.

24 Henriques JF, Caseiro R, Martins P, *et al*. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583–596. [doi: 10.1109/TPAMI.2014.2345390]

25 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of 2005 IEEE Computer

Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005. 886–893.

26 Danelljan M, Khan FS, Felsberg M, *et al*. Adaptive color attributes for real-time visual tracking. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1090–1097.

27 Danelljan M, Häger G, Khan FS, *et al*. Accurate scale estimation for robust visual tracking. Proceedings of British Machine Vision Conference. Nottingham, UK. 2014. 65.1–65.11.

28 Danelljan M, Häger G, Khan FS, *et al*. Learning spatially regularized correlation filters for visual tracking. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4310–4318.

29 Kuen J, Lim KM, Lee CP. Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle. Pattern Recognition, 2015, 48(10): 2964–2982. [doi: 10.1016/j.patcog.2015.02.012]

30 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

31 Li HX, Li Y, Porikli F. Deeptrack: Learning discriminative feature representations online for robust visual tracking. IEEE Transactions on Image Processing, 2016, 25(4): 1834–1848. [doi: 10.1109/TIP.2015.2510583]

32 Wang NY, Wang JD, Yeung DY. Online robust non-negative dictionary learning for visual tracking. Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia. 2013. 657–664.

33 Vincent P, Larochelle H, Bengio Y, *et al*. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland. 2008. 1096–1103.

34 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

35 Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 2342–2350.

36 Bromley J, Guyon I, LeCun Y, *et al*. Signature verification using a “siamese” time delay neural network. Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, CO, USA. 1993. 737–744.

37 Hong S, You T, Kwak S, *et al*. Online tracking by learning discriminative saliency map with convolutional neural network. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 597–606.

38 Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence. Vancouver, BC, Canada. 1981. 674–679.

39 Horn BKP, Schunck BG. Determining optical flow. Artificial Intelligence, 1981, 17(1-3): 185–203. [doi: 10.1016/0004-3702(81)90024-2]

40 Ma C, Huang JB, Yang XK, *et al*. Hierarchical convolutional features for visual tracking. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 3074–3082.

41 Qi YK, Zhang SP, Qin L, *et al*. Hedged deep tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4303–4311.

42 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

43 Russakovsky O, Deng J, Su H, *et al*. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

44 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv: 1412.7062, 2014.

45 Danelljan M, Robinson A, Khan FS, *et al*. Beyond correlation filters: Learning continuous convolution operators for visual tracking. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands. 2016. 472–488.

46 Danelljan M, Bhat G, Khan FS, *et al*. ECO: Efficient convolution operators for tracking. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6931–6939.

47 Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4293–4302.

48 Wang LJ, Ouyang WL, Wang XG, *et al*. STCT: Sequentially training convolutional networks for visual tracking.

Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1373–1381. [doi: 10.1109/CVPR.2016.153]

49 Nam H, Baek M, Han B. Modeling and propagating cnns in a tree structure for visual tracking. arXiv: 1608.07242, 2016.

50 Wang LJ, Ouyang WL, Wang XG, *et al*. Visual tracking with fully convolutional networks. Proceeding of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 3119–3127.

51 Han B, Sim J, Adam H. BranchOut: Regularization for online ensemble tracking with convolutional neural networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 521–530. [doi: 10.1109/CVPR.2017.63]

52 Breiman L. Bagging predictors. Machine Learning, 1996, 24(2): 123–140.

53 Wu Y, Lim J, Yang MH. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834–1848. [doi: 10.1109/TPAMI.2014.2388226]

54 Danelljan M, Häger G, Khan FS, *et al*. Convolutional features for correlation filter based visual tracking. Proceedings of 2015 IEEE International Conference on Computer Vision Workshop. Santiago, Chile. 2015. 621–629.

55 Bertinetto L, Valmadre J, Henriques JF, *et al*. Fully-convolutional siamese networks for object tracking. In: Hua G, Jégou H. eds. European Conference on Computer Vision. Cham: Springer. 2016. 850–865.

56 Valmadre J, Bertinetto L, Henriques J, *et al*. End-to-end representation learning for correlation filter based tracking. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 5000–5008.

57 Wang Q, Gao J, Xing JL, *et al*. DCFNet: Discriminant correlation filters network for visual tracking. arXiv: 1704.04057, 2017.

58 Huang C, Lucey S, Ramanan D. Learning policies for adaptive tracking with deep feature cascades. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 105–114.

59 Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

60 He AF, Luo C, Tian XM, *et al*. A twofold siamese network for real-time object tracking. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. 2018. 4834–4843.

61 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

62 Wang Q, Teng Z, Xing JL, *et al*. Learning attentions: Residual attentional siamese network for high performance online visual tracking. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 4854–4863.