深度学习在视频目标跟踪中的应用进展与展望

管 皓 [1] 薛向阳 [1] 安志勇 [1]

摘 要 视频目标跟踪是计算机视觉的重要研究课题, 在视频监控、机器人、人机交互等方面具有广泛应用. 大数据时代的到来及深度学习方法的出现, 为视频目标跟踪的研究提供了新的契机. 本文首先阐述了视频目标跟踪的基本研究框架. 对新时期视频目标跟踪研究的特点与趋势进行了分析, 介绍了国际上新兴的数据平台、评测方法. 重点介绍了目前发展迅猛的深度学习方法, 包括堆叠自编码器、卷积神经网络等在视频目标跟踪中的最新具体应用情况并进行了深入分析与总结. 最后对深度学习方法在视频目标跟踪中的未来应用与发展方向进行了展望.

关键词 目标跟踪, 视频分析, 在线学习, 深度学习, 大数据

引用格式 管皓, 薛向阳, 安志勇. 深度学习在视频目标跟踪中的应用进展与展望. 自动化学报, 2016, **42**(6): 834−847

DOI 10.16383/j.aas.2016.c150705

Advances on Application of Deep Learning for Video Object Tracking

GUAN Hao[1] XUE Xiang-Yang[1] AN Zhi-Yong[1]

Abstract Video object tracking is an important research topic of computer vision with numerous applications including surveillance, robotics, human-computer interface, etc. The coming of big data era and the rise of deep learning methods have offered new opportunities for the research of tracking. Firstly, we present the general framework for video object tracking research. Then, we introduce new arisen datasets and evaluation methodology. We highlight the application of the rapid-developing deep-learning methods including stacked autoencoder and convolutional neural network on video object tracking. Finally, we have a discussion and provide insights for future.

Key words Object tracking, video analysis, online learning, deep learning, big data

Citation Guan Hao, Xue Xiang-Yang, An Zhi-Yong. Advances on application of deep learning for video object tracking. *Acta Automatica Sinica*, 2016, **42**(6): 834−847

视频目标跟踪是计算机视觉领域的重要研究课题, 其主要任务是获取视频序列中感兴趣的目标的位置与运动信息, 为进一步的语义层分析 (动作识别、场景识别等) 提供基础. 其定义是: 给定视频序列初始帧中目标的位置框 (一般为矩形框), 在接下来的视频序列中自动给出该目标的位置框或者在目标离开视域时给出提示. 视频目标跟踪研究在智能视频监控、人机交互、机器人等领域有广泛应用, 具有很强的实用价值. 视频目标跟踪同视频目标检测、视频分类 (识别) 一样, 都是视频内容分析的重要方面. 在一个实用的计算机视觉系统中, 跟踪的初始状态由检测结果所提供, 同时其所给出的运动信息为语义层的分类 (识别) 等任务所使用. 因此, 视频目标跟踪是处于视频内容分析研究的中间层次模块.

视频目标跟踪研究有较多分支, 内容十分丰富. 按照跟踪目标是否已知, 可分为特定目标跟踪与非特定目标跟踪. 特定目标的跟踪可以利用先验知识对目标外观进行建模, 典型代表有手的跟踪、人眼跟踪、头或脸部跟踪等, 其中手的跟踪在人机交互方面有重要应用, 是未来非接触式交互工具的基础. 非特定目标跟踪对目标无任何先验知识, 只能利用第一帧所给出的标注信息, 因其较高的难度一直以来都是跟踪研究的重点. 按照跟踪目标的数量, 可分为单目标跟踪和多目标跟踪. 单目标跟踪是最早、最基础也是目前研究最多的分支. 多目标跟踪研究随着近年来数据关联等方法的出现也日益增多并发展较快. 按照获取目标数据的摄像头的特点, 可以分为单摄像头跟踪、多摄像头跟踪和跨摄像头跟踪 (也称为重识别). 单摄像头跟踪最为基础, 其特点是无法获取目标的深度信息. 多摄像头跟踪可以捕获目标多个视角的图像, 从而获取深度信息, 但图像融合难度较大. 跨摄像头跟踪是近年来跟踪领域里面新兴的研究课题, 旨在弥补目前固定摄像头的视域局限, 在目前的安防领域中具有重要的实用价值.

收稿日期 2015-10-26 录用日期 2016-05-03
Manuscript received October 26, 2015; accepted May 3, 2016
国家自然科学基金 (61572138), 上海市科技创新行动计划项目 (15511104402) 资助
Supported by National Natural Science Foundation of China (61572138) and Science and Technology Commission of Shanghai Municipality (15511104402)
本文责任编委 柯登峰
Recommended by Associate Editor KE Deng-Feng
1. 复旦大学计算机科学技术学院上海市智能信息处理重点实验室 上海 201203
1. Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 201203

此外, 还有刚体跟踪与非刚体跟踪、离线跟踪与在线跟踪、RGBD 跟踪、红外小目标跟踪等研究分支. 限于篇幅, 本文不再一一列举. 本文主要以单摄像头下的单目标跟踪进行说明, 该部分研究的历史较长, 成果最为丰富, 是目前视频目标跟踪的主流内容, 最能体现跟踪的本质特点, 而其他分支的内容则多与图形学、图像识别以及具体领域知识等有所交叉融合.

将视频中目标的运动信息进行提取一直以来都是多媒体内容分析研究中的重要方面, 因此视频目标跟踪是一个研究历史并不短的课题. 许多经典的视频目标跟踪算法如均值漂移 (Mean shift) 已经作为标准模块集成到影响较大的计算机视觉开发库如 OpenCV 等当中. 虽然其发展一直较为缓慢, 但是随着目前大数据时代的到来, 在新时期下视频目标跟踪研究取得了突飞猛进式的发展并呈现出许多新的特点. 这主要得益于机器学习理论和技术的发展以及较大规模跟踪数据集和评测平台的建设. 尤其值得重视的是, 目前机器学习的前沿领域, 在多媒体识别领域中取得了巨大成功的深度学习方法也开始在视频目标跟踪研究中得以应用并取得了良好效果. 本文在介绍视频目标跟踪研究的基本框架及自身特点的基础上, 重点介绍深度学习方法在视频目标跟踪研究中的最新应用情况. 通过结合视频目标跟踪自身的特点, 对具体应用深度学习时存在的困难与挑战进行了分析和探讨. 最后对其未来发展进行分析和展望.

1　视频目标跟踪系统框架及关键技术

一般性视频目标跟踪系统的运行流程及框架如图 1 所示.

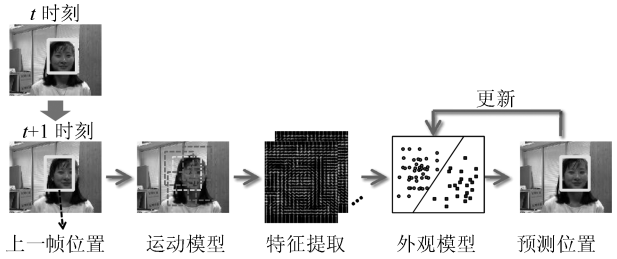


图 1　视频目标跟踪系统框架

Fig. 1　The framework of video object tracking

从整体上分为输入视频、运动模型、特征提取、外观模型、位置确定、模型更新等几个步骤. 初始化由视频序列中的第一帧给定, 一般由一个矩形框来标定待跟踪的目标. 运动模型利用视频序列的时空关联性, 在目标潜在空间范围内进行搜索或采样, 为后面的特征提取、外观模型提供样本. 特征提取是对目标外观进行有效编码, 从二维图像空间映射到某一特征空间, 从而为后面不同外观模型的处理提供基础. 外观模型旨在对目标外观进行有效建模与描述, 从而将目标以最大的区分度被跟踪系统搜索到. 具体跟踪时, 通过计算候选样本的相似度、可信度, 得分最高的样本被确定为最终的预测结果.

目标在新一帧视频中的位置最终确定以后, 一般要利用新得到的数据对目标的外观模型进行更新操作, 这样做的目的是适应目标在线运动过程中外观的变化.

1.1　运动模型

在视频序列中对目标的位置进行预测时, 会在上一帧跟踪框的基础上, 在原目标位置周围产生一定数量的候选位置. 跟踪算法就是要在这些候选位置中寻找出一个最优解. 运动模型在此过程中起到核心作用, 即按照一定规则产生候选位置样本. 连续两帧之间目标的位置不会相距过远, 运动模型就是依据这个基本约束来以较高效率提供候选, 这是与基于全图像扫描的目标检测的根本不同之处. 目前运动模型主要分为三种:

1) 均值漂移 (Mean shift)

均值漂移, 是一种基于核密度估计的非参数估计方法. 文献 [1] 中首先将均值漂移算法应用于跟踪问题, 此后成为经典跟踪方法. 在跟踪时, 需要设定一个目标函数来计算目标与候选窗口的核密度, 而后利用 Bhattacharyya 准则作为匹配条件, 通过移动均值向量来不断优化目标函数从而完成目标搜索. 由于通过梯度优化来完成搜索, 因此基于均值漂移的跟踪算法运行速度快、实时性高.

2) 滑动窗口 (Slide window)

在目标周边正方形或者圆形范围内进行穷举搜索的采样策略, 也称为密集采样. 这种方式将搜索范围内所有可能的潜在位置都予以考虑, 但是要付出较大的计算代价.

3) 粒子滤波 (Particle filter)

粒子滤波在经典的卡尔曼滤波的基础上发展而来[2], 先验概率密度用加权粒采样样本 (粒子) 来近似表示. 每个粒子的权值表示了该样本的重要程度. 每次跟踪结果确定后, 会根据不同粒子的重要程度进行重采样. 粒子滤波方法具有较高的计算效率, 同时可以融入仿射变换信息, 因此目前在一些较好的跟踪算法中应用较多.

1.2　特征提取

特征是对目标的抽象化表示, 即从目标原始空间映射到某一特征空间. 特征提取过程就是将原始图像数据通过转换得到更有利于描述需求的表达方式. 在多媒体内容分析的各个领域, 特征表达与提取都是最重要的内容之一. 对于视频目标跟踪而言, 好

的特征应当具备两个基本性质: 1) 具有较强的区分度; 2) 要具有较高的计算效率, 以满足跟踪的实时性要求.

目前跟踪算法采用的特征分为人工特征和学习特征两类. 人工特征可以分为外观特征和运动特征. 外观特征是从目标的物理直观出发, 通过结合数学工具设计出来的特征. 运动特征是针对视频的特点, 从视频帧之间的时间关联性出发设计的特征, 这些特征是静态图像中所没有的. 由机器自动学习到的特征为学习特征. 这些特征通过机器学习的方式自动提取, 无需事先知道目标的物理性质, 从而可以大大提高特征提取的效率. 目前以深度学习为代表的特征学习方法已经成为计算机领域的前沿和热点.

1.2.1 人工特征

人工特征包含外观特征和运动特征. 目前跟踪算法广泛采用的外观特征总体上可以分为四类: 灰度特征、颜色特征、梯度特征和纹理特征.

灰度特征是最为简单和直观的特征表达方式, 计算效率高, 可以分为原始灰度特征、灰度直方图特征、区域灰度变化特征 (Haar 特征) 三种表征形式. 原始灰度特征就是将输入视频图像转换为灰度图, 而后将标准化处理后的灰度图作为模板来表示目标. 这种方式较简单, 运算速度快. 灰度直方图通过统计手段来反映目标图像整体或局部的灰度分布特征. Haar 特征是一种反映目标图像中区域灰度变化的特征表示手段, 于文献 [3] 中首次提出并成功应用于人脸检测. Haar 特征由于计算效率高, 同时对于边缘、水平、垂直敏感等优点被广泛应用到目标检测与跟踪当中.

颜色特征主要分为两种: 一种以颜色直方图来表征[4−5]; 另一种则是近年来兴起的具有更好表征能力的 Color name 特征[6]. 颜色特征对姿态、尺度等不敏感, 用于非刚体跟踪时具有一定优势. 但其受光照影响较大, 同时易受颜色相近背景的干扰.

纹理特征通过外观表面的微观变化来描述目标, 是对目标外观细节、规则程度的量化. 目前跟踪算法中常用的纹理特征是局部二值模式 (Local binary pattern, LBP)[7]. 纹理特征可以较好地描述目标外观的细节, 但是对于纹理细节少、小尺度、远距离或者背景纹理复杂的目标描述能力较差, 此时跟踪效果往往不理想.

梯度特征通过统计目标图像局部的梯度分布来表征外观. 文献 [8−9] 中采用在图像中广泛采用的 SIFT (Scale invariant feature transform) 特征及其加速版本 SURF (Speeded up robust features) 特征来表征跟踪目标, 但实时性较差. 一种更为广泛应用的梯度特征是 HOG (Histogram of oriented gradient) 特征, 它于文献 [10] 中首次被提出并成功用于行人检测. HOG 特征的思想是利用分块单元对梯度进行统计, 能够非常好地反映局部像素之间的关联. 梯度特征对光照变化等具有不变性, 性能稳定. 其主要不足是无法描述外观精确尺寸、角度、姿态等信息.

运动特征旨在挖掘视频帧之间的时空关联性, 因此有效提取运动特征, 可在外观特征的基础上增添辅助信息, 有利于提高跟踪性能. 目前跟踪中最重要的运动特征提取方法是光流法. 光流 (Optical flow) 是对局部图像运动的一种近似表达, 主要通过计算给定视频中局部图像的时间与空间导数, 近似得出二维运动场. 两种经典的光流算法是 LK 算法[11] 和 HS 算法[12]. 前者更具运算效率优势, 在跟踪中应用更多.

光流法效率较高, 能够应对摄像头与目标相对运动的情况, 但其计算存在一些光强、位移的限定条件. 目前对复杂场景的跟踪较少单独使用光流特征, 而是同其他外观特征结合在一起, 最典型的例子是 TLD (Tracking learning detection) 算法[13−14].

1.2.2 学习特征

研究者们在一直努力, 试图让机器能够自动学习到特征. 主成分分析法 (Principle component analysis, PCA) 可以视为最早的自动特征提取方法.

近三年来, 深度学习 (Deep learning) 在图像分类、目标检测等领域取得了突出成绩, 成为目前最强有力的自动特征提取方法. 深度神经网络通过多层级的学习和映射, 可以从边缘、颜色等底层特征逐步得到高层的抽象特征. 这些抽象特征维数高、区分力强, 利用简单的分类器即可实现高准确率的分类、回归等任务. 目前已经有一些基于学习特征的跟踪方法被提出, 利用离线训练好的深度卷积网络, 在跟踪时通过截取目标在网络不同卷积层的特征来辅助实现目标定位.

关于深度学习在视频跟踪方向的研究进展是本文的核心内容, 将在稍后部分进行详细介绍.

1.3 外观模型

外观模型是视频目标跟踪研究中的重要内容[15−16]. 好的外观模型能较大提升跟踪性能. 近年来, 外观模型得到了极大发展, 这主要得益于图像处理、机器学习、目标检测等相关领域所取得的丰硕成果. 目前跟踪算法的外观模型分为两类: 产生式模型和判别式模型.

1.3.1 产生式模型

产生式模型是一种自顶向下的处理方法[17]. 首先建立目标的外观数据先验分布, 而后在候选区域中搜索与先验模型最为匹配、重构误差最小的区域

作为下一帧中目标的位置, 如图 2 所示. 产生式模型总体上分为三类: 基于模板的模型[18−20]、基于子空间的模型[21−22] 和基于稀疏表示的模型[23−28]. 近年来, 稀疏表示逐渐成为多媒体领域的热门研究课题[27], 它通过基函数字典表示的稀疏向量来建立目标的外观模型. 文献 [28] 首次将稀疏表示方法引入到视频目标跟踪领域中, 其核心思想是将跟踪转化为求解 L_1 范数最小化问题.

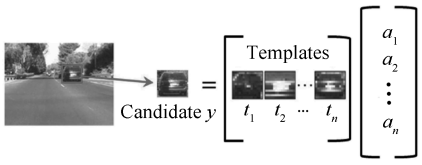


图 2　产生式外观模型

Fig. 2　The generative appearance model

产生式模型着眼于对目标外观数据内在分布的刻画, 具有很强的表征能力. 其最大不足是没有利用背景信息, 在遇到遮挡等情况时容易通过错误更新将噪声混入模型中从而最终导致误差和漂移.

1.3.2　判别式模型 (基于目标检测的模型)

判别式模型也称为基于检测的模型 (Tracking by detection), 是近年来逐渐兴起并逐渐占据主流的方法. 其直接借鉴了机器学习理论及其在目标检测中的成功应用. 与产生式模型不同, 判别式模型并不对目标外观分布做事先的刻画, 而是将跟踪问题等同于一个分类问题, 利用一个在线分类器 (目标检测器) 将跟踪目标与背景分离, 如图 3 所示.

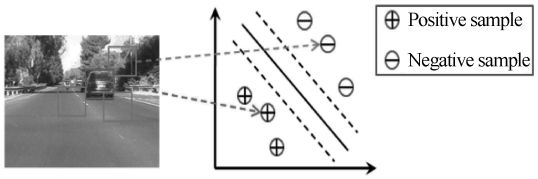


图 3　判别式外观模型

Fig. 3　The discriminative appearance model

判别式模型充分利用了前景与背景信息, 可以将两者更好的区分, 因而具有较强的鲁棒性, 这是较之于产生式模型的优势所在. 但在利用样本进行在线学习与更新的过程中, 也容易因样本的标注错误影响分类器的性能, 造成误分类. 尽管如此, 各种改进与优化措施的出现, 使得基于判别式模型的跟踪器显示出越来越强的优势.

判别式模型有基于支持向量机的模型[29−31]、基于 Boosting 的模型[32−34]、基于多示例学习的模型[35]、基于岭回归的模型[36−37]、基于随机蕨的模型[13]、基于朴素贝叶斯的模型[38] 等.

1.4　更新

相比于离线训练模型 (目标及视频都是已知的), 在线跟踪的优势在于可以实时地获取目标外观变化并做出在线调整, 体现出更大的灵活性与适应性. 在线跟踪的这种优势主要体现于在线更新环节. 利用第一帧给出的标注信息以及随后各帧的跟踪结果, 在线外观模型可以增量式更新. 对于产生式模型, 主要是对模板或基函数的更新; 对于判别式模型, 主要利用新采样的样本来对分类器进行增量式在线训练, 通过不断融入的正负样本, 使分类器能够不断适应目标与背景的变化. 目前更新策略研究相对较少, 主要的更新策略有:

1) 每一帧都进行更新. 该方式较简单, 目前应用较多. 但由于太过频繁, 增加了漂移的可能性.

2) 每隔一定的帧数才更新一次.

3) 当响应分数 (匹配或分类得分) 低于一定阈值时才更新. 低于阈值往往说明目标外观已发生较大变化. 在该策略中, 增加了对外观变化程度的判断, 减少了更新频率, 因而比策略 1) 效果好一些.

4) 分别计算正负样本的响应分数, 当两者的差值低于一定阈值时更新. 该方式在判别式模型中采用. 由于考虑了前景与背景的差异度量, 可使跟踪器具有更好的鉴别能力.

2　跟踪算法评测平台与方法的新发展

2.1　跟踪算法的最新评测数据平台

一个有代表性的数据集对于跟踪算法性能进行全面而公正的评测是至关重要的. 随着大数据时代的到来, 对训练与测试数据集的重视与日俱增[39], 如图像识别领域的 ImageNet[40]、目标检测领域的 Pascal VOC[41]、视频检索领域的 TRECVID[42] 等. 具体到视频目标跟踪研究领域, 权威的数据集与测试平台的建立也是大势所趋.

该方面突破性的工作是文献 [43] 中所提出 VTB 数据集, 它是目前最具影响力的视频目标跟踪算法测试数据集. 起初包含 50 个测试视频, 随后扩展到 100 个[44]. 该数据集的建立具有里程碑式的意义, 结束了跟踪算法在零散视频集上测试的局面, 使众多跟踪算法第一次有了真正意义上统一的测试平台.

另一个有影响力的视频目标跟踪数据集是 VOT 数据集[45]. 该平台效法著名的目标检测数据平台 Pascal VOC, 从 2013 年开始每年进行一次跟踪算法的竞赛并作排名. 其规模与 VTB 相当, 但在算法的性能评测指标上有一些不同.

以上两个是目前最具影响力的视频跟踪数据集.

2.2 跟踪算法的评测准则与方法

对目标跟踪算法有三个要求: 准确性、鲁棒性、高效性. 目前很少算法能同时在这三点上表现优异.

准确性 (Accuracy): 有三个指标可反映跟踪准确性. 如果一个跟踪器能尽量降低这三种误差, 则其准确性较高. 这三个指标分别是: 1) 偏移 (Deviation): 预测位置同实际位置的距离; 2) 误检 (False positive): 将非目标物体视为物体; 3) 漏检 (False negative): 没有正确识别出目标.

鲁棒性 (Robustness): 如果一个跟踪器在一个视频序列中取得高精度, 但在另一些视频中表现差, 则其不够鲁棒. 一个有较高鲁棒性的跟踪器应能在大多数的测试视频序列中表现出较高性能, 即能应对复杂多样的场景.

高效性 (Efficiency): 视频目标跟踪是一个对实时性要求极高的研究领域, 这是与检测、识别的重要不同点. 一个真正实用的跟踪器必须实时运行.

对应于跟踪器的总体性能要求, 很多测量准则与方法被提出. 下面进行详细介绍.

1) 中心误差 (Center location error): 每一帧中跟踪器输出的矩形框中心与实际中心位置的欧氏距离. 加和后取平均值为平均中心误差. 中心误差越小, 说明跟踪效果越好.

2) 重叠率 (Overlap rate): 设 S_T 是跟踪器输出的跟踪框区域, S_G 为实际目标区域, 则重叠率的定义为两者的交集与并集的比值, 即: $R = \frac{area(S_T \cap S_G)}{area(S_T \cup S_G)}$, 重叠率越高, 说明跟踪效果越好.

3) 成功率 (Success rate): 对于每一帧而言, 若中心误差小于一定阈值或重叠率大于一定阈值则认为该帧跟踪成功. 跟踪成功的帧数同视频序列总帧数的比值称为成功率.

4) 精度图 (Precision plot) 与成功图 (Success plot): 将 3) 中所设置的阈值在一定范围内变动时, 会得到一系列的成功率数值所构成的曲线图, 当对应于中心误差时构成的曲线称为精度图; 对应于重叠率时称为成功图.

5) 时间鲁棒性度量 (Temporal robustness evaluation, TRE) 和空间鲁棒性度量 (Spatial robustness evaluation, SRE). 这两个指标是在文献 [43] 中为衡量跟踪器鲁棒性而提出的. TRE 跟踪器用测试视频序列中的随机的一帧进行初始化而不是第一帧, 作出其相应的成功图, 以此来衡量跟踪器在时间轴上的鲁棒性. SRE 跟踪器用第一帧初始化, 但对初始跟踪框位置进行了一定的平移、缩放等微小扰动, 做出相应的成功图, 以此来测试跟踪器能否在随后帧中稳定跟踪住目标.

6) FPS (Frames per second): 每秒处理的帧数, 是一个用来衡量跟踪算法处理效率和速度的常用指标.

3 深度学习方法在跟踪中的应用

3.1 深度学习概述

深度学习 (Deep learning) 是近年来机器学习领域的一个新的研究方向. 由于其在语音、文本、图像、视频等诸多方面相较于传统方法所取得的巨大进展和突破, 使得其成为目前计算机科学中最引人注目的研究课题, 在某种程度上可以说是引领了一场大数据时代下的科技革命.

深度学习的产生和崛起并非一日之功, 而是有着深厚的历史积淀. 直观上, 它是神经网络在大数据时代新的发展, 然而从“浅”走到“深”却经历了很长的曲折与积累. 20 世纪 80 年代, Rumelhart、Hinton 和 Williams 三位科学家完整而系统的提出了基于反向传播算法 (Back propagation, BP) 的神经网络[46]. 此成果掀起了神经网络研究的巨大浪潮. 但 BP 神经网络只能含有较“浅”的层次结构, 原因是随着层数的增加网络很容易陷入局部最小和出现过拟合现象. 随着 20 世纪 90 年代以支持向量机 (Support vector machine, SVM) 为代表的更优秀的“浅”层模型的出现, 神经网络的研究相对沉寂. 此局面在 2006 年被 Hinton 及其学生发表在著名的《科学》上的研究成果所打破[47]. 该文提出了深度网络与深度学习的概念, 拉开了深度学习的序幕. 深度学习首先在语音识别领域取得突破[48], 在图像识别领域取得的突破性成果[49], 其作者用深层卷积神经网络在大规模图像识别问题上取得了巨大成功. 随后在目标检测任务中也超越了传统方法[50−51], 继而在视频分类方面也取得突破[52−53].

深度学习之所以在其产生和发展过程中不断取得惊人的成功, 根本原因在于其强大的特征表达能力. 如图 4 所示, 在多媒体识别领域, 一个最为基本和核心的问题就是如何对多媒体信息 (图像、语音等) 进行有效表达. 一个强有力的特征表达, 对于多媒体内容识别和分析的效果是事半功倍的.

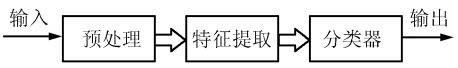


图 4 多媒体内容识别的框架

Fig. 4 The framework of recognition in multimedia

传统的特征表达是通过人们手工设计的特征来实现的, 比如上文所提到的 HOG 特征、LBP 特征等, 这样做的缺点是费时费力, 需要根据具体问题和任务的不同而重新设计. 而深度学习则可以自动学习到反映目标的良好特征, 完全不需要人的参与. 同时, 神经学的研究表明人对信息的处理是分级

的[54−55], 而深度学习的分层架构在某种程度上正是对人脑机制的模拟. 相比于浅层模型, 深度学习方法对于如图像这种高度非结构化、分布复杂的数据的刻画能力和泛化性能要强大很多.

特别需要指出的是, 深度模型的成功有赖于两个重要基础条件, 一个是容量巨大的训练和测试数据集, 它们为深度模型的训练提供了数据保障; 另一个是通用计算芯片 GPU 的发展, 它为深度模型的训练提供了硬件支持. GPU 原本用于计算机图形显示, 后来在大规模并行计算中的优势使其成为深度学习的计算硬件基础. 目前主流的深度学习研究开发平台如 Caffe[56]、Theano[57] 都已将对 GPU 的支持作为必备功能.

3.2　深度学习基本模型

深度学习按照学习方法可以分为无监督学习模型和有监督学习模型. 无监督深度学习模型主要包括基于受限玻尔兹曼机的深度置信网络 (Deep belief net, DBN)[58] 和基于自动编码器的深度网络 (Stacked autoencoder)[59] 两大类. 监督学习深度模型包括多层感知机 (Multilayer perceptron) 和深度卷积神经网络 (Convolutional neural network, CNN)[60].

按照深度网络中的组成单元之间是否存在闭环, 可将深度学习模型分为前馈型深度网络 (Feed-forward neural network, FNN) 和递归型深度网络 (Recurrent neural network, RNN)[61], 如图 5 所示. 值得一提的是递归型深度网络是较其他类型深度网络更加特殊的类型, 它将着眼点放在“时间”的深度建模上. 尤其是目前递归型神经网络的主要代表之一 — 长短时记忆网络 (Long and short term memory, LSTM)[62−64], 能够对数据相对较长的时间跨度内的状态进行记忆和学习, 因此在序列问题的处理, 如语音识别、自然语言处理、手写体识别等方面表现优异, 成为又一引人注目的深度模型.

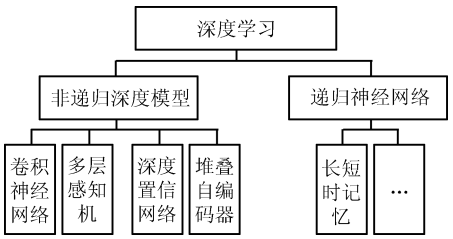


图 5　深度学习的基本模型

Fig. 5　The basic models of deep learning

3.3　深度学习方法在跟踪中的应用概述

深度学习是一种强大的特征学习方法. 本节对深度学习在视频目标跟踪领域中的应用做一个整体性的介绍与分析. 尽管在多媒体领域诸多方面取得了巨大成功, 但在视频目标跟踪这一特殊领域, 深度学习的应用却受到一定限制, 成果数量较视频识别、视频目标检测要少很多. 主要原因是:

1) 视频目标跟踪中, 严格意义上讲仅有第一帧的数据是真正的标注数据, 在其后的在线跟踪过程中, 正负样本的量级仅有几百个. 所以, 视频目标跟踪是典型的小样本在线学习问题, 这使得以处理大数据见长的深度学习方法难以发挥优势.

2) 视频目标跟踪对实时性要求极高. 而规模庞大的深度网络很难达到实时性要求. 这就需要在网络规模和运行速度方面做综合考虑.

尽管存在以上困难, 由于深度学习在特征提取、外观建模上的优势, 研究者们仍然通过不同手段, 结合视频目标跟踪任务的特点, 设计出一些基于深度学习的跟踪算法. 从目前的研究成果来看, 研究者们在将深度学习应用于目标跟踪的过程中主要遵循两种思路:

1) 利用深度神经网络所学习到的特征的可迁移性, 首先在大规模的图像或视频数据集上离线训练某一特定类型的深度神经网络. 而后在具体的在线跟踪时, 利用之前基本训练好的网络对目标进行特征提取, 并利用在线获取的数据对该深度网络进行微调节, 以适应在线时目标外观的具体变化.

2) 将深度神经网络的结构做一定的改变, 使其能够适应在线跟踪的要求. 主要的方法包括将网络的层数维持在一个兼顾性能与效率的数量水平、将网络中费时的训练过程做适度简化等. 目前该方面的工作还处于起步阶段, 探索空间较大.

3.4　堆叠自编码器在跟踪中的应用

3.4.1　自编码器基本原理

堆叠自编码器是典型的非监督深度学习网络, 它的基本构成单元是自编码器 (Autoencoder). 自编码器的示意图如图 6 所示. 其基本过程是将输入信号进行编码, 而后利用解码器在编码后的信号的基础上对原始信号进行重构, 目标函数是使重建信号与原始信号的重构误差最小. 自编码器的思想是通过对原始信号进行编码的方式将其以更为简洁的形式加以表达, 从而去除冗余, 反映信号更加本质的属性.

将自编码器逐层叠加就构成了堆叠自编码器 (Stacked autoencoder) 这一深度学习网络模型. 在堆叠自编码器中, 下一层的输出作为上一层的输入, 每一层进行单独优化. 这样通过每一层编码器的映射, 逐步得到反映原始信号更本质属性的高层特征. 为了利用数据中的标注信息, 还可以使用监督学习的方法对网络参数进行微调, 此时需在顶层增加一

个逻辑斯谛回归 (Logistic regression) 层.

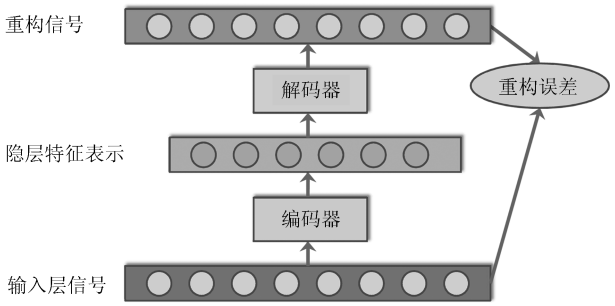


图 6 自编码器示意图

Fig. 6 The illustration of autoencoder

自编码器的一个重要改进是去噪自编码器[59]. 其提出的目的是使深度网络对于噪声更加鲁棒. 去噪自编码器的原理示意如图 7 所示. 它的核心思想是在原始信号上施加一定噪声后作为训练数据对深度网络进行训练. 将重构信号与原始未加噪声的信号作对比作为重构误差. 通过最小化重构误差, 使得去噪自编码器可以适应一定程度的噪声干扰, 从而增强了网络的鲁棒性.

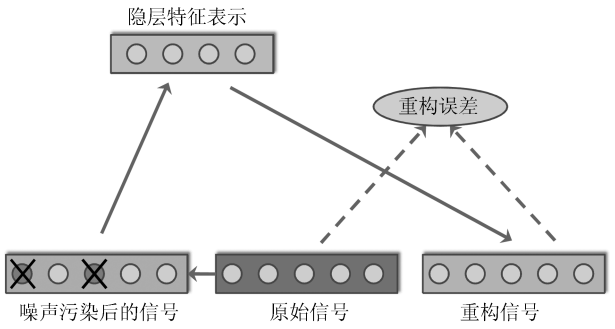


图 7 去噪自编码器示意图

Fig. 7 The illustration of denoise autoencoder

3.4.2 自编码器在跟踪中的应用

由于堆叠自编码器, 尤其是去噪堆叠自编码器的特征学习能力和抗噪声性能, 它被首先应用到非特定目标的在线视频目标跟踪当中. 该方面的经典工作来自于文献 [65]. 该文作者首先在大规模的小尺度图像样本数据集[66] 上对一个堆叠去噪自编码器进行离线训练. 其深度网络的结构如图 8 中左图所示. 而后将训练好的网络用于跟踪时对目标外观的特征提取. 为了利用在线标注信息, 在网络的顶端加入逻辑斯谛回归二值分类器, "1" 指示目标, "0" 指示背景, 如图 8 中右图所示. 初始化时, 利用第一帧给出的标注信息, 对网络进行微调整. 在线跟踪时, 继续通过实时采集的正负样本对深度网络进行微调 (更新), 以达到适应目标外观变化的目的.

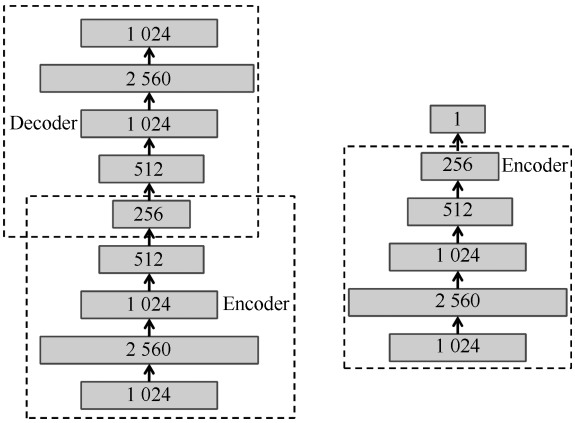


图 8 用于跟踪的去噪自编码器架构[65]

Fig. 8 Denoise autoencoder for video tracking[65]

为减少计算量, 系统更新并非每一帧都进行, 而是每隔一定帧数或系统置信度小于一定阈值时才更新一次. 整个跟踪系统的运动模型基于粒子滤波框架. 实验结果表明其跟踪效果好于部分基于传统特征表示的方法. 该文工作首次将深度网络用于非特定目标在线跟踪问题, 是典型的 "离线训练 + 在线微调" 架构下的深度学习跟踪方法, 框架具有示范性. 其网络结构相对简单, 训练容易. 其不足主要是: 1) 对网络进行离线的预训练所使用的图像数据都是较低分辨率下的小图像, 虽然网络可以学习到一些一般性的图像特征, 但对于跟踪任务而言, 核心要求是对跟踪目标特征的有效描述而非对整个图像的描述. 这些基于低分辨率图像重构意义下所学习到的特征能否最大化区分目标与背景并没有理论上保证. 2) 其网络后端为一个二值分类器, 即只将跟踪视为二值分类问题. 在线样本标注时, 将与当前目标较近的样本标为正样本, 较远的标为负样本. 由于 "近" 和 "远" 都需要设定具体的阈值, 因此非常容易引入误样本从而使网络得到错误的训练信息. 3) 实时性低, 特别是目标遭遇背景中较强干扰时, 网络频繁的更新操作使得运行效率很低.

文献 [67] 在文献 [65] 的基础上, 将深度网络同在线 AdaBoost 框架进行融合, 将 4 个基于堆叠自编码器的跟踪器组成集成系统, 将置信度最大的候选区域作为最终预测的目标位置. 而后根据跟踪结果在线调节每个自编码器网络的权重从而达到增强鲁棒性的目的. 该方法通过几个网络的融合互补, 一定程度上弥补了单个网络跟踪时易受干扰而漂移的问题, 但代价是使得计算负担进一步加重.

文献 [68] 同样采用了先离线训练深度堆叠自动编码器, 而后在线微调的策略. 与文献 [65] 不同的是, 文献 [68] 中的工作强调了深度网络对于时间关联性图像的学习. 在离线训练阶段并未利用离散的静态图像作为训练样本, 而是采用带标注的视频序

列图像来训练深度网络. 在网络训练算法上, 除了增加重构误差最小的约束项外, 还增加了基于独立子空间分析 (Independent subspace analysis, ISA) 的相邻帧之间的时间连续性约束 (Temporal slowness constraint). 通过这样的策略, 使得训练出的网络在进行在线跟踪时可以更好地提取运动不变性特征. 实验结果表明其效果要好于文献 [65] 中的方法.

文献 [69] 将深度自编码器网络用于跟踪含有运动模糊的视频目标. 快速运动和运动模糊是视频目标跟踪中的一大类困难因素. 该文通过高斯函数对模糊图像建模与深度网络进行特征提取相结合, 在一定程度上克服了模糊帧对跟踪器的影响.

文献 [70] 的重点放在解决深度学习用于跟踪时的实时性问题. 文章作者的出发点有两个: 1) 视频跟踪中的目标都是较小尺度的图像, 因此没有必要用过多层数的深度网络, 这样会加大在线计算负担, 作者认为用较少层数的深度网络足可以充分表达目标特征. 2) 作者认为由于只有视频第一帧是真正的标注数据, 而在线运行时的标注数据都或多或少存在不准确性, 因此在对离线训练好的深度网络进行在线微调时, 第一帧与后继帧采用不同的训练策略, 即在后继帧中更新微调时, 采用较少的训练周期和较大的学习率, 这样可以进一步加快网络的运行速度.

总体而言, 作为优秀的非监督深度学习模型, 堆叠自编码器理论直观而优美, 体量适中, 因此在视频跟踪中最先得到应用并取得了优良效果.

3.5　卷积神经网络在跟踪中的应用

3.5.1　卷积神经网络基本原理

与堆叠自编码器不同, 深度卷积神经网络 (Convolutional neural network, CNN) 是一种监督型的前馈神经网络. 鉴于其出色的效果, 卷积神经网络成为目前图像与视频识别领域的研究热点.

卷积神经网络的生理学理论基础来自 20 世纪 60 年代科学家 Hubel 和 Wiesel 通过对猫视觉皮层的研究成果. 他们提出了感受野 (Receptive field) 的概念[71]. 基于此发现, 文献 [72] 中提出的神经认知机 (Neocognitron) 首次将感受野概念应用于人工神经网络, 该模型可视为卷积神经网络的初级版本. 随后 LeCun 等设计出基于 BP 算法的卷积神经网络[60, 73], 该网络集成了局部感受野、权值共享、降采样三大特性, 在计算机视觉的许多方面都获得了很好的效果[74]. 在大数据时代, 随着大规模带标注的图像数据平台 ImageNet 等的出现以及计算硬件水平的发展, 卷积神经网络在模式识别, 特别是计算机视觉任务中体现出强大性能. 革命性的标志是文献 [49] 中, Krizhevsky 等利用深层卷积神经网络大幅度提高了图像识别成功率. 此后在目标检测、视频分类等任务中都取得了超越传统方法的成果.

卷积神经网络的基本结构如图 9 所示, 总体上分为特征提取部分、全连接部分和输出部分. 特征提取部分是卷积神经网络的核心, 由卷积、非线性变换和降采样三种操作的周期性交替进行而组成. 卷积操作就是通过卷积核来获取特征图 (图 9 中的 C1, C2 层), 卷积核需要通过训练优化得到. 非线性变换就是将卷积阶段得到的特征按照一定的原则进行筛选, 提高模型的特征表达能力. 降采样操作采用池化 (Pooling, 通常的做法是取一定邻域内像素的平均值或最大值) 得到分辨率降低的图像, 目的是获取一定的位移不变性, 提高图像识别的鲁棒性. 经过特征提取层后, 得到的多个特征图构成特征向量后通过全连接层与最终的输出层相连.

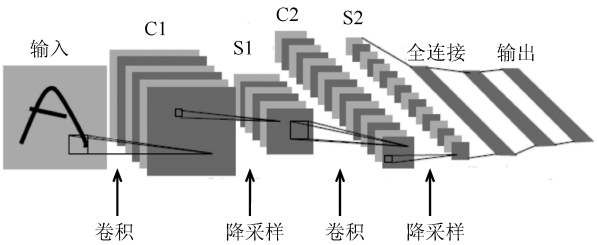


图 9　卷积神经网络的基本架构示意[60, 73]

Fig. 9　The illustration of convolutional neural network[60, 73]

卷积神经网络通过误差反向传播算法进行有监督的学习和训练. 随着当前一些技术实力强大的科技公司的推动, 卷积神经网络的层数在不断加深, 规模越来越庞大[75−76], 但需耗费大量的训练时间.

3.5.2　卷积神经网络在跟踪中的应用

目前卷积神经网络在跟踪中的应用, 主要研究思路有两种: 一种是先离线训练好所采用的网络, 而后在线运行时微调; 另一种则是设计简化版的卷积神经网络, 力图摆脱离线训练而能够完全在线运行.

文献 [77] 中采用两卷积层和两降采样层的卷积神经网络进行特征提取. 网络后端接径向基神经网络来实现分类. 该工作的主要不足是在线跟踪时没有采用实时更新的策略, 因此对目标外观变化的适应性不强.

文献 [78] 中首先在辅助数据集上离线训练一个两层级的卷积神经网络, 而后将其应用于在线跟踪当中. 为使网络学习到能够应对复杂运动的特征, 作者提出在视频图像而非离散图像上进行离线训练. 在线跟踪时, 利用在线采集的样本对网络进行微调、更新. 该工作的主要创新在于注重了网络对于运动不变性特征的学习, 因而对于跟踪而言更具启发意义.

文献 [79] 中作者设计了一个含有 7 个卷积层和

2 个全连接层的深度卷积神经网络. 与大部分用于跟踪的卷积神经网络不同, 作者所设计的网络并不是二值化输出 (1 代表目标, 0 代表背景), 而是结构化输出. 通过一张响应图来指示目标潜在区域的可能性. 首先在 ImageNet 上离线训练网络, 而后通过迁移学习将其用于在线特征提取. 通过两个卷积神经网络的相互融合互补来实现稳定的跟踪. 该工作的主要创新在于对深度网络用于跟踪时的输出端进行了关注.

文献 [80] 中利用离线训练好的深度卷积神经网络在线提取目标的显著性图, 跟踪系统通过存储若干帧跟踪目标的显著性特征图, 在线维护一个外观模型模板, 通过相关匹配来实现定位目标. 该文的研究着眼点较为新颖, 没有直接利用深度卷积网络给出跟踪结果, 而是先通过其得到目标的显著性特征图再进行操作, 这在很大程度上避免了网络误分类造成跟踪漂移的问题.

文献 [81−82] 都借鉴了卷积网络的最新发展, 将更深层数、特征学习能力更强的卷积网络引入到视频目标跟踪中. 与之前的工作相比, 两者都注重了对不同层级特征的充分利用, 在对跟踪中应用深度网络的理解上更进了一步.

以上工作都是首先离线训练卷积神经网络而后以在线数据对网络进行微调和更新. 除了这种思路外, 还有少数工作试图通过以完全在线的方式来利用卷积神经网络进行目标跟踪.

文献 [83−84] 中提出了一种在线卷积神经网络架构, 其特点在于完全不依赖离线学习而只进行在线学习. 其在采样、训练、更新等几个方面都做了一定改进, 主要考虑在线运行效率问题. 其采用含有两个卷积层和两个降采样层的卷积神经网络. 为获取尽可能多的在线样本, 增加了一个预处理环节, 得到若干不同参数的局部正则化图像及梯度图像作为多通道输入. 跟踪系统维持一个记忆池, 在线存储跟踪到的目标样本作为网络训练和更新之用.

文献 [85] 中对卷积神经网络做了较大的简化, 没有通过监督训练的方式获取卷积核, 而是通过预先设计的滤波器作为卷积核来获取层级特征. 这些方式往往需要在特征表达能力与运行速度之间做权衡以便设计简化版网络.

3.6 对比分析

3.6.1 卷积神经网络与传统方法的对比分析

能够将目标与其周边背景有效区分的特征向量对视频目标跟踪的最终效果起到关键作用. 传统方法的局限首先在于往往只着眼于目标某一方面物理特性的刻画, 而忽视了其他特性. 例如 Haar 特征在对人脸进行跟踪时的效果较好, 但应用于行人跟踪时效果则不够理想. 这就使得这些方法的应用范围受到很大限制, 在含有各种干扰因素的最新跟踪数据平台上很难获得全面优异的表现. 而深度学习方法在辅助训练数据的支撑下可以获取普适性更高的特征[86].

其次, 传统方法如 HOG 特征几乎都只着眼于底层特征, 而卷积神经网络可以通过层级映射提取从边缘、纹理等底层特征到高层抽象语义特征等一系列不同层次的特征表示. 与图像分类等任务仅利用最后的语义性特征不同, 卷积神经网络所提取的不同层级的特征都可以为跟踪任务所采用, 这等同于为目标的位置分析提供了更多的视窗, 这一点是传统方法无法比拟的.

当然, 传统方法的主要优势在于运行速度和对辅助数据的较少依赖, 在目前而言更具工程实用价值, 随着硬件加速技术的进步, 相信这种差距会逐步缩小. 同时非深度学习跟踪方法中的优秀思想也值得借鉴[87], 如文献 [29] 中提出的 Struck 算法所采用的结构化学习与输出思想, 体现出对目标跟踪问题更深刻的理解, 对于深度学习跟踪方法而言非常值得借鉴.

3.6.2 卷积神经网络与堆叠编码器的对比分析

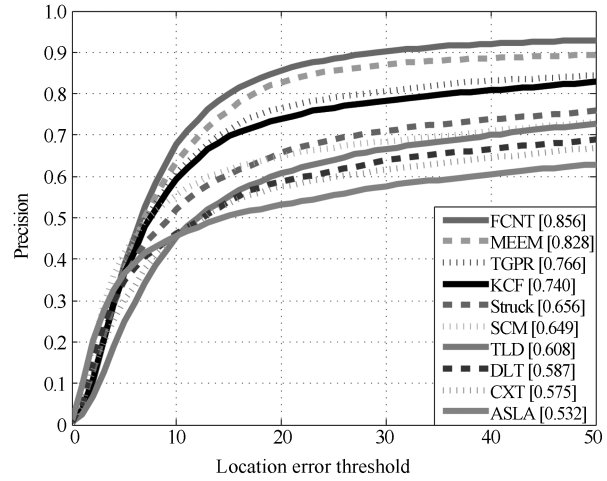
通过对目前的研究成果的对比分析, 基于卷积网络的跟踪架构比基于堆叠自编码器的方法具有更大的优势和更广阔的发展空间. 首先, 卷积网络的结构决定了其具有处理图像数据的先天优势, 这是目前其他深度学习架构所不及的. 同时, 卷积网络的架构具有很强的可拓展性, 可以达到非常“深”的层数. 相比深度卷积网络而言, 目前堆叠自编码器的中间层数就少很多. 卷积网络的这种优势使得其具有更强大的特征学习能力, 可以为跟踪任务提供更多的特征分析视窗.

图 10 所示是一份基于卷积网络最新成果的跟踪方法与优秀的传统方法及基于堆叠自编码器方法的实验对比结果图 (引自文献 [81]). 其中, 图 10 (a) 为精度图, 图 10 (b) 为成功图, 其物理意义分别是在不同的中心误差阈值和重叠度阈值下, 成功跟踪到的帧数的百分比 (具体定义详见第 2.2 节). 其中每个图中的说明框是各个算法的性能排名, 越靠上的算法性能越好. 从结果可以看出, 基于卷积网络的跟踪方法优于目前性能较好的基于传统方法的跟踪器, 同时对比于基于堆叠自编码器的跟踪方法[65] 也表现出明显优势.

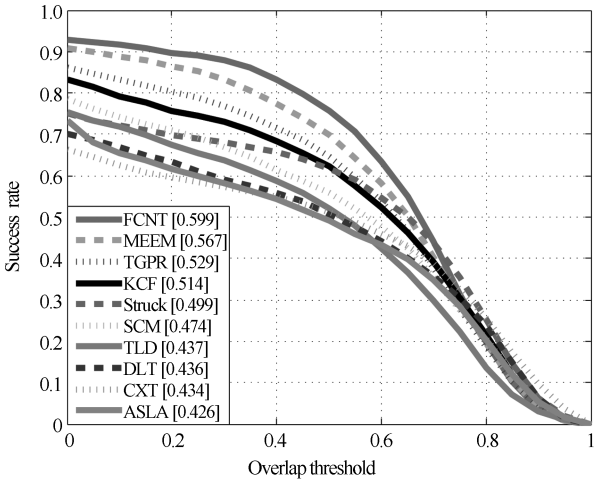
3.7 应用总结与困难分析

上面的一些工作尽管取得了一些成果, 但是深度学习在视频目标跟踪中的应用仍然较少, 尽管部分算法跟踪效果很好, 但总体而言, 此方面仍有很大

的探索空间. 目前的问题和困难主要有:



(a) 精度图
(a) Precision plot



(b) 成功图
(b) Success plot

图 10　基于卷积网络的跟踪算法与其他方法的对比实验[81] (FCNT 为基于卷积网络的跟踪器, DLT 为基于堆叠自编码器的跟踪器[65].)

Fig. 10　Comparison of CNN-based tracking method and other trackers[81] (FCNT is a CNN-based tracker and DLT is an autoencoder-based tracker[65].)

1) 通过预训练深度网络的方式需要耗费大量的时间, 且此种方式更加适合于特定目标的跟踪, 如行人跟踪等. 当应用场合是非特定目标的跟踪时, 一个重要问题是选取什么样的辅助训练集能够获取更稳定的跟踪效果. 有些研究者认为应选取如 ImageNet 这样包含物体类别丰富的海量图像训练集, 这样可以获取更一般的图像特征, 另一些工作则更倾向于视频数据集, 认为可以获取更好的时间特征表达能力. 目前针对辅助训练数据集的选取并没有明确的理论指导, 也没有工作进行此方面的实验来验证, 总体上训练数据集的选取有着较大的随意性.

2) 卷积神经网络的传统架构在图像识别、检测等领域取得了巨大成功, 但并不适用于跟踪. 这主要是因为其中的降采样、池化等操作会降低图像的分辨率. 这些操作的目的是获取图像位移不变性从而降低因物体形变等因素对于识别的影响. 然而降低分辨率后会损失空间位置信息, 而这些信息对于视频目标跟踪来说是至关重要的. 因此简单套用卷积神经网络未必会取得非常好的效果, 必须对网络结构进行一定的改进, 不能够在特征提取过程损失空间信息.

3) 目前深度学习在视频目标跟踪中的应用中, 大都以二值分类器作为最终的输出, 即在线跟踪过程中所采样的样本都是以 0 和 1 作为样本, 这种在线标注方式显得过“硬”, 非常容易引入误标签, 从而引起深度网络的误分类, 最终导致误差积累直致漂移. 此时单纯使用深度学习方法并不能解决跟踪漂移问题, 需要同其他方法相结合才能更好地发挥深度网络的作用.

4) 深度网络用于视频目标跟踪的实时性问题是其应用的一大挑战. 由于深度学习算法及架构固有的性质, 其实时性往往很难达到实用要求. 一些工作对深度网络作了过大的简化, 以牺牲特征表达能力来加速系统, 似乎并不可取. 如何做真正合理的简化和改进, 使得深度学习方法真正适用于实时应用, 是值得深入研究的课题.

5) 深度网络的重要形式 — 递归神经网络目前在视频目标跟踪中还没有应用. 递归神经网络, 尤其是其重要变体 — 长短时记忆网络在序列识别问题上已取得了较大的成功. 由于具有对序列的记忆能力, 这种网络是一种时间轴上的深度学习方法, 也是对人类智能的一种重要的模拟形式. 具体到视频目标跟踪领域, 由于当前数据集中各种干扰因素的存在, 如摄像机晃动等, 使得跟踪视频序列往往成为很不规则的序列信号, 这与语音信号等不同. 因此目前对于非特定目标、非特定环境的视频目标跟踪问题, 应用递归神经网络还非常困难, 仅有一些研究工作试图从其他方面进行模拟[88].

4　发展趋势展望

作为多媒体内容分析的重要子领域, 视频目标跟踪是一个复杂且困难的研究课题, 因为在现实环境中有太多因素对跟踪过程进行干扰. 经过数十年的努力, 虽然对一些简单场景已经能够很好处理, 但面对更多更复杂环境时跟踪效果仍不够理想. 深度学习方法的出现, 为构建更加鲁棒的目标外观模型提供了可能. 但为了设计出高精度、高鲁棒性和实时性的跟踪算法, 仍然需要开展大量研究工作, 目前的

研究重点和发展趋势主要集中于以下几点:

1) 深度学习与在线学习的融合. 视频目标跟踪本质上是一个在线学习问题, 最显著的特点是在线数据集是在不断扩充的. 深度学习应用中所采用的先逐层训练而后全局微调的训练方式在纯粹的在线环境是否真正适用, 如何避免陷入局部极小值, 都是值得深入研究的问题.

2) 构建适合视频目标跟踪的深度网络. 需要在目标表征能力和实时性之间有所权衡, 既要保持深度学习特征学习的优势, 同时也要兼顾跟踪的高实时性要求. 同时, 如卷积神经网络中的降采样等损失空间信息的操作都是应用于跟踪任务的障碍, 因此要进行必要改进, 才能使深度网络真正适用于跟踪问题.

3) 跟踪数据平台的创建. 目前建立大型的训练与测试数据平台并举行定期的比赛, 已经成为图像与视频研究的流行趋势. 因此如何根据视频目标跟踪研究的特点, 建立起大规模、具有代表性、测试方法严谨、适合深度网络训练、测试的跟踪视频数据平台, 仍然是一个值得研究的课题.

4) 递归神经网络的应用. 尽管应用于一般性目标及开放环境的视频目标跟踪问题困难较大, 但作为对于时间序列建模的重要深度模型, 递归神经网络仍然可以在跟踪中有所作为. 可以预见, 在特定目标、固定镜头等限定情况下, 应用递归神经网络可以帮助跟踪系统更好地进行轨迹预测, 从记忆角度来防止漂移发生. 这方面有很大探索空间.

5 结束语

本文在对视频目标跟踪的研究框架进行说明的基础上, 首先介绍了跟踪算法评测数据平台与方法的最新发展. 而后作为核心, 本文重点介绍了目前在多媒体领域发展迅猛的深度学习方法在视频目标跟踪领域的应用情况. 在已有工作的基础上, 对深度学习方法应用于跟踪时的特点、问题及难点进行了深入分析和总结. 文章最后对未来深度学习方法在跟踪中的进一步应用进行了展望, 相信对相关领域的研究人员会有较好的参考价值.

References

1 Comaniciu D, Ramesh V, Meer P. Real-time tracking of non-rigid objects using mean shift. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island, SC: IEEE, 2000. 142−149

2 Risfic B, Arulampalam S, Gordon N. Beyond the Kalman filter-book review. *IEEE Aerospace and Electronic Systems Magazine*, 2004, **19**(7): 37−38

3 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2001. I-511−I-518

4 Pérez P, Hue C, Vermaak J, Gangnet M. Color-based probabilistic tracking. In: Proceedings of the 7th European Conference on Computer Vision. Copenhagen, Denmark: Springer, 2002. 661−675

5 Possegger H, Mauthner T, Bischof H. In defense of color-based model-free tracking. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 2113−2120

6 Danelljan M, Khan F S, Felsberg M, van de Weijer J. Adaptive color attributes for real-time visual tracking. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 1090−1097

7 Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of the 12th IAPR International Conference on Pattern Processing. Jerusalem: IEEE, 1994. 582−585

8 Zhou H Y, Yuan Y, Shi C M. Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding*, 2009, **113**(3): 345−352

9 Miao Q, Wang G J, Shi C B, Lin X G, Ruan Z W. A new framework for on-line object tracking based on SURF. *Pattern Recognition*, 2011, **32**(13): 1564−1571

10 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

11 Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981. 674−679

12 Horn B K P, Schunck B G. Determining optical flow. *Artificial Intelligence*, 1981, **17**(2): 185−203

13 Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(7): 1409−1422

14 Kalal Z, Mikolajczyk K, Matas J. Forward-backward error: automatic detection of tracking failures. In: Proceedings of the 20th IEEE International Conference on Pattern Recognition. Istanbul: IEEE, 2010. 2756−2759

15 Li X, Hu W M, Shen C H, Zhang Z F, Dick A, van den Hengel A. A survey of appearance models in visual object tracking. *ACM Transactions on Intelligent Systems and Technology*, 2013, **4**(4): Article No. 58

16 Zhang Huan-Long, Hu Shi-Qiang, Yang Guo-Sheng. Video object tracking based on appearance models learning. *Journal of Computer Research and Development*, 2015, **52**(1): 177−190
(张焕龙, 胡士强, 杨国胜. 基于外观模型学习的视频目标跟踪方法综述. 计算机研究与发展, 2015, **52**(1): 177−190)

17 Hou Zhi-Qiang, Han Chong-Zhao. A survey of visual tracking. *Acta Automatica Sinica*, 2006, **32**(4): 603−617
(侯志强, 韩崇昭. 视觉跟踪技术综述. 自动化学报, 2006, **32**(4): 603−617)

18 Adam A, Rivlin E, Shimshoni I. Robust fragments-based tracking using the integral histogram. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, NY, USA: IEEE, 2006. 798−805

19 Alt N, Hinterstoisser S, Navab N. Rapid selection of reliable templates for visual tracking. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1355−1362

20 He S F, Yang Q X, Lau R W H, Wang J, Yang M H. Visual tracking via locality sensitive histograms. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 2427−2434

21 Black M J, Jepson A D. EigenTracking: robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 1998, **26**(1): 63−84

22 Ross D A, Lim J, Lin R S, Yang M H. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 2008, **77**(1−3): 125−141

23 Zhang T Z, Liu S, Xu C S, Yan S C, Ghanem B, Ahuja N, Yang M H. Structural sparse tracking. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 150−158

24 Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 1822−1829

25 Zhang T Z, Ghanem B, Liu S, Ahuja N. Robust visual tracking via multi-task sparse learning. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2042−2049

26 Zhang S P, Yao H X, Sun X, Lu X S. Sparse coding based visual tracking: review and experimental comparison. *Pattern Recognition*, 2013, **46**(7): 1772−1788

27 Wright J, Ma Y, Mairal J, Sapiro G, Huang T S, Yan S C. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 2010, **98**(6): 1031−1044

28 Mei X, Ling H B. Robust visual tracking using L_1 minimization. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto: IEEE, 2009. 1436−1443

29 Hare S, Saffari A, Torr P H S. Struck: structured output tracking with kernels. In: Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona: IEEE, 2011. 263−270

30 Avidan S. Support vector tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(8): 1064−1072

31 Bai Y C, Tang M. Robust tracking via weakly supervised ranking SVM. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 1854−1861

32 Grabner H, Grabner M, Bischof H. Real-time tracking via on-line boosting. In: Proceedings of the British Machine Vision Conference. Edinburgh, UK: BMVA Press, 2006. 47−56

33 Grabner H, Leistner C, Bischof H. Semi-supervised on-line boosting for robust tracking. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 234−247

34 Stalder S, Grabner H, van Gool L. Beyond semi-supervised tracking: tracking should be as simple as detection, but not simpler than recognition. In: Proceedings of the 12th IEEE International Conference on Computer Vision Workshops. Kyoto: IEEE, 2009. 1409−1416

35 Babenko B, Yang M H, Belongie S. Visual tracking with online multiple instance learning. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 983−990

36 Henriques J F, Caseiro R, Martins P, Batista J. Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 702−715

37 Henriques J F, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(3): 583−596

38 Zhang K H, Zhang L, Yang M H. Real-time compressive tracking. In: Proceedings of 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 864−877

39 Huang Kai-Qi, Ren Wei-Qiang, Tan Tie-Niu. A review on image object classification and detection. *Chinese Journal of Computers*, 2014, **37**(6): 1225−1240
(黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述. 计算机学报, 2014, **37**(6): 1225−1240)

40 Deng J, Dong W, Socher R, Li J J, Li K, Li F F. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 248−255

41 Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303−338

42 Smeaton A F, Over P, Kraaij W. Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval. Santa Barbara, CA, USA: ACM, 2006. 321−330

43 Wu Y, Lim J, Yang M H. Online object tracking: a benchmark. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 2411−2418

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

45 Kristan M, Matas J, Leonardis A, Felsberg M, Cehovin L, Fernández G, Vojír T, Häger G, Nebehay G, Pflugfelder R. The visual object tracking VOT2015 challenge results. In: Proceedings of the 2015 IEEE International Conference on Computer Vision Workshops. Santiago: IEEE, 2015. 564−586

46 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

47 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

48 Hinton G, Deng L, Yu D, Dahl G E, Mohamed A R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, **29**(6): 82−97

49 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceeding of Advances in Neural Information Processing Systems. Nevada, USA: MIT Press, 2012. 1097−1105

50 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 580−587

51 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Proceeding of Advances in Neural Information Processing Systems. Montréal, Canada: MIT Press, 2015. 91−99

52 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F F. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 1725−1732

53 Ji S W, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 221−231

54 Lee T S, Mumford D, Romero R, Lamme V A F. The role of the primary visual cortex in higher level vision. *Vision Research*, 1998, **38**(15−16): 2429−2454

55 Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A: Optics Image Science and Vision*, 2003, **20**(7): 1434−1448

56 Jia Y Q, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, FL, USA: ACM, 2014. 675−678

57 Bergstra J, Bastien F, Breuleux O, Lamblin P, Pascanu R, Delalleau O, Desjardins G, Warde-Farley D, Goodfellow I J, Bergeron A, Bengio Y. Theano: deep learning on GPUS with python. In: Advances in Neural Information Processing Systems Workshops. Granada, Spain: MIT Press, 2011. 1−4

58 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

59 Vincent P, Larochelle H, Bengio Y, Manzagol P A. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 1096−1103

60 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324

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

62 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

63 Gers F A, Schraudolph N N, Schmidhuber J. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research*, 2003, **3**: 115−143

64 Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(5): 855−868

65 Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking. In: Proceeding of Advances in Neural Information Processing Systems. Nevada, USA: MIT Press, 2013. 809−817

66 Torralba A, Fergus R, Freeman W T. 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(11): 1958−1970

67 Zhou X Z, Xie L, Zhang P, Zhang Y N. An ensemble of deep neural networks for object tracking. In: Proceedings of the 2014 IEEE International Conference on Image Processing. Paris, France: IEEE, 2014. 843−847

68 Kuen J, Lim K M, Lee C P. Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle. *Pattern Recognition*, 2015, **48**(10): 2964−2982

69 Ding J W, Huang Y Z, Liu W, Huang K Q. Severely blurred object tracking by learning deep image representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, **26**(2): 319−331

70 Dai L, Zhu Y S, Luo G B, He C. A low-complexity visual tracking approach with single hidden layer neural networks. In: Proceedings of the 13th IEEE International Conference on Control Automation Robotics and Vision. Singapore: IEEE, 2014. 810−814

71 Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat′s visual cortex. *Journal of Physiology*, 1962, **160**(1): 106−154

72 Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 1980, **36**(4): 193−202

73 LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, **1**(4): 541−551

74 LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. Paris, France: IEEE, 2010. 253−256

75 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 1−9

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

77 Jin J, Dundar A, Bates J, Farabet C, Culurciello E. Tracking with deep neural networks. In: Proceedings of the 47th Annual Conference on Information Sciences and Systems (CISS). Baltimore, MD, USA: IEEE, 2013. 1−5

78 Wang L, Liu T, Wang G, Chan K L, Yang Q X. Video tracking using learned hierarchical features. *IEEE Transactions on Image Processing*, 2015, **24**(4): 1424−1435

79 Wang N Y, Li S Y, Gupta A, Yeung D Y. Transferring rich feature hierarchies for robust visual tracking. arXiv: 1501.04587, 2015.

80 Hong S, You T, Kwak S, Han B. Online tracking by learning discriminative saliency map with convolutional neural network. In: Proceedings of the 32th International Conference on Machine Learning. Lille, France: JMLR, 2015. 597−606

81 Wang L J, Ouyang W L, Wang X G, Lu H C. Visual tracking with fully convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3119−3127

82 Ma C, Huang J B, Yang X K, Yang M H. Hierarchical convolutional features for visual tracking. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3074−3082

83 Li H X, Li Y, Porikli F. DeepTrack: learning discriminative feature representations online for robust visual tracking. *IEEE Transactions on Image Processing*, 2016, **25**(4): 1834−1848

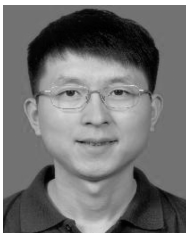
84 Li H X, Li Y, Porikli F. Robust online visual tracking with a single convolutional neural network. In: Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer, 2015. 194−209

85 He Y, Dong Z, Yang M, Chen L, Pei M T, Jia Y D. Visual tracking using multi-stage random simple features. In: Proceedings of the 22nd International Conference on Pattern Recognition. Stockholm: IEEE, 2014. 4104−4109

86 Danelljan M, Häger G, Khan F S, Felsberg M. Convolutional features for correlation filter based visual tracking. In: Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop. Santiago: IEEE, 2015. 621−629

87 Wang N Y, Shi J P, Yeung D Y, Jia J Y. Understanding and diagnosing visual tracking systems. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3101−3109

88 Hong Z B, Chen Z, Wang C H, Mei X, Prokhorov D, Tao D C. MUlti-Store tracker (MUSTer): a cognitive psychology inspired approach to object tracking. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 749−758



管 皓 复旦大学计算机科学技术学院博士研究生. 主要研究方向为多媒体内容分析, 深度学习. 本文通信作者.
E-mail: guanh13@fudan.edu.cn
(**GUAN Hao** Ph. D. candidate at the School of Computer Science, Fudan University. His research interest covers video analysis and deep learning. Corresponding author of this paper.)

薛向阳 复旦大学计算机科学技术学院教授. 主要研究方向为视频大数据分析, 计算机视觉, 深度学习.
E-mail: xyxue@fudan.edu.cn
(**XUE Xiang-Yang** Professor at the School of Computer Science, Fudan University. His research interest covers big video data analysis, computer vision, and deep learning.)

安志勇 复旦大学计算机科学技术学院博士后. 2008 年获得西安电子科技大学博士学位. 主要研究方向为图像与视频内容分析、检索.
E-mail: azytyut@163.com
(**AN Zhi-Yong** Postdoctor at the School of Computer Science, Fudan University. He received his Ph. D. degree from Xidian University in 2008. His research interest covers image and video content analysis and retrieval.)