小视频内容分析技术发展探讨

Short Video Content Analysis Technology

薛向阳 /XUE Xiangyang，李斌 /LI Bin
（复旦大学，中国 上海 200433）
(Fudan University, Shanghai 200433, China)

DOI：10.12142/ZTETJ.202101012
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20210118.1639.002.html
网络出版日期：2021-01-18
收稿日期：2020-12-15

摘要：小视频数量呈爆炸式增长态势，并引发了许多技术需求，包括小视频的编辑、搜索、推荐、溯源、审查和监管等。介绍了小视频数据的主要特点和小视频内容分析技术面临的挑战，并对目标检测、追踪、重识别等小视频内容分析技术的研究进展做了综合分析。认为只有构建一个整合多种不同算法的系统，才能够更准确、更鲁棒地解决分析问题，才能系统性地完成小视频内容分析任务。

关键词：小视频；内容分析技术；视频目标检测；多目标追踪；人物重识别

Abstract: The number of short videos has increased explosively, which has led to more technical requirements, such as editing, searching, recommendation, sourcing, censoring, and monitoring of short videos. The main features of short video data and the challenges faced by the short video content analysis technology are introduced. The research progress of short video content analysis technologies such as object detection, tracking, and re-identification is comprehensively analyzed. It is considered that only by building a system that integrates multiple algorithms, can the analysis problems be solved more accurately and robustly, and the short video content analysis task can be completed systematically.

Keywords: short video; content analysis technology; video object detection; multi-object tracking; person re-identification

1 小视频数据类型与特点

1.1 小视频数据类型

随着抖音、快手、腾讯微视、西瓜视频等小视频应用平台的兴起，小视频已经随处可见。在激烈的竞争下，市场上涌现出了不同类别的小视频内容。

（1）社交生活类

以快手、抖音、腾讯微视等为代表的应用平台，鼓励用户拍摄、制作、上传小视频，分享自己的生活点滴，这方便了用户拓宽自己的社交范围。此类小视频主题多为生活记录，如拍摄宠物、烹饪、服饰等。通过分享生活点滴，用户可以找到与自己趣味相投的朋友，拓宽社交圈。

（2）内容服务类

以西瓜视频、梨视频为代表的应用平台，依靠大数据分析为用户提供精准内容服务，如感兴趣的话题、认识的朋友、关心的产品等。此类小视频主题多为行业热点资讯、育儿经验或家教信息、“双十一”优惠活动等。

（3）剪辑技术类

以小咖秀等为代表的应用平台，为对视频制作感兴趣的用户提供制片剪辑等功能，使用户以更灵活幽默的方式记录自己的生活。此类小视频主题多为宣传视频、纪念视频、情景短剧以及其他具有特殊意义的视频（如高考加油视频）等。

1.2 小视频数据特点

小视频数据除了具有规模海量这一特点之外，其余还包括类型繁多、特效复杂、姿态多变等。

（1）类型繁多

类型繁多是小视频数据的一大特点。小视频数据包含的物体类别为开集，除人物之外，还涵盖宠物、电子产品、音乐器材、体育用品等。此外，与图像数据集（ImageNet）[1]的1 000类和目标检测数据集（COCO）[2]的80类相比，小视频数据的类别更丰富，包含更多的子类，如不同品种的猫和狗、不同品牌的电子产品等。

（2）特效复杂

与其他视频相比，小视频往往包含更多的特效，以使自身更具有吸引力和娱乐性，如各种幻灯片转场、人物美颜特效、多屏镜面特效等。这对目标检测和追踪等分析任务而言，是一个不可忽视的巨大挑战。

（3）姿态多变

在小视频中，各目标的外观姿态往往变化较大。小视频记录生活点滴，包含大量特写镜头。一段小视频主题可能聚焦于人、动物、产品等。这些目标围绕的主题包含较多姿态和外观变化，例如人的换装小视频、宠物成长记录小视频等。

除前文提到的3种特点外，由于小视频的拍摄设备多为智能手机，故小视频数据的特点还包括画面清晰度相对较低、镜头抖动、视野较窄等。

2 小视频分析技术面临的挑战

学术界对视频内容分析技术已进行大量且系统的深入研究。例如，针对视频盗用转载和重复出现问题的视频拷贝检测技术，对视频进行分割以提取感兴趣或关键场景的镜头分割技术，对视频中主要物体进行检测、分类和追踪的语义提取技术等。其中，小视频语义提取是最受关注的技术，是后续各种应用的基础。

在对小视频中的主要物体进行语义抽取时，涉及的技术模块主要包括视频目标检测、多目标追踪、人物重识别（也称Person ReID）等。视频目标检测是指，从视频图像帧中自动定位事先定义好的类别集合中的物体，并推断其类别。多目标追踪是指，利用目标的外观特征和位置信息来将相邻帧中的相同目标关联起来，以构成目标序列，实现对目标的持续追踪。人物重识别是指，在多个非重叠摄像头拍摄的场景下，在一段视频或者某个图片集合中筛检出感兴趣的人物。当然，重识别技术也可以用于筛检某一动物、某一物品等。

2.1 小视频目标检测

目前，人们对视频目标检测的研究主要集中在类似ImageNet VID[3]（VID指视频目标检测）的数据集上。这些数据集往往包含相对较少的物体类别，背景相对简单，前景物体容易与背景区分。小视频场景下的目标检测任务面临的巨大挑战具体包括：（1）类别繁多。小视频中出现的物体类别数以万计，且物体类别的分布呈现长尾效应。大量物体类别严重缺乏训练数据，极大地影响了目标检测算法的性能。（2）剪辑与特效带来较大干扰。镜头切换和视频特效使得物体外观信息被严重干扰，前后帧中主要物体的外观连续性被严重破坏。（3）背景复杂、物体运动难预测。小视频来自用户上传，其背景和人物姿态变化往往更复杂。

2.2 小视频多目标追踪

考虑到业界的实际需求，传统的多目标追踪任务主要聚焦于交通监控等应用场景中对行人和车辆的追踪。这导致目前学术界广泛研究的数据集更多是通过监控设备来采集的，并且主要针对行人目标进行追踪。目前，多目标追踪算法解决的焦点主要是监控场景中的常见问题，如行人目标密集、遮挡等。

在小视频场景中，多目标追踪任务面临前所未有的挑战。与交通监控场景相比，小视频创作偏爱近景。人物在视频上占据区域较大，很难被简单地视为刚体。人物姿态变化直接影响追踪效果。除此以外，频繁的镜头切换也打破了物体帧间位置连续性的假设。

因此，小视频目标追踪任务面临的挑战可归纳为：（1）镜头切换。这使得时空连续性只能在局部窗口内有效。（2）场景不确定性。目标的距离、大小难以预测，很难依据先验信息进行算法性能优化。（3）制作特效问题。小视频有电脑特效或叠加字幕，这给目标追踪带来很多干扰。

2.3 目标重识别

通用目标的重识别是一个十分困难的研究课题，主要是因为每类目标的特征各不相同。在对小视频分析时，我们通常从人物等特定类别目标重识别开始研究，而这面临的挑战包括：每个镜头中人物的入镜区域存在很大不同，上一个镜头出现的是一个完整的人物，下一个镜头中可能只有上半身入镜；人物在小视频画面中的复杂运动姿态与传统监控画面中的行走姿态有很大差别。这些挑战使得小视频场景下的目标重识别与相机固定监控场景下的行人重识别有很大的不同。

针对小视频场景的人物重识别任务主要包括两点：（1）视频内人物重识别。根据某段小视频前几帧出现的主要人物目标，将后续帧出现的相同人物目标与之一一匹配起来。这类任务的挑战主要是人物局部入镜、姿态变化大、遮挡情况复杂多样（如障碍物遮挡、人物相互遮挡、随机字幕遮挡）。（2）视频间的人物重识别。根据（1）中得到的某个人物图片序列，搜寻其他小视频中出现的相同着装的该人物。这类任务的挑战主要是解决人物着装变化、背景风格差异大、面部遮挡模糊等问题。

2.4 算法性能需求

（1）计算速度

对于现有海量规模的小视频数据，如果算法处理不够快，对用户请求的响应不及时，用户的使用体验将极大降低。以小视频搜索为例，如果搜索算法能为用户即时提供新的热点视频，用户体验无疑将会得到提升。

（2）算法精度

由于小视频包含的物体种类繁多，且姿态外观等变化较大，如果分

析算法的精度不够高，用户体验将受到显著影响。这对小视频内容分析算法提出了很高的要求，即必须在面临各种挑战的情况下保持稳定且很高的精度，才可获得良好的应用效果。

（3）泛化能力

小视频类别很多，其包含的物体类别也是开放的，这对分析技术的泛化能力提出更高要求。小视频分析算法只有具备了良好的泛化能力，才能很好地适应各种应用场景，从而才能真正满足用户时刻变化的应用需求。

3 小视频分析技术研究进展

本章分别从小视频分析任务涉及的技术研究进展，和针对第 2 章所述的小视频数据特殊难点的解决方案出发，对相关方法进行详细介绍。

3.1 视频目标检测

目标检测从计算机视觉兴起时便一直是基础性的研究任务。随着 2015 年面向视频目标检测任务的数据集 ImageNet VID 的发布，深度学习在目标检测研究中开始发挥巨大作用。当前学术界主流研究思路有：

（1）将检测与追踪相结合

基于检测与追踪结合的方法在图像级别的目标检测结果的基础上，辅以目标追踪方法来将各帧中相同物体的检测框关联起来。2017 年由 KANG K. 等提出具有卷积神经网络的小管（T-CNN）[4] 的方法，通过图像目标检测器对输入视频完成目标检测，再通过目标追踪算法得到目标的检测框序列。2019 年由 LUO H. 等提出的分布式对象技术（DoT）[5] 框架则进一步地对视频目标检测任务进行有选择性地检测和追踪，充分利用检测算法和追踪算法各自的优点，在速度和质量上取得平衡。

（2）利用光流信息

光流可描述物体的运动状态和轨迹。2015 年和 2017 年 P. FISCHER 等分别提出了光流网络（FlowNet）[6] 和 FlowNet 2.0[7]，通过卷积神经网络直接计算出光流，用来代替目标追踪模块。ZHU X. 等在 2017 年提出的流引导特整体聚合（FGFA）[8] 算法，利用光流描述的运动轨迹将相邻帧的特征聚合到当前帧的特征上，可得到更鲁棒的物体特征，能明显减少由于视频中物体运动模糊和亮度变化带来的影响。光流适用于对局部时空域内的物体运动进行建模，但难以对全局时空域内的物体特征进行整合。

（3）利用循环神经网络

视频是一种典型的序列数据，用循环神经网络来对帧序列和物体的运动进行建模是一种常见的选择。2017 年，LU Y. 等提出关联长短期记忆（LSTM）[9] 结构，对视频目标检测任务中的相邻帧间物体的关联信息进行专门建模。通过与检测网络相结合，该方法可直接回归获得物体的位置和类别，同时还能将物体在不同帧之间的特征在时空上都关联起来，最终可得到融合了时序运动信息的关联特征。然而，这类方法的缺点是大量增加了模型训练难度和计算耗时。

（4）利用全局帧特征融合

WU H. P. 等不仅考虑到从局部时域中提取物体的运动信息，还更加关注物体在全局时域上的时序信息，并在 2019 年提出了序列级语义聚合（SELSA）[10] 算法。该算法在整个视频的完整序列内提取各帧所有感兴趣区域的特征，通过一个聚类模块和变换模块将不同帧之间具有相似语义信息的候选框匹配，从而得到一个全局时域内综合的特征，随后与各帧中提取得到的局部特征相聚合，可得到一个更鲁棒的特征。CHEN Y. H. 等在 2020 年提出基于记忆增强的全局 – 局部整合（MEGA）[11] 算法，同时利用局部时域和全局时域内物体的时序信息，即在局部更加关注物体的运动信息，在全局更加关注物体的外观信息，并将两者结合得到最终的融合特征。

3.2 视频目标追踪

目前，视频多目标追踪主要分为 3 个模块：目标检测、特征提取 / 运动预测、亲和力计算与关联。

（1）目标检测模块

目标检测模块负责提供目标位置信息，并将其作为后续处理的先验信息。检测模块提供位置信息，用于确定目标的外观特征，为运动预测提供目标初始位置信息。针对目标检测的研究已经取得长足进步：从传统的可变形部件模型（DPM）[12] 到深度学习方法，从视觉几何网络（VGGNet）[13] 到最新的高分辨网络（HRNet）[14]，ImageNet 数据集的精度不断被刷新，位置预测方式从一阶段的快速区域卷积神经网络（Faster R-CNN）[15] 到两阶段的 YOLOv4（指对象检测算法）[16]，在精度和速度上都取得了巨大突破。

（2）特征提取 / 运动预测

特征提取 / 运动预测模块主要负责从外观特征提取高层语义特征和充分利用运动信息。多目标跟踪算法 DeepSort[17] 利用简单残差网络构成的重识别（ReID）模型，大幅度改善 Sort[18] 算法的性能。而 HRNet 等方法则采用姿态评估模型来挖掘目标姿态等更为丰富的信息。在运动预测方法中，目前采用比较多的是简单高效的卡尔曼滤波算法。卡尔曼滤波算法可预测接近匀速直线的运动，也有些方法采用更为复杂的粒子滤波，以拟合目标的复杂运动。

（3）亲和力计算与关联

亲和力计算模块从物体区域的特征信息中计算出匹配对，即当前检测

区域与预测结果区域之间的相似度，以此作为依据来进行关联计算。关联模块从相似度矩阵中求解出最佳的匹配方式，尽量将同一目标的检测区域匹配到对应的轨迹上，通过关联形成新的轨迹。网络流算法、匈牙利匹配算法、多假设追踪算法等都是通过以降低全局匹配为代价来提升匹配效果的。此外，基于深度学习的方法也有所进展：多趟近邻排序（MPN）[19]算法以及深度多目标跟踪（DeepMOT）[20]算法利用卷积神经网络分别模拟传统的网络流算法和匈牙利匹配算法来实现关联匹配，并取得了出色的效果。

3.3 视频物体重识别

对于小视频场景下的通用物体重识别，学术界目前还没有找到很好的解决方法。对于复杂场景下的人物等特定物体重识别来说，我们一般将人物局部入镜的重识别问题定义为局部人物重识别，即利用局部人物图片来检索其完整的人物图片。此外，还有不少关于遮挡人物重识别的研究工作，下面我们将分别进行介绍。

（1）局部人物重识别

早期处理局部人物重识别的方法是直接将局部人物图片和完整人物图片缩放到同样尺寸，这会导致特征不对齐等问题。有的研究则采用滑动窗口方法，利用局部人物图片大小相同的滑动窗口在完整人物图片上进行区域检索，找到最相近的区域进行相似度计算。当局部人物图片的宽度大于完整人物图片时，这类方法就会失效，同时也耗费了很多计算资源。

为了解决局部人物重识别的问题，HE L. X. 等提出了一种深度空间特征重构（DSR）的方法[21]。该方法首先利用全卷积网络生成固定尺寸的特征图，然后利用字典学习模型中的重建误差来计算不同特征图的相似度。SUN Y. F. 等提出一种自监督的方法[22]来解决局部人物重识别的特征不对齐问题。该方法将图片划分为上、中、下3个抽象模块区域，得到每个区域中像素点的区域标签，并以此来训练模型对每个区域的观察能力。在推理阶段，模型通过预测区域可见得分，判断图片是否发生了身体部位的缺失，进而通过自监督的注意力机制实现对人物图片间对应区域的相似度比较。

（2）遮挡人物重识别

不同于局部人物重识别，遮挡人物重识别主要的问题在于图片中包含的遮挡区域会使得直接提取的全局特征包含大量的干扰噪声，进而影响两张图片的相似度计算结果。针对这一点，MIAO J. X. 等[23]通过引入额外的姿态检测模型来获得人体关键点信息，进而引导重识别模型关注人物的非遮挡区域。具体思路是，首先通过关键点的位置信息来提取人物的局部特征，然后利用关键点的置信度信息来判断哪些关键点是处于遮挡区域的。在重识别的推断阶段，模型只会计算两张图片未被遮挡的区域之间的相似度，以此来消除遮挡噪声的干扰。

3.4 针对小视频的研究工作

目前，学术界专门针对小视频特点的研究工作比较少。本文中，我们挑选一些比较突出的相关研究工作进行介绍。

（1）针对小视频复杂特效问题的研究

针对不同镜头间添加的视频特效导致物体外观信息不匹配问题，ZHONG Z. 等于2018年在行人重识别领域提出了相机风格自适应[24]算法。该算法假定，在不同相机风格下拍摄所得的人物数据属于不同的数据域，同时通过引入循环生成对抗网络（CycleGAN）[25]，对每一对具有不同风格的同一人物图像，生成图像到图像的风格转移模型。生成不同相机风格下的人物图像为重识别模型提供额外的训练数据。为了防止重识别模型受到由CycleGAN风格转移得到的伪图像中噪声的影响，算法引入了一个标签平滑修正（LSR）机制，以降低在重识别模型损失函数中对伪图像评判的权重。

（2）针对小视频物体类别繁多的研究

针对物体类别繁多所带来的长尾分布效应，POOJAN O. 与 VISHAL P. 于2019年在图像分类领域提出了基于多任务的开集物体识别（MLOSR）[26]算法。该算法通过使用权值共享的分类网络和解码网络，同时进行分类与重构任务。此外，算法依据极值理论[27]通过一个极值模型来对重构误差分布的尾部部分建模，使得模型对未出现在训练集中的类别更为敏感。

（3）针对小视频镜头切换频繁的研究

针对不同镜头下的物体空间位置变化不连续问题，HSU H. M. 等于2019年在目标追踪领域提出一个多摄像机目标追踪系统[28]，将多个摄像机下的目标追踪问题划分为镜头内的目标追踪问题和镜头间的目标追踪问题。对于镜头内的目标追踪问题，该研究团队采用踪片网络追踪器（TNT）[29]。对于镜头间的目标追踪问题，该研究团队首先将镜头内追踪得到的踪片输入到Mask R-CNN[30]网络中，以得到去除背景后的结果，然后再通过一个时间注意力模型，对各踪片提取踪片级别的特征，最后通过比较特征相似度的方式来匹配不同摄像机下的同一物体。

4 小视频内容分析系统

要系统性完成小视频内容分析任务，单纯依靠某一个算法模块是困难

的。只有构建一个整合多种不同算法的系统，才能够更准确、更鲁棒地解决分析问题。本文在此抛砖引玉，提出一个小视频内容分析系统的构成框图。结合此前提到的小视频数据的特点，以及当前对于视频分析技术的研究成果，我们认为小视频内容分析系统至少应包括镜头分割、视频目标检测、视频目标追踪、视频目标重识别等模块，如图 1 所示。

对于输入的小视频，首先，镜头分割模块将不同镜头分割开来，使得每个镜头内物体运动能基本满足帧间位置连续性假设；接着，目标检测模块获得各帧内物体的定位框和物体分类结果，并将结果输入到后续镜头内的目标追踪模块，同时属于同一物体的检测框在相邻帧中将被关联起来；最后，系统再进行跨镜头目标重识别，得到各物体在小视频中完整的时空运动轨迹。小视频内容分析系统的输出结果可被应用到后续更多的应用处理中，例如实现视频结构化、完成以视频搜索视频等任务。

视频结构化应用的主要目标是，仅从无结构视频数据中解析主要物体的语义属性和时空轨迹等结构化的语义信息，就可以实现人车信息检索以及行为研判等，为交通安全和社会治安提供风险评估和事件预警。以视频搜视频是小视频的一大类应用。常规文字、图片搜索等不能完全满足用户需求，而以视频搜索类似视频的功能在各大应用软件的出现，有助于提升用户体验。小视频内容分析结果使小视频搜索成为可能。此外，小视频查重、溯源等也是类似应用。基于小视频内容分析的各种衍生应用正在日益增多，这将大大改善小视频的用户体验。

5 结束语

小视频应用的兴起是互联网技术发展的必然结果，也是人工智能技术广泛服务人们生活的发展趋势。目前，越来越多的巨头公司和科研机构开始研发小视频内容分析技术，旨在更好地应用人工智能技术分析海量视频数据，以更好地服务社会。随着小视频研究和应用的不断发展，在为受众提供更高质量服务的同时，对小视频数据规范化利用、确保个人隐私和数据安全，正在成为社会大众非常关注的热点问题。

致谢

感谢复旦大学计算机科学技术学院邱泰儒、徐僖禧、王浔彦、陈冠先等为本文写作而做出的大量贡献。

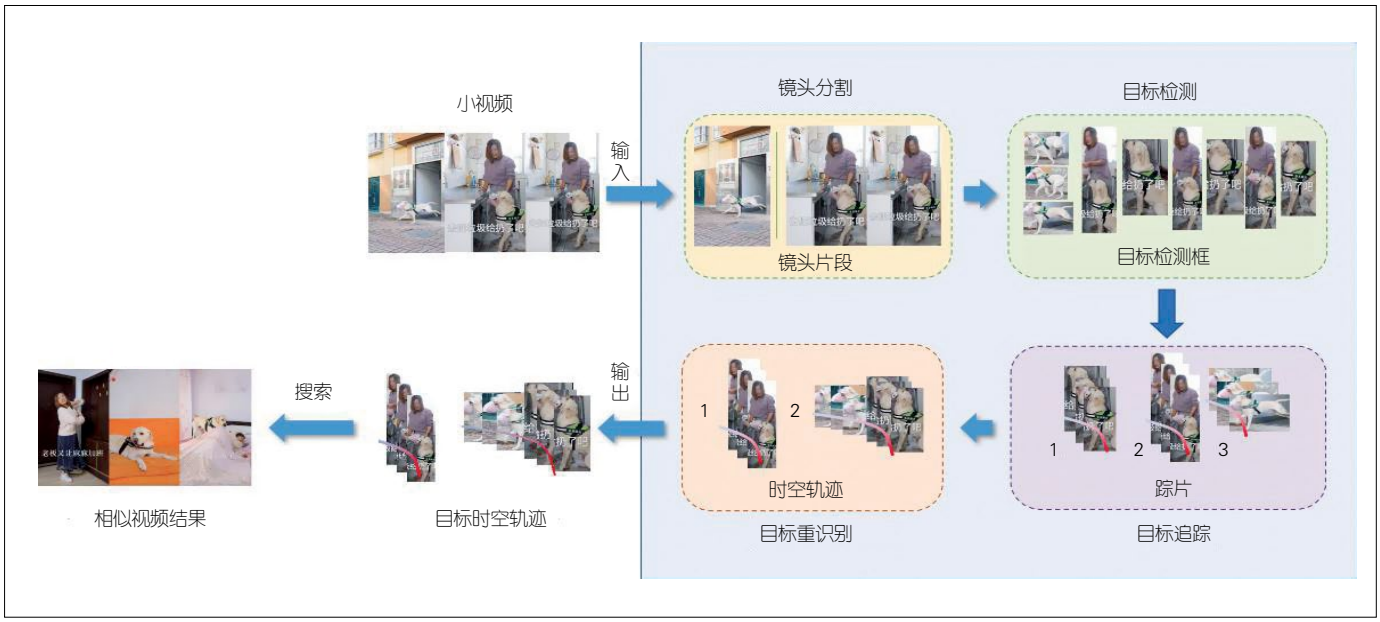
参考文献

[1] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248–255. DOI: 10.1109/cvprw.2009.5206848

[2] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//European conference on computer vision. Zurich, Switzerland: Springer, 2014: 740–755. DOI: 10.1007/978-3-319-10602-1_48

[3] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International journal of computer vision, 2015, 115(3): 211–252. DOI: 10.1007/s11263-015-0816-y

[4] KANG K, LI H S, YAN J J, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos [J]. IEEE transactions on circuits and systems for video technology, 2018, 28(10): 2896–2907. DOI: 10.1109/tcsvt.2017.2736553

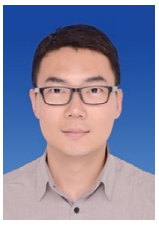


▲图 1 小视频内容分析系统构成框图

[5] LUO H, XIE W X, WANG X G, et al. Detect or track: towards cost–effective video object detection/tracking [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, HI, USA: AAAI, 2019, 33: 8803–8810. DOI: 10.1609/aaai.v33i01.33018803
[6] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: learning optical flow with convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 2758–2766. DOI: 10.1109/iccv.2015.316
[7] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2462–2470. DOI: 10.1109/cvpr.2017.179
[8] ZHU X, WANG Y, DAI J, et al. Flow–guided feature aggregation for video object detection [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 408–417
[9] LU Y, LU C, TANG C K. Online video object detection using association LSTM [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2344–2352
[10] WU H P, CHEN Y T, WANG N Y, et al. Sequence level semantics aggregation for video object detection [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 9217–9225. DOI: 10.1109/iccv.2019.00931
[11] CHEN Y H, CAO Y, HU H, et al. Memory enhanced global–local aggregation for video object detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 10337–10346. DOI: 10.1109/cvpr42600.2020.01035
[12] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008. DOI: 10.1109/cvpr.2008.4587597
[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large–scale image recognition [EB/OL]. [2020–12–05]. https://arxiv.org/abs/1409.1556v1
[14] SUN K, XIAO B, LIU D, et al. Deep high–resolution representation learning for human pose estimation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5693–5703. DOI: 10.1109/cvpr.2019.00584
[15] REN S Q, HE K M, GIRSHICK R, et al. Faster R–CNN: towards real–time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 39(6): 91–99. DOI: 10.1109/tpami.2016.2577031
[16] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: Optimal speed and accuracy of object detection [EB/OL]. [2020–12–05]. https://arxiv.org/abs/2004.10934
[17] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric [C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE, 2017: 3645–3649. DOI: 10.1109/icip.2017.8296962
[18] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking [C]//2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 3464–3468
[19] BRASÓ G, LEAL–TAIXÉ L. Learning a neural solver for multiple object tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 6247–6257
[20] XU Y H, SEP A, BAN Y T, et al. How to train your deep multi–object tracker [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 6787–6796. DOI: 10.1109/cvpr42600.2020.00682
[21] HE L X, LIANG J, LI H Q, et al. Deep spatial feature reconstruction for partial person Re–identification: alignment–free approach [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 7073–7082. DOI: 10.1109/cvpr.2018.00739
[22] SUN Y F, XU Q, LI Y L, et al. Perceive where to focus: learning visibility–aware part–level features for partial person Re–identification [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 393–402. DOI: 10.1109/cvpr.2019.00048
[23] MIAO J X, WU Y, LIU P, et al. Pose–guided feature alignment for occluded person Re–identification [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 542–551. DOI: 10.1109/iccv.2019.00063
[24] ZHONG Z, ZHENG L, ZHENG Z D, et al. Camera style adaptation for person Re–identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 5157–5166. DOI: 10.1109/cvpr.2018.00541
[25] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image–to–image translation using cycle–consistent adversarial networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2223–2232. DOI: 10.1109/iccv.2017.244
[26] OZA P, PATEL V M. Deep CNN–based multi–task learning for open–set recognition [EB/OL]. [2020–12–05]. https://arxiv.org/abs/1903.03161
[27] DE HAAN L, FERREIRA A. Extreme value theory: an introduction [M]. Springer Science & Business Media, 2007
[28] HSU H M, HUANG T W, WANG G, et al. Multi–camera tracking of vehicles based on deep features re–ID and trajectory–based camera link models [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 416–424
[29] WANG G A, WANG Y Z, ZHANG H T, et al. Exploit the connectivity: multi–object tracking with TrackletNet [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 482–490. DOI: 10.1145/3343031.3350853
[30] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R–CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2961–2969. DOI: 10.1109/iccv.2017.322

作 者 简 介

薛向阳，复旦大学计算机科学技术学院教授、博士生导师；主要从事计算机视觉、视频大数据分析、机器学习等研究；发表论文 200 余篇，其中 90 余篇发表在国际权威期刊（如《IEEE Transactions on Pattern Analysis and Machine Intelligence》《IEEE Transactions on Image Processing》等）和顶级国际会议（如 ICCV、CVPR、ICML、NeurIPS、ACM MM、IJCAI、AAAI 等）上。



李斌，复旦大学计算机科学技术学院青年研究员、博士生导师，上海高校特聘教授（东方学者）；研究领域为机器学习、类脑人工智能及其在机器视觉与大数据分析中的应用；在《IEEE Transactions on Knowledge and Data Engineering》《IEEE Transactions on Cybernetics》等知名期刊与 ICML、NeurIPS、IJCAI、AAAI 等一流机器学习和人工智能会议上发表论文 60 余篇。