

基于多粒度特征的行人跟踪检测结合算法

王子晔¹ 苗夺谦^{1,2} 赵才荣^{1,2} 罗 晟¹ 卫志华^{1,2}

¹(同济大学计算机科学与技术系 上海 201804)
²(嵌入式系统与服务计算教育部重点实验室(同济大学) 上海 201804)
(yeziwang@tongji.edu.cn)

A Pedestrian Tracking Algorithm Based on Multi-Granularity Feature

Wang Ziyue¹, Miao Duoqian^{1,2}, Zhao Cairong^{1,2}, Luo Sheng¹, and Wei Zhihua^{1,2}

¹(Department of Computer Science and Technology, Tongji University, Shanghai 201804)
²(Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai 201804)

Abstract Recently in some popular applications, such as video scene surveillance, long-term effective pedestrian tracking is the basis of these applications. Although the related technology of target detection and target tracking have a long history, how to achieve real-time and accurate pedestrian tracking is still an active research field and needs to be solved. At present, most pedestrian tracking methods only use hand-designed features to track or only use deep learning to extract features, which are not good ways to represent the features of the target because the use of one single feature will restrict the expression of the features. Therefore, multi-granularity hierarchical features are used in this paper to achieve more stable pedestrian tracking. This paper proposes an improved pedestrian tracking algorithm. The algorithm adopts the idea of multi-granularity, combines convolutional feature with bottom color feature, makes decision on the tracking result obtained by GOTURN, a tracking algorithm based on deep learning, and modifies the tracking result with target detection. This paper uses Pascal VOC data set for model training, and uses OTB-100 and VOT 2015 data sets for testing. The experimental results show that the tracking algorithm based on multi-granularity decision can track target pedestrians more accurately than a single tracking algorithm and the tracking accuracy is improved obviously.

Key words pedestrian tracking; object detection; deep learning; multiple-granularity; robustness

摘 要 对于一些较为流行的应用,例如视频场景监控,对行人的长期有效跟踪是应用的基础.尽管对目标检测与跟踪的相关技术研究已经有了很长的历史,但是如何实时并较为准确地实现目标行人跟踪目前仍然是一个活跃的研究领域.基于多粒度的思想,提出了一种改进的行人跟踪算法,将卷积特征与底层颜色特征结合,对基于深度学习的跟踪算法 GOTURN(generic object tracking using regression networks)得到的跟踪结果进行判断决策,结合目标检测对跟踪结果进行修正.实验结果表明:与单一的跟踪算法相比,多粒度决策的跟踪算法能够更加准确地对目标行人进行跟踪,可以显著提高跟踪精度.

关键词 行人跟踪;目标检测;深度学习;多粒度;鲁棒性

中图法分类号 TP181

目前,低成本视觉采集机器逐渐发展成熟,可以便捷地获取大量的视频数据,极大地促进了视觉多媒体研究与应用.例如在智能监控系统中,对监控视频中可疑的目标人物进行长期跟踪已经成为计算机视觉研究领域中的重要课题.行人跟踪任务的目的是在一段连续视频中根据第1帧中已经标注的某个特定目标行人的位置及相关信息,来对这个特定的目标行人进行跟踪,最终记录得到目标行人在视频每一帧中的位置信息.

传统的目标跟踪方法^[1]普遍采用概率密度方法、人工设计特征等方法直接对图像进行检测^[2],从而在视频的每一帧中定位出感兴趣的运动目标.例如使用手工设计的方向梯度直方图(histogram of oriented gradient, HOG)特征^[3]进行跟踪的方法,这种方法不依赖先验知识,因此很难适应复杂的环境变化,鲁棒性与精度都有待提高.相反,基于深度学习的方法可以自动从大数据中学习得到有效的特征表示,因此跟踪效果较好.

深度学习^[4]在计算机视觉的各个领域一直都有很广泛的应用,例如图像分类、目标检测、语义分割等.随着大数据的不断发展以及计算机计算能力的不断提高,近年来深度学习也逐渐开始被应用于目标跟踪领域.2013年Wang等人^[5]提出使用叠加去噪自动编码器(stacked denoising autoencoder, SADE),通过无监督预训练从大量数据中提取目标特征,再使用粒子滤波器来进行在线跟踪.这也是第一个引入深度学习模型的单目标跟踪任务.2014年提出的结构化输出深度学习跟踪器(structured output deep learning tracker, SO-DLT)^[6]算法,成功使用大规模卷积神经网络来提取特征.Long和Shelhamer^[7]通过探索来自不同层的卷积特征来表达目标属性中的区别与联系,并且通过微卷积神经网络使其稀疏化,从而得到更好的特征来进行追踪,有效避免了跟踪器漂移的现象.POSTECH团队提出多域学习的概念,将每一个视频数据作为一个单独的域,通过多个域中目标的共享表示来进行跟踪,也就是MDNet(multi-domain convolutional neural networks)算法^[8].Held等人^[9]于2016年提出一种基于回归网络的深度视觉跟踪算法GOTURN(generic object tracking using regression networks),使用大量的视频与图像数据进行离线学习,使网络

能够学习到对象的外观模型以及运动模型,这也是第1次使用深度学习的目标跟踪算法在速度上达到100 fps.

GOTURN算法使得基于深度学习的实时跟踪方法成为可能,然而行人是非刚性的目标,场景以及形状变化、视角变化等复杂性使得跟踪器无法很好的根据先前帧中的信息有效地锁定目标行人.相反,采用深度学习的Faster R-CNN(region proposal convolutional neural network)目标检测算法^[10],可以在运动模糊或部分遮挡的情况下对行人有较好的检测结果.因此我们考虑结合目标跟踪与目标检测的优点,提出一种新的模型来实现对行人的长期稳定跟踪.

目前存在的跟踪检测结合算法主要为TLD(tracking-learning-detection)算法^[11],是一种新的单目标长时间跟踪算法,将传统的跟踪算法和检测算法结合,并通过一种改进的在线学习机制对跟踪模块的相关参数进行不断更新,从而得到更加可靠稳定的跟踪.但是同时,TLD算法也存在目标形变易丢失等问题.

为了提高模型的跟踪精度,我们采用多粒度特征的方法对跟踪结果进行判定,从而判断跟踪结果与目标是否为同一个行人.由于使用单一特征对图像判定会造成局限性,因此采用多粒度层次的特征,来实现对跟踪目标的更稳定准确的判定.

本文将粒计算的思想应用到目标跟踪过程中,采用多粒度特征的方法对GOTURN目标跟踪算法进行改进,结合目标检测算法实现了长期稳定的行人跟踪算法,取得了较好的跟踪效果.

1 相关工作

在本节中,我们主要介绍多粒度特征思想以及基于深度学习的目标跟踪方法与目标检测方法.

1.1 多粒度特征

彩色图像在不同粒度下进行表示可以得到不同粒度层次的特征^[12].例如图像的颜色、纹理、形状等底层特征以及使用卷积神经网络得到的深层特征.每一种特征都具有其自身的局限性,无法对所有的图像都有良好的表示效果.因此可以考虑将卷积特

征与图像底层特征进行结合,对图像进行判别.与单一的特征表示相比,多粒度特征对图像 的表示更加准确^[13].

1.2 GOTURN 跟踪算法

目前已经有多种使用深度学习方法对视频中单个物体进行跟踪的算法,但是大部分算法的速度难

以保证,无法满足实时性的需求.而 GOTURN 算法的特点是使用大量的图片和视频,使用完全离线的方式进行预训练,再对目标进行在线跟踪^[9].这种方法可以使跟踪的速度达到 100fps,具有一定的商业应用价值.图 1 所示为 GOTURN 目标跟踪算法流程图.

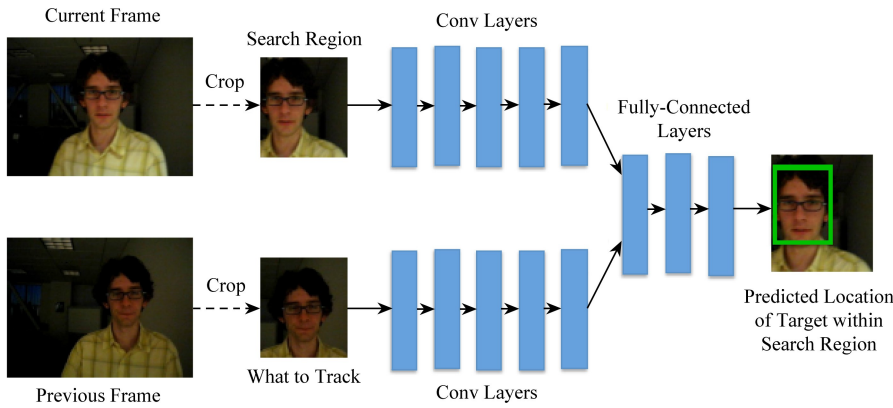


Fig. 1 Network architecture for GOTURN^[9]
图 1 GOTURN 算法流程图^[9]

算法假定目标的运动速度不会非常快,也就是在相邻的 2 帧中,目标的位置距离相对较小.假设目标在上一帧的所在位置中心为 $c = (c_x, c_y)$,并且处在一个宽为 w 、高为 h 的矩形框中,则在当前帧和上一帧中,同样以 $c = (c_x, c_y)$ 为中心,切割大小为 (k_1w, k_1h) 的图像块.将 2 个切割好的图像块分别输入到卷积神经网络中,提取相应的特征.将提取得到的特征输入全连接层进行回归,比较当前帧的特征与上一帧的特征,从而找到目标在当前帧的位置.在实际的实验中,多设定 $k_1 = 2$.

1.3 Faster R-CNN 检测算法

目前许多计算机视觉任务,特别是目标检测领域中都引入了深度学习的概念,提出了许多基于 CNN (convolutional neural network) 的算法. R-CNN^[14] 和 Fast R-CNN^[15] 都依赖 CNN 从提案框中提取特征,但提案框生成的速度还不够快.因此,为了减少生成提案的计算负担,提出了 Faster R-CNN.它可以被视为区域提案网络 (region proposal network, RPN)^[10] 与 Fast R-CNN^[15] 的一种组合, RPN 网络可以快速获取提案,而 Fast R-CNN 网络的目的则是对提案进行特征提取. Faster R-CNN 最重要的思想是 RPN 与 Fast R-CNN 共享卷积层,如图 2 所示.通过这种方式,图像仅通过 1 次卷积神经网络,并且能够有效地提取出提案框的位置信息.

在 Faster R-CNN 模型中,输入图像首先经过

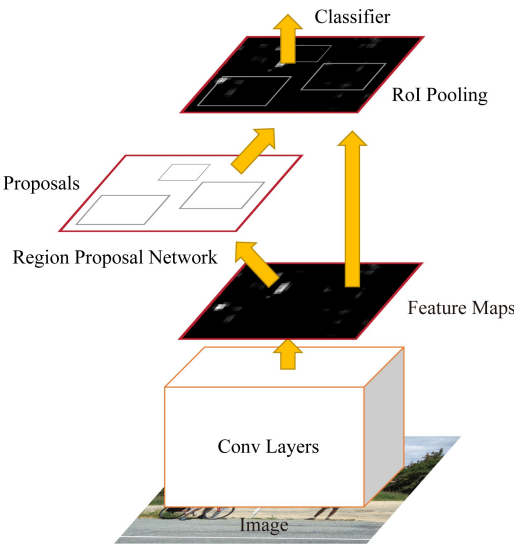


Fig. 2 Network architecture for Faster R-CNN^[10]
图 2 Faster R-CNN 算法流程图^[10]

由 13 个卷积层、13 个激励层和 4 个池化层组成的卷积网络,提取出整个图像的特征图.然后将特征图作为 RPN 网络的输入,得到区域方案.ROI (region of interest) 池化层收集提案和功能图,并创建提案框对应特征图,将其发送给分类器,以计算类的分数值并进行回归处理,获得准确的边界框.本文中使 用 Faster R-CNN 来对出现错误或较大误差的目标跟踪结果进行修正.

2 改进的行人跟踪算法

近年来,基于深度学习的跟踪算法得到了深入的研究.然而,跟踪器可能会在复杂多变的现实情况中发生误差积累、跟踪漂移、目标丢失等情况.这是因为大多数跟踪器使用前一帧中的信息在当前帧中获得结果,如果目标物体移动过快或被其他物体遮挡,跟踪器无法在搜索区域中获取匹配信息,因此很可能会丢失目标或得到错误的跟踪结果.

本文的研究重点是在复杂情况下,例如目标快速运动、目标部分遮挡等情况下,对目标行人的跟踪效果进行优化.大部分基于深度学习的目标跟踪算法通过卷积提取深度层次特征后,再对特征进行回归运算最终的到相应的跟踪结果.其余的目标跟踪算法则主要基于图像内容来进行检索,主要利用图像底层特征,例如颜色、纹理、形状、空间分布信息等.其中,颜色是彩色图像中最显著的视觉特征之一,可以为图像的分析提供丰富的信息.同时,图像的颜色特征稳定性相比其他特征更好,对于图像的方向、大小等变化较为不敏感,可以较好地捕捉图像的全局特征,因此颜色特征是图像检索领域中最为广泛应用的特征.常见的提取颜色特征的方法为颜色直方图方法(color histogram)^[16].为了提升跟踪效果,需要增强对跟踪目标的特征表示,基于多粒度的思想,考虑将深层卷积特征与底层颜色特征结合,从而实现对目标的稳定跟踪.

在使用多粒度特征联合的方法得到跟踪结果后,为了对跟踪结果进行修正,采用目标跟踪与目标检测结合的方法.如果目标跟踪结果较差,则调用目标检测器对跟踪结果进行修正;如果目标跟踪结果较好,则直接根据当前帧跟踪结果信息进行下一帧的跟踪.

算法 1. 多粒度特征联合目标跟踪算法.
输入:上一帧图片 Img_{t-1} 、当前帧图片 Img_t ;
输出:当前帧的跟踪结果 $bbox_t$.
① 根据上一帧 Img_{t-1} 的跟踪结果,即中心 $c_{t-1} = (c_{xt-1}, c_{yt-1})$,大小为 (w_{t-1}, h_{t-1}) ,在当前帧 Img_t 与上一帧中同时切割中心为 c_{t-1} 大小为 (kw_{t-1}, kh_{t-1}) 的区域,得到 $crop_t$ 和 $crop_{t-1}$.
② 将 $crop_t$ 和 $crop_{t-1}$ 分别输入卷积层,得到相应特征图,并将特征图作为全连接层的输入,进行回归的到当前帧上的目标框 $bbox_t$.

- ③ 根据 $bbox_t$ 计算对应位置图片的颜色直方图.
- ④ 与上一帧中 $bbox_{t-1}$ 对应位置图片的颜色直方图进行比较,如果差距小于阈值 Th_corr ,认为当前帧跟踪结果正确,进入下一帧的跟踪.
- ⑤ 如果差距大于阈值 Th_corr ,认为当前帧跟踪结果误差较大,调用目标检测器.
- ⑥ 选取目标检测的多个检测结果中与上一帧跟踪结果 $bbox_{t-1}$ 最相似的一个作为当前帧的跟踪结果.

改进算法的流程图如图 3 所示:

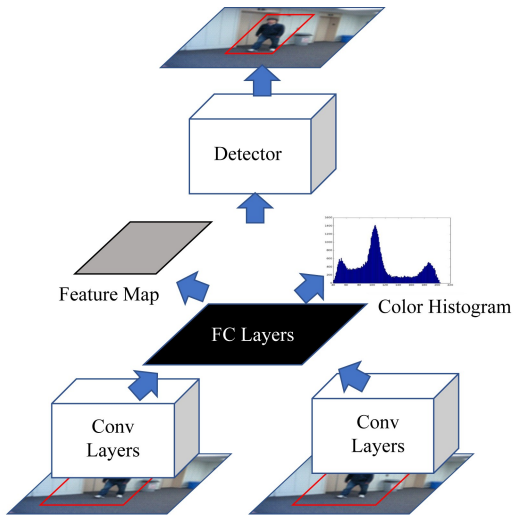


Fig. 3 Framework of our model

图 3 行人跟踪模型网络结构

总的来说,本文提出的改进跟踪算法可以有效地处理一些特殊情况,例如目标遮挡、运动模糊等,避免目标丢失.

3 实验及结果分析

3.1 训练

本文中使用 Caffe 深度学习框架,使用 Pascal VOC2007 数据集对目标检测器进行训练. Pascal VOC2007 数据集将目标分为 20 类,包含 9 963 张图片以及 26 460 个 bounding box.在本文实验中仅使用标签类别为行人的相关数据对模型进行训练,选择 VGG16 卷积神经网络模型并设置迭代次数为 [80 000, 40 000, 80 000, 40 000].训练后的模型对行人进行检测的 mAP(mean average precision)为 0.777.

3.2 测试数据集

本文使用的测试数据集包含 OTB-100 数据

集^[17-18]中的16个视频序列和VOT 2015数据集^[19]中的14个视频序列,OTB-100数据集和VOT 2015数据集都是目标跟踪领域的基准数据集,其中每个视频序列均有不同的目标类别以及属性,例如目标遮挡、运动模糊等.我们从中挑选以行人作为目标的视频序列来进行测试,使用精度和成功率作为评估标准.

3.3 实验结果分析

本文分别使用近年来基于深度学习且跟踪效果

较为出众的MDNet目标跟踪模型以及GOTURN目标跟踪模型与改进的多粒度目标跟踪检测结合模型来对测试数据进行实验,获取指定目标行人的位置以及相应的边界框坐标,如图4所示.将得到的跟踪结果边界框与实际位置边界框进行误差分析.采用one pass evaluation(OPE)的方法对跟踪结果进行评估,使用ground-truth中目标的位置对第一帧进行初始化,然后运行跟踪算法得到平均精度和成功率.

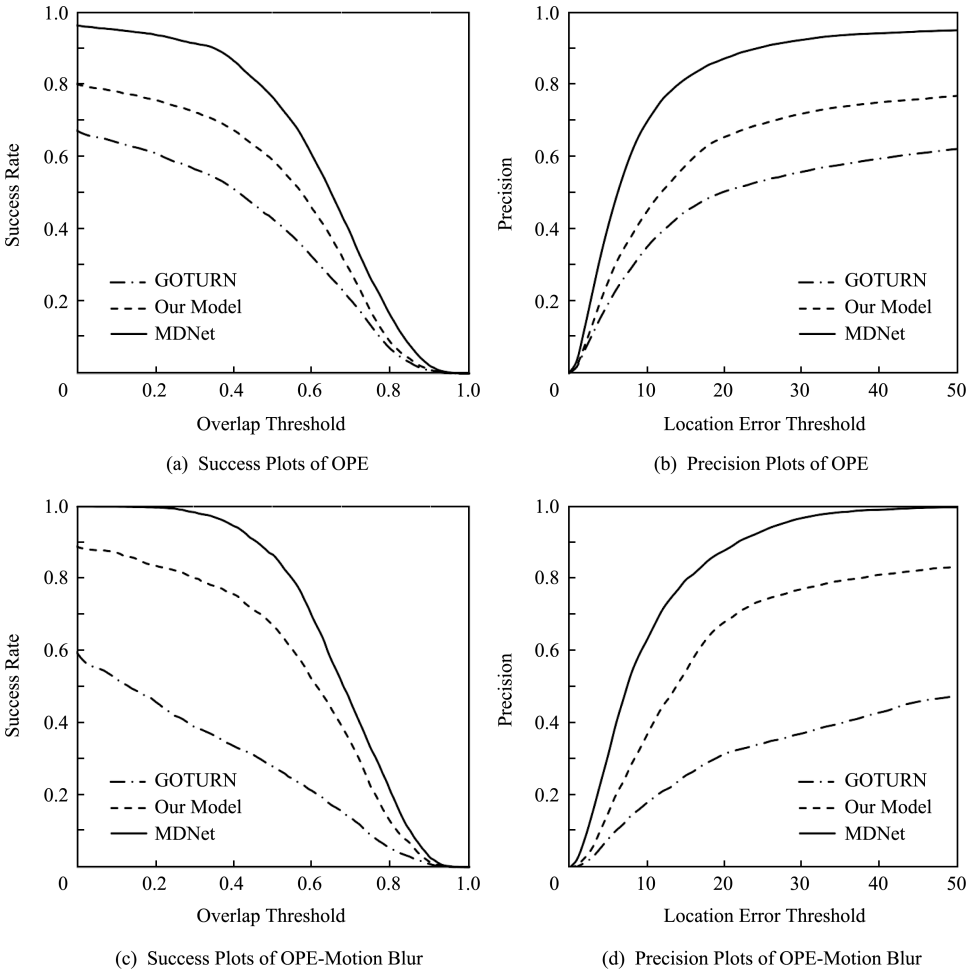


The white solid boxes are the ground truth b-boxes, the dark dotted boxes are the tracking results.

Fig. 4 Target pedestrian tracking results
图4 目标行人跟踪结果

如图5所示,设置目标遮挡情况的位置误差阈值为20时,可以看到本文提出模型的精度为0.642,而GOTURN的精度为0.416,MDNet的精度则为0.950.当目标遮挡情况的重叠阈值为0.5时,本文模

型和GOTURN的成功率分别为0.556和0.345.设定运动模糊情况的位置误差阈值为20时,本文模型和GOTURN的精度分别为0.682和0.315.当运动模糊情况的重叠阈值为0.5时,本文模型和GOTURN



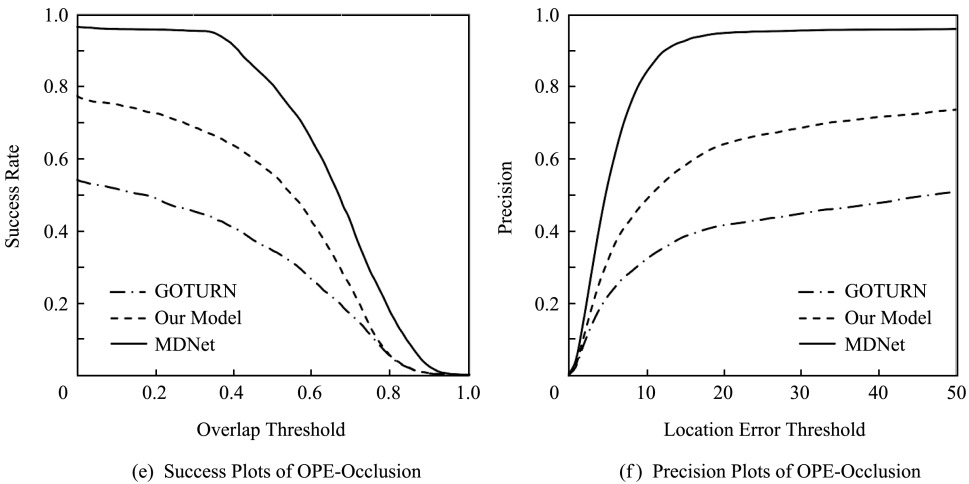


Fig. 5 Success rate plots and precision plots for sequences with different attributes

图 5 不同属性视频序列组的成功率与精度

Table 1 Success Rate and Precision of Different Models

表 1 不同模型的成功率、精度、平均跟踪速度

Situation	Our Model		GOTURN		MDNet	
	Precision	Success Rate	Precision	Success Rate	Precision	Success Rate
All	0.655	0.589	0.504	0.427	0.872	0.765
Motion Blur	0.682	0.670	0.315	0.279	0.879	0.866
Occlusion	0.642	0.556	0.416	0.345	0.950	0.803

的成功率分别为 0.670 和 0.279.从表 1 中可以非常直观地看出,在各种复杂情况中,本文模型的跟踪精度相比于 GOTURN 模型有很大的提升.虽然精度上与 MDNet 模型仍然存在一定差距,但是在平均跟踪速度上,本文模型具有很大的优势,如表 2 所示,MDNet 的平均跟踪速度为 2.403 fps,而本文改进模型跟踪速度可达 16 fps.

色直方图特征共同使用,目标跟踪检测结合的方法建立了长期稳定的目标行人跟踪系统.实验表明,提出的方法可以有效地对跟踪结果进行判定,并且对行人跟踪系统的精度有很大的提升.在接下来的研究工作中,可以尝试加入其他不同粒度不同层次的特征,来丰富图像的特征表示,从而对图像实现更优的判定效果,进一步提升对目标行人的跟踪精度,并且可以对不同大小的粒度对跟踪模型所起到的作用进行更进一步的探索,从而选取更有效的粒度特征进行融合来提升跟踪模型的性能.

Table 2 Average Tracking Speed of Different Models

表 2 不同模型的平均跟踪速度

Model	Average Tracking Speed/fps
Our Model	16
GOTURN	148
MDNet	2.403

总的来说,本文提出的跟踪模型在较为复杂的情况下跟踪效果相对较好.实验结果表明,与单独使用 GOTURN 的跟踪器相比,本文构建的模型可以有效地提高长期行人跟踪的精度和稳定性.

4 结 论

本文采用多粒度特征的思想,将卷积特征与颜

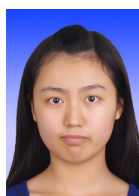
参 考 文 献

[1] Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient convolution operators for tracking [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2017: 6638-6646

[2] Zhang Huanlong, Hu Shiqiang, Yang Guosheng. Video object tracking based on appearance models learning [J]. Journal of Computer Research and Development, 2015, 52(1): 177-190 (in Chinese)

(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[J]. 计算机研究与发展, 2015, 52(1): 177-190)

- [3] Kaaniche M B, Brémond F. Tracking hog descriptors for gesture recognition [C] //Proc of the 6th IEEE Int Conf on Advanced Video and Signal Based Surveillance. Piscataway, NJ: IEEE, 2009: 140-145
- [4] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444
- [5] Wang Naiyan, Yeung D Y. Learning a deep compact image representation for visual tracking [C] //Proc of the 26th Advances in Neural Information Processing Systems. New York: Curran Associates, 2013: 809-817
- [6] Wang Naiyan, Li Siyi, Gupta A, et al. Transferring rich feature hierarchies for robust visual tracking [J]. arXiv preprint, arXiv:1501.04587, 2015
- [7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2015: 3431-3440
- [8] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2016: 4293-4302
- [9] Held D, Thrun S, Savarese S. Learning to track at 100 fps with deep regression networks [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 749-765
- [10] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [C] //Proc of the 28th Advances in Neural Information Processing Systems. New York: Curran Associates, 2015: 91-99
- [11] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection [J]. IEEE Transactions on Software Engineering, 2011, 34(7): 1409-1422
- [12] Xu Xiangli, Zhang Libiao, Yu Zhezhou, et al. Application of multi-granularity color features in image retrieval [J]. Journal of Applied Sciences—Electronics and Information Engineering(in Chinese), 2009, 27(1): 56-61
(Xu Xiangli, Zhang Libiao, Yu Zhezhou, et al. Application of multi-granularity color features in image retrieval [J]. 应用科学学报, 2009, 27(1): 56-61)
- [13] Guo Qingwen, Wang Guoyin, Zhang Qinghua. Research on multi-granularity image retrieval method [J]. Computer Science, 2013, 40(1): 298-301 (in Chinese)
(郭庆文, 王国胤, 张清华. 多粒度的图像检索方法研究[J]. 计算机科学, 2013, 40(1): 298-301)
- [14] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2014: 580-587
- [15] Girshick R. Fast R-CNN [C] //Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 1440-1448
- [16] Zivkovic Z, Krose B. An EM-like algorithm for color-histogram-based object tracking [C] //Proc of the 17th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2004: 798-803
- [17] Wu Yi, Lim J, Yang MingHsuan. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1834-1848
- [18] Wu Yi, Lim J, Yang M H. Online object tracking: A benchmark [C] //Proc of the 26th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2013: 2411-2418
- [19] Kristan M, Matas J, Leonardis A, et al. The visual object tracking vot2015 challenge results [C] //Proc of the 15th IEEE Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2015: 564-586



Wang Ziyi, born in 1997. PhD candidate. Her main research interests include object tracking, computer vision, multiple-granularity.



Miao Duoqian, born in 1964. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include rough set, deep learning and data mining.



Zhao Cairong, born in 1981. PhD, professor, PhD supervisor. Member of CCF. His main research interests include pattern recognition, machine learning, data mining.



Luo Sheng, born in 1982. PhD. His main research interests include multiple-granularity, rough set, pattern recognition.



Wei Zhihua, born in 1979. PhD, professor, PhD supervisor. Her main research interests include natural language processing, machine learning, data mining.