文章编号:1671-6833(2021)05-0019-06

融合图像深度的抗遮挡目标跟踪算法

王希鹏,李 永,李 智,张 妍

(武警工程大学 信息工程学院,陕西 西安 710086)

摘 要:由于视频信息的局限性,在遮挡情况下的目标跟踪依然是一个很难解决的问题。针对目标跟踪过程中的遮挡问题,提出将图像深度引入单目标跟踪算法。首先应用单目图像深度估计算法对图像进行深度估计,获取图像的深度信息;其次,将基于孪生区域推荐网络的目标跟踪算法与图像深度相结合,构建遮挡判别模块,利用目标深度信息的变化判断遮挡情况;最后,将遮挡判别得分与锚点框响应得分进行加权融合,根据最终响应得分对目标跟踪器的候选框重新排序,避免目标跟踪被遮挡物干扰而产生跟踪漂移。在 OTB-2015 数据集上的实验结果表明:该算法能有效地应对遮挡情况对跟踪性能的影响,平均跟踪成功率为 0.623,平均跟踪精确度为 0.853,相比基准算法分别提高了 1.8%和 0.9%。

关键词:孪生网络;深度学习;目标跟踪;单目图像深度估计;抗遮挡

中图分类号:TP391.4 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2021.05.011

0 引言

目标跟踪是计算机视觉研究领域的热点之一,被广泛应用于自动驾驶、智能视频监控、人机交互等多个方面。单目标跟踪是指在给定第一帧目标框的情况下,在视频的后续帧中自动地标出该目标的位置和大小。早期的单目标跟踪算法以相关滤波为主,CSK[1]相关滤波算法采用灰度特征,KCF[2]算法在 CSK 算法的基础上做出了改进,采用了 HOG 特征。近几年,随着目标跟踪数据集的扩充、跟踪标准的完善、深度学习模型的不断优化,基于深度学习的目标跟踪方法取得了很好的成绩。SINT[3]是第一个使用孪生网络解决目标跟踪问题的算法。SiamFC[4]算法由于是端到端的跟踪网络,速度方面有了很大的提升,这使得基于孪生神经网络的跟踪器真正地流行起来。SiamRPN[5]在孪生网络的基础上将目标检测算法 Faster R-CNN 中的 RPN 模块应用到跟踪任务上来,回归分支代替了原始的尺度金字塔,因此该算法在提升精度的同时,速度也得到了提升。

在单目标跟踪问题中,算法的性能受到环境因素的影响,主要包括光照变化、尺度变化、目标遮挡等。尽管基于孪生网络的跟踪器已经取得了优异的性能,但依然存在一些缺陷。目标遮挡在目标跟踪任务中经常出现,在很大程度上影响了目标跟踪算法的性能。大部分基于孪生网络的跟踪器选取第 1 帧目标为模板,之后每一帧的搜索区域与目标模板比较,计算相似度,由于不更新目标模板,在目标被遮挡或发生形变时就会发生跟踪漂移。基于孪生网络的跟踪器仅提取图像深度特征用于视觉跟踪,忽略了图像中的语义信息,这导致基于深度学习的跟踪器在目标遇到遮挡或运动模糊的时候,跟踪性能只能依赖离线训练时特征的质量。Zhu 等[6]提出了 FlowTrack 跟踪器,将光流信息整合到端到端的深度网络中,光流在计算上比较慢,而且光流仅提取连续帧中的运动信息。Wu 等[7]用卡尔曼滤波器构造目标运动模型。

上述跟踪器使用了图像深度特征和目标的运动特征,但是忽略了图像中高层的语义信息。人眼在跟踪目标时,直觉上会估计出图像中场景的深度信息。场景的深度估计是机器人视觉领域的研究重点之一。很多研究将场景深度信息与跟踪任务相结合[8],但场景深度信息的提取都基于 RGB-D 相机,对硬件设备要求较高。单目深度估计领域相关技术的发展使得不需要 RGB-D 相机

收稿日期:2020-07-29;**修订日期**:2020-10-09

基金项目:全国教育科学"十三五"规划课题(JYKYB2019012)

通信作者:李永(1981—),男,陕西永寿人,武警工程大学副教授,博士,主要从事模式识别和视频智能分析研究,E-mail:llililil819@163.com。

也可以得到场景的深度信息。早期单目深度估计的方法大多基于机器学习。Godard 等[9]提出Monodepth2算法,使用深度估计和姿态估计网络的组合来估计单目图像中的深度,提升了深度估计的性能。通过单目图像深度估计可以准确得到图像场景中的深度信息,并利用像素值进行表示:当目标离相机较远时,像素值较小;离得较近时,像素值较大;当目标被遮挡时,跟踪框内的像素值会发生改变,从而可以判断跟踪目标是否被遮挡,并对跟踪器得到的目标位置进行修正。

本文主要针对目标跟踪中的遮挡问题,提出将 Monodepth2 算法得到的场景深度信息融合到 SiamRPN 跟踪器中,根据 Monodepth2 得到每一帧图像的深度信息:当目标没有被遮挡时,目标深度值在时序上的变化是平滑的;当发生遮挡时,目标深度值会快速变化。将 SiamRPN 与 Monodepth2 得到的响应图进行融合,克服目标被遮挡时产生的跟踪漂移问题。为验证本文算法的性能,利用 OTB-2015[10]数据集中部分存在遮挡的视频序列进行测试。

1 融合图像深度的抗遮挡目标跟踪算法

1.1 基于孪生网络的目标跟踪算法

孪生网络主要用来衡量输入样本的相似性。SiamFC[4]分为模板分支和搜索区域分支:模板分支输入 x,经过特征提取网络 φ,可以得到一个卷积核 $\varphi(x)$;搜索区域分支输入 z,经过特征提取网络 φ,得到一个候选区域 $\varphi(z)$。$\varphi(x)$ 与 $\varphi(z)$ 进行互相关操作,得到一个响应图,如式(1)所示,其中⊗代表互相关运算。从响应图中选取响应最大的位置,作为目标当前的位置,进行多尺度测试,得到目标当前的尺度,如式(2)所示。

$$f(x,z)=\varphi(x)\otimes\varphi(z); \tag{1}$$

$$p=\operatorname{argmax}[\varphi(x)\otimes\varphi(z)]。 \tag{2}$$

SiamRPN 将原来目标跟踪任务中的相似度计算转化为回归和分类问题,其中的 RPN 模块可以理解为一种全卷积网络,该网络最终目的是为了推荐候选区域。RPN 中最重要的是 anchor 机制,通过预先定义 anchor 的尺寸和长宽比引入多尺度方法。SiamRPN 中 anchor 有 5 种长宽比:0.33、0.5、1、2、3。通过平移和缩放对原始的 anchor 进行修正,使 anchor 更接近真实的目标窗口。

1.2 单目图像深度估计算法

由于像素级深度数据集难以获取,使得监督学习在单目深度估计中的应用受到限制,所以基于自监督学习或无监督学习的单目图像深度估计的研究越来越多。单目深度估计的输入为一帧 RGB 图像,输出为深度图,深度图中每个像素值表示该像素在空间中的位置 L_p。Monodepth2 算法利用单目和双目图像序列在自监督框架上进行训练。采用最小化像素投影损失,对每个像素进行计算,解决遮挡问题:

$$L_p=\sum_{t'}pe(I_t,I_{t'\to t})。 \tag{3}$$

式中:pe 表示光度重建误差;I_t 表示每个目标视图;$I_{t'\to t}$ 表示相对位姿。采用 auto-masking 方法过滤掉序列中相邻两帧固定不变的像素。选用多尺度结构,将低分辨率深度图上采样到输入分辨率,在高分辨率下重投影、重采样,计算光度重建误差。

在目标跟踪中,计算目标框区域内深度平均值作为跟踪目标的平均深度值:当目标没有被遮挡时,目标平均深度值在时序上的变化是平滑的;当目标被遮挡时,目标平均深度值会快速变化。本文采用在单目和双目数据集中训练的模型。Monodepth2 算法深度图如图 1 所示,两张图片均来自目标跟踪的 OTB-2015 数据集,左侧均为原始图像,右侧均为由 Monodepth2 得到的深度图。

(a) Girl2视频序列原图 (b) Girl2视频序列深度图

(c) Walking2视频序列原图 (d) Walking2视频序列深度图

图 1 Monodepth2 算法效果

Figure 1 Monodepth2 algorithm effect

1.3 融合图像深度的抗遮挡目标跟踪算法

在目标跟踪任务中,解决部分遮挡通常有两种思路:一种是利用检测机制判断目标是否被遮挡,如果被遮挡,则更新模板,提升模板对遮挡的鲁棒性[11];另一种是把目标分成多个块,利用没有被遮挡的块进行跟踪[12]。根据人的视觉知觉,当人在对视频中目标进行视觉跟踪时,会估计出视频中场景的层次关系,判断目标和干扰物的位

置关系,减小跟踪过程中遮挡对目标跟踪的影响。根据上述思路,本文提出将图像深度信息引入到单目标跟踪算法中,构建遮挡判别模块,利用目标深度信息的变化判断遮挡情况并修正跟踪结果。

本文在 SiamRPN 跟踪算法中引入单目图像深度估计,利用深度信息进行遮挡判别,在发生遮挡时对 SiamRPN 的跟踪结果进行修正。算法框架如图 2 所示。算法输入为第 t 帧图像和第 $t-1$ 帧跟踪目标的深度图。将原图像同时输入孪生网络跟踪器和深度估计算法,分别得到搜索区域内所有的锚点框、对应的响应得分和搜索区域深度图,将以上输出信息和前一帧跟踪目标的深度图输入遮挡判别模块得到预测的目标位置。

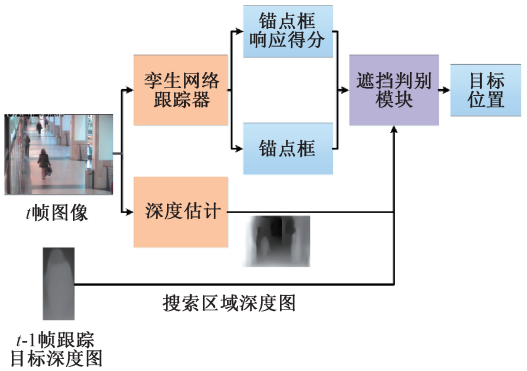


图 2　融合图像深度的抗遮挡目标跟踪算法框架

Figure 2　Framework of anti-occlusion target tracking algorithm based on image depth

每个锚点框的位置表示为 (x_c, y_c, w, h),其中 (x_c, y_c) 表示锚点框的中心位置坐标,w 和 h 分别表示锚点框的宽和高。$D_t(x,y)$ 表示当前深度图 (x,y) 位置上的像素值。D_i 为锚点框内的平均深度值。遮挡判别模块将所有 N 个锚点框的平均深度值 D_i 与前一帧的跟踪目标平均深度值 D_p 作比较,得到深度差值。对所有锚点框的所有深度差值求平均值,得到平均深度差值 M,计算式为

$$D_i = \mathrm{mean}\left(\sum_{x,y} D_t(x,y)\right); \tag{4}$$

$$M = \frac{1}{N}\sum_{i=1}^{N} |D_i - D_p|。 \tag{5}$$

通过锚点框与对应的响应得分 S_1 得到当前最佳锚点框。由于目标跟踪任务中,环境因素的影响会导致深度估计准确率降低,固定的遮挡阈值会造成过多的遮挡误判,所以本文算法将平均深度差值 M 设置为遮挡阈值,遮挡判别计算过程如下:

$$B = \begin{cases} 1, & |D_s - d| \geqslant M; \\ 0, & |D_s - d| < M。 \end{cases} \tag{6}$$

式中:d 为当前帧前 15 帧跟踪目标的累计深度平均值;D_s 为最佳锚点框的平均深度值;当 $B=1$ 时,判断目标为遮挡状态,$B=0$ 时,判断目标未被遮挡。

当判断目标被遮挡时,算法对孪生网络跟踪器生成的锚点框进行修正,将深度差值加权融合到锚点框响应得分中。锚点框响应得分分布在 0 到 1 之间,而深度差值数值变化较大,所以首先对深度差值 $|D_i - D_p|$ 进行归一化处理,将深度差值数值变换至 0 到 1 之间,将其作为遮挡判别的响应得分 S_2。将 RPN 模块输出的锚点框响应得分 S_1 与遮挡判别响应得分 S_2 进行加权融合,计算式如下:

$$S = S_1 \cdot \lambda + S_2 \cdot (1-\lambda)。 \tag{7}$$

式中:λ 为权重系数。

2　实验结果与分析

本文算法基于 Python3.6 实现,硬件实验环境为 Intel Core i7-6700K CPU,主频 4 GHz,内存 8 GB,显卡 GeForce GTX 1060 配置的计算机。

为验证算法的有效性,本文采用 OTB-2015[10] 中部分存在目标遮挡的视频作为测试数据集,视频序列名称分别为 Singer1、Walking、Walking2、Skating2-2、Bolt、David3、Girl2、Woman、FaceOcc2、Jogging-1、Human5。OTB 数据集采用跟踪精确度和跟踪成功率两种评价指标。跟踪精确度计算了跟踪算法估计的目标位置中心点与标注的中心点之间的距离小于给定阈值的视频帧所占的百分比。跟踪成功率反映了算法估计的目标位置与标注位置之间的重合程度,当某一帧的重合程度大于设定的阈值时,则该帧被视为成功的,成功帧的总数占所有帧的百分比即为跟踪成功率。

本文 anchor 设置与 SiamRPN 算法一致,为 0.33、0.5、1、2、3 共 5 种长宽比。式(7)中权重参数 λ 的取值非常重要,需要通过实验确定,λ 太大或太小都会引起跟踪漂移。本文赋予基准跟踪器响应得分更大的权重,$\lambda \in [0.7, 1]$。表 1 为 λ 取不同值时的跟踪成功率和跟踪精确度。根据表 1 中性能对比,本文的融合权值 λ 取 0.85。得到修正的响应得分 S 后,求出 S 中的最大值和对应的锚点框位置,将此锚点进行修正得到最终的坐标。

测试跟踪器包括 SiamRPN[5]、SRDCF[13]、Staple[14]、CFNet[15]、SiamFC[4]、fDSST[16]。表 2 中展示了 7 种算法在 11 个视频序列上的跟踪精确度,其中 10 个视频序列中本文算法的精确度均不小于基准跟踪器 SiamRPN。SiamRPN 作为孪生网络跟踪器,只学习了离线的通用特征,在跟踪目

标被遮挡时,跟踪器判别能力不足,无法区分跟踪目标与遮挡物。本文提出的遮挡判别模型能有效地利用图像深度信息提升跟踪算法在目标被遮挡时的跟踪性能。

图3展示了跟踪算法在不同属性视频序列下的跟踪精确度对比,在这些属性的视频序列下,本文算法的跟踪精确度均高于基准跟踪器SiamRPN,在4种属性的视频序列中,平均跟踪精确度分别

表1 λ 取不同值时的跟踪结果

Table 1 The tracking results with different λ values

λ	跟踪成功率	跟踪精确度
0.70	0.590	0.805
0.75	0.610	0.830
0.80	0.612	0.831
0.85	0.623	0.853
0.90	0.620	0.847
0.95	0.620	0.845
1.00	0.612	0.845

表2 不同算法在11个视频序列上的跟踪精确度

Table 2 Accuracy on 11 video sequences using different algorithms

视频序列	本文算法	SiamRPN	SiamFC	CFNet	fDSST	SRDCF	Staple
Bolt	0.804	0.795	0.043	0.020	0.898	0.020	**0.917**
David3	0.900	0.899	0.576	0.876	0.554	**0.919**	**0.919**
FaceOcc2	0.815	0.813	0.776	0.821	**0.858**	0.814	0.834
Girl2	**0.642**	0.629	0.405	0.342	0.094	0.069	0.096
Human5	0.912	**0.922**	0.692	0.818	0.232	0.913	0.899
Jogging-1	0.851	0.847	0.855	0.820	0.220	**0.894**	0.216
Singer1	0.864	0.864	0.900	0.877	0.921	**0.937**	0.926
Skating2-2	**0.534**	0.528	0.417	0.218	0.069	0.484	0.110
Walking	0.953	0.943	0.937	0.945	**0.957**	0.951	0.950
Walking2	0.378	0.375	0.908	0.892	0.928	**0.933**	0.931
Woman	0.921	0.899	0.821	0.840	0.877	0.897	**0.938**

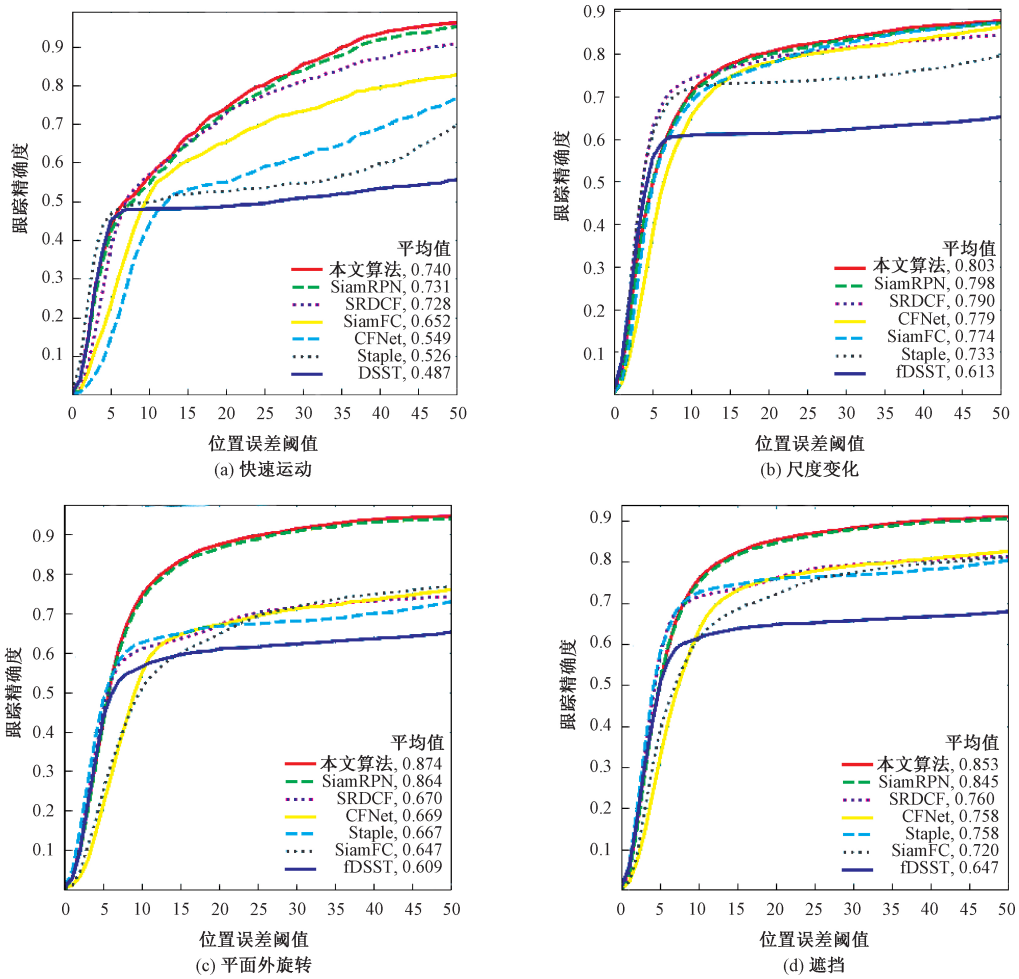


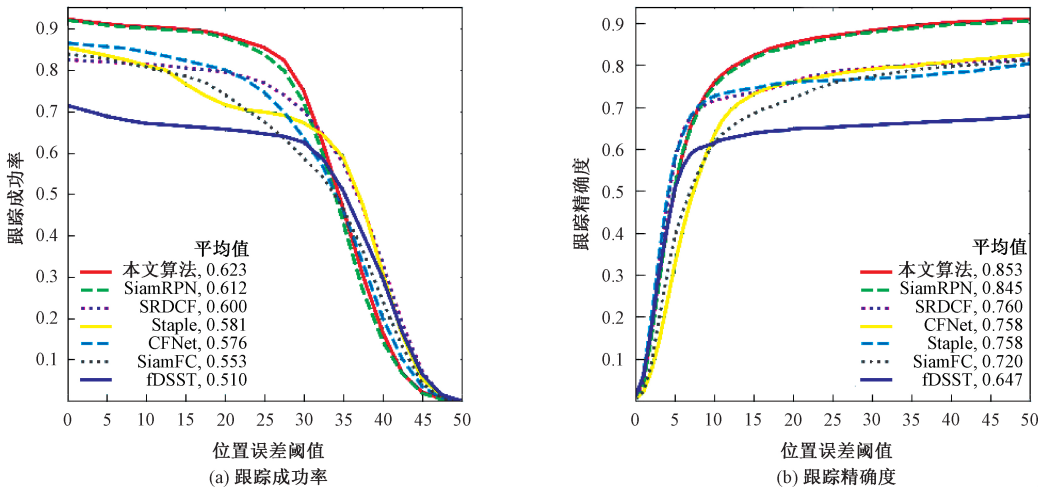
图3 部分属性视频序列下的精确度曲线对比图

Figure 3 Comparison of accuracy rates of video sequences with different attributes

提升了 0.009、0.005、0.01、0.008,图像深度信息在目标跟踪中起到了辅助作用。

图 4 为跟踪器在 11 个视频序列下的成功率和精确度曲线。在平均成功率上,本文算法(0.623)在 SiamRPN 算法(0.612)的基础上提升了 0.011;在平均精确度上,本文算法(0.853)在 SiamRPN 算法(0.845)的基础上提升了 0.008,分别提高了 1.8%和 0.9%。

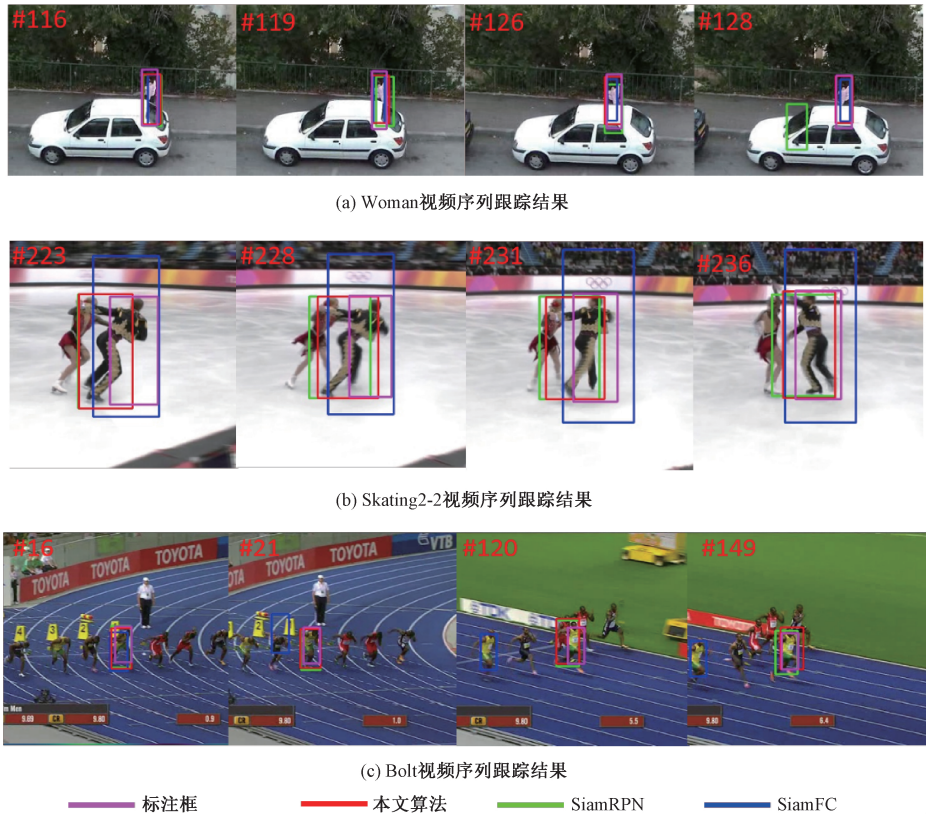
图 5 记录了 SiamRPN、SiamFC 和本文算法在 Woman、Skating2-2 和 Bolt 这 3 个视频序列下的实际跟踪效果对比。紫色框为目标的标注框,红色框为本文算法结果,绿色框为 SiamRPN 算法结果,蓝色框为 SiamFC 算法结果。从图 5 中可以看出,本文算法在部分遮挡时的目标跟踪上取得了不错的效果。当目标被遮挡时,SiamRPN 和 SiamFC 均会出现跟踪漂移的现象,而本文算法能有效缓解或避免此问题。



(a) 跟踪成功率　　(b) 跟踪精确度

图 4　跟踪器在 11 个视频序列上性能对比

Figure 4　Tracker performance comparison on 11 video sequences



(a) Woman视频序列跟踪结果

(b) Skating2-2视频序列跟踪结果

(c) Bolt视频序列跟踪结果

图 5　在部分视频序列下的算法实际效果对比

Figure 5　Comparison of actual effects of algorithms in video sequences

3 结论

为提高目标跟踪算法在目标遮挡场景下的适应性,提出将孪生网络跟踪器 SiamRPN 与单目图像深度估计算法 Monodepth2 结合。本文提出的基于深度信息的遮挡判别模型判断出目标是否被遮挡,有效地避免了跟踪漂移。若出现遮挡,算法会将原跟踪器的锚点响应得分与遮挡判别响应得分进行加权融合得到最终响应得分,重新选择锚点计算目标框的位置。针对 OTB-2015 中具有遮挡属性的 11 个视频序列进行测试。实验结果表明,在目标遮挡场景下,与 6 种主流跟踪算法相比,本文算法具备更优的跟踪性能。同时也说明,图像深度信息可以辅助提升目标跟踪的性能。下一步研究工作可针对遮挡判别策略进行改进或采用性能更优异的图像深度估计算法,进一步提高算法的跟踪性能,也可以尝试将此遮挡判别模块应用于其他跟踪算法中。

参考文献:

[1] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]// European Conference on Computer Vision. Berlin:Springer, 2012: 702-715.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(3): 583-596.

[3] TAO R, GAVVES E, SMEULDERS A W. Siamese instance search for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1420-1429.

[4] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 850-865.

[5] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8971-80.

[6] ZHU Z, WU W, ZOU W, et al. End-to-end flow correlation tracking with spatial-temporal attention[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 548-557.

[7] WU C L, ZHANG Y, ZHANG Y, et al. Motion guided siamese trackers for visual tracking [J]. IEEE access, 2020, 8:7473-7489.

[8] MUNARO M, BASSO F, MENEGATTI E. OpenPTrack: open source multi-camera calibration and people tracking for RGB-D camera networks [J]. Robotics & autonomous systems, 2016, 75:525-538.

[9] GODARD C, MAC AODHA O, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE 2019:3828-3838.

[10] WU Y, LIM J, YANG M H. Object tracking benchmark [J]. IEEE analysis and machine intelligence, 2015, 37(9): 1834-1848.

[11] 毛晓波, 周晓东, 刘艳红. 基于 FAST 特征点改进的 TLD 目标跟踪算法 [J]. 郑州大学学报(工学版), 2018, 39(2): 1-5,17.

[12] 刘明华, 汪传生, 胡强, 等. 多模型协作的分块目标跟踪 [J]. 软件学报, 2020, 31(2): 511-530.

[13] DANELLJAN M, HAGER G, SHAHBAZ KHAN F S, et al. Learning spatially regularized correlation filters for visual tracking[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE,2015: 4310-4318.

[14] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: complementary learners for real-time tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1401-1409.

[15] VALMADRE J, BERTINETTO L, HENRIQUES J F, et al. End-to-end representation learning for correlation filter based tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway: IEEE, 2017: 2805-2813.

[16] DANELLJAN M, HAGER G, KHAN F S, et al. Discriminative scale space tracking [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(8): 1561-1575.

(下转第 31 页)

each channel to establish the interdependence between the channels. Finally, the channel is weighted to the original in terms of features to complete the reconstruction of the original features. In addition, different sizes of downsamplers are used to capture and fuse feature information of different scales to enhance the detailed feature information of microfibers, and to improve the model's learning ability and recognition effect of microfibers. The improved MobileNetV2 network's microfiber recognition accuracy rate reaches 97.96%. Compared with the original MobileNetV2 network, the recognition accuracy rate is increased by 2.54%. At the same time, the false recognition rate and the missed recognition rate are also significantly reduced. In comparison to ResNet, DenseNet, VGG16 and NasNet networks, the model size is compressed several times, the accuracy of microfiber recognition is improved, and the false recognition rate and missed recognition rate are greatly reduced. Experimental results show that the network model can extract more complete feature information for microfiber. While strengthening the microfiber feature to identify the directivity, the model is reduced, and the difficulty of deployment in mobile devices is reduced as well. The improved model recognizes microfibers with higher accuracy and better stability.

Key words: waterbody; microfiber recognition; MobileNetV2; pooling fusion; feature reconstruction

(上接第 24 页)

Anti-occlusion Target Tracking Algorithm Based on Image Depth

WANG Xipeng, LI Yong, LI Zhi, ZHANG Yan

(School of Information Engineering, Engineering University of People's Armed Police, Xi'an 710086, China)

Abstract: Due to the limitation of video information, target tracking in the case of occlusion is still a difficult problem to solve. Aiming at the problem of occlusion in the target tracking process, it is proposed to introduce image depth into single target tracking algorithm. Firstly, the monocular image depth estimation algorithm is used to estimate the depth of the image to obtain the depth information of the image. Secondly, the target tracking algorithm based on the siamese region proposal network is combined with the image depth to construct an occlusion discriminating module, which uses the change of the target depth information to determine the occlusion. Finally, the occlusion discrimination score and the anchor response score are weighted integrated. According to the final response score, the anchor of the target tracker is reordered to avoid interference by obstructions. Experimental results on the OTB-2015 dataset show that the algorithm can effectively deal with the influence of occlusion on tracking performance, with an average success rate of 0.623 and an average tracking accuracy of 0.853, which is 1.7% and 0.9% higher than the benchmark algorithm, repectively.

Key words: siamese network; deep learning; target tracking; monocular depth estimation; anti-occlusion