基于特征融合的 RGBT 双模态孪生跟踪网络

申亚丽

(运城学院 数学与信息技术学院，山西 运城 044000)

摘　要：热红外成像技术被广泛地应用于军事、遥感和安防等领域中的目标跟踪，但热红外图像对对比度较低、目标模糊等跟踪场景效果一般。因此，将热红外图像与可见光图像进行融合提高跟踪性能具有重要意义。与基于可见光或热红外图像的单模态跟踪算法相比，基于可见光/热红外(RGB/Thermal, RGBT) 图像的双模态跟踪算法对光照变化、云雾遮挡具有更强的鲁棒性。提出了一种基于特征融合的 RGBT 双模态孪生跟踪网络架构。该网络将双模态图像中提取的深度特征进行融合，提高目标外观特征的判别力。该网络可以利用训练数据进行端到端的离线训练。公开数据集 RGBT234 上的实验结果表明，所提出的 RGBT 双模态孪生特征融合跟踪网络能够实现复杂场景下鲁棒持续的目标跟踪。

关键词：可见光/热红外；　双模态跟踪；　孪生网络；　特征融合

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.3788/IRLA20200459

RGBT dual-modal Siamese tracking network with feature fusion

Shen Yali

(School of Mathematics and Information Technology, Yuncheng University, Yuncheng 044000, China)

Abstract: Infrared imaging technology has been widely used for object tracking in military, remote sensing, security and other fields. However, thermal infrared images generally suffer from low contrast and blurry targets. Therefore, it has great importance of fusing infrared images with visible images. Compared with single-modal RGB trackers, dual-modal RGBT(RGB/Thermal infrared) trackers are more robust to illumination variation and fog. In this paper, a RGBT dual-modal siamese tracking network with feature fusion was proposed. Convolutional features extracted from the visible image and infrared image were fused to improve the appearance feature discrimination. The network can use the training data for end-to-end off-line training. Experimental results on the public RGBT234 dataset demonstrate that our tracker achieves robust and persistent tracking in complex scenarios.

Key words: RGB/Thermal infrared;　dual-modal tracking;　Siamese network;　feature fusing

收稿日期:2020−11−28;　**修订日期:**2020−12−10

基金项目:山西省高等院校科技创新项目 (2019L0868); 山西省教育科学‘十三五’规划 2020 年度互联网+教育研究专项课题 (HLW-20096)

0 引 言

红外传感器利用目标物与背景的温度差异从而获得热红外图像[1−2]。和传统的可见光图像相比，红外成像技术能够在夜间和恶劣天气环境下工作且隐秘性和保密性较强，对云、雾、烟和伪装网具有一定的穿透性。基于红外图像的以上优势，红外传感器被广泛应用于卫星遥感、导引头制导、消防、红外夜视、资源探测和安防等领域。

文中关注计算机视觉领域的短时单目标跟踪任务[3−4]。一般来讲，视觉跟踪研究者往往利用可见光图像序列对图像中的感兴趣目标进行跟踪。近年来许多研究者利用热红外图像进行目标跟踪[2]。然而，传统的红外传感器在目标跟踪中一直存在探测失效或误判概率较高的问题。当目标和背景温度相似时，往往会发生热交叉效应，这使得跟踪算法难以从背景中发现前景目标。不同于红外图像，可见光图像具有较高的分辨率，可以更好地区分目标和背景的局部细节。由于可见光图像和红外图像具有很强的互补性，双模态目标跟踪越来越受到研究者们的重视和欢迎[5−9]。

近年来，孪生网络在视觉跟踪领域受到了广泛的关注，基于孪生网络的目标跟踪算法被不断提出[10−15]。基于孪生网络的跟踪算法是通过图像相似性度量的方法来进行目标跟踪，其基本思想是学习一个相似性函数计算样本图像和候选图像之间的相似性，如果两幅图像描述相同的目标则得到高分，否则为低分。具体来说，一般将初始帧图像中的目标区域作为样本图像，为了找到目标新的位置，将详细地测试候选图像中所有可能的位置，并选择与样本图像具有最大相似性的位置作为目标新的位置。在实际跟踪场景中，基于可见光图像和孪生网络的视觉跟踪算法会受到光照变化、局部遮挡、烟雾等因素的影响而发生漂移，从而导致跟踪失败，如图 1 所示。

为了解决双模态目标跟踪的以上难点，文中对可见光图像和热红外图像进行像素级和特征级融合并提出了一种可见光-热红外 (RGB-Thermal, RGBT) 双模态孪生跟踪网络。公开数据集上的实验证明了可见光图像和热红外图像融合对于跟踪任务的有效性。

图 1 可见光图像 (上) 和红外图像 (下)

Fig.1 Visible image (up) and infrared image (down)

1 基准算法 Siamfc

近年来，由于卷积神经网络具有强大的目标表征能力，在计算机视觉各个任务中都取得了惊人的性能，所以一般使用深度卷积网络来学习相似性函数 f。其中，孪生网络 (Siamese) 是使用深度网络进行相似性学习的最佳选择，其结构如图 2 所示。孪生网络对两个输入使用相同的变换φ，然后使用另一个函数 g组合它们的表示来计算相似性$f(z,x)=g(\varphi(z),\varphi(x))$，其中函数$g$是一个简单的距离或者相似性度量，$\varphi$是一个映射。

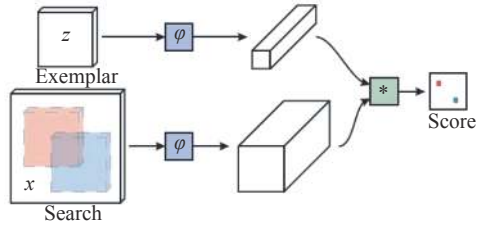


图 2 孪生网络的基本框架图

Fig.2 Flowchart of the Siamese network

Siamfc 是基于孪生网络跟踪器的典型代表之一，它选择全卷积网络作为基网络。与其他孪生网络跟踪器不同的是，Siamfc 跟踪器使用了一个互相关层 (cross-correlation layer) 来组合两个输入的特征图，以计算搜索图像与目标图像之间的相关性：

$$f(z,x)=\varphi(z)*\varphi(x)+b \tag{1}$$

式中：b是一个偏置量；$*$表示循环相关运算。需要注意的是，Siamfc 跟踪器的网络输出是一张分数图，而不是一个单一的分数值。除此之外，大多数的孪生网络跟踪器都采用多尺度搜索 (通常选择 3 个尺度) 的策略来完成目标尺度的估计，但是这种简单的策略不仅会影响算法的实时性，而且不能适应目标的长宽比变化。

基于孪生网络的跟踪算法使用多个正负图像以进行网络的训练，并采用了 logistic 损失函数：

$$l(y,v)=\log(1+\exp(-yv)) \tag{2}$$

式中：v表示单个样本对的输出值；$y\in\{+1,-1\}$为标签。公式 (2) 表示分数图上单个点的损失值，所以整张分数图的损失值采用的是全部点损失值的平均值，即：

$$L(y,v)=\frac{1}{|D|}\sum_{u\in D}l(y[u],v[u]) \tag{3}$$

然后，通过对下面的问题应用随机梯度下降 (SGD) 来计算网络的参数值θ：

$$\arg\min_{\theta}\ \underset{(z,x,y)}{E}\ L(y,f(z,x;\theta)) \tag{4}$$

训练的样本图像对是从标注视频数据集中获得的，它们是通过提取以目标为中心的示例图像和搜索图像来组成图像对。图像是从同一段视频中提取出来的，这两帧图像都包含对象并且最多相隔 T 帧 (T 通常设为 100)。

2 文中算法

该节提出了一种基于 Siamese 网络的端到端训练双模态 (Dual-modal) 跟踪网络，能够同时学习可见光图像和热红外图像深度特征，并对深度特征进行堆叠，从而达到可见光和红外双模态融合的效果，通过目标的融合特征表示进行自适应和鲁棒跟踪。

DMSiam 网络架构包括样本分支和搜索分支，每个分支又分为可见光子分支和红外子分支，如表 1 和图 3。子分支是一个特征提取的主干网络，由五个卷积层和两个最大池化层组成，网络参数的详细维度和不同层的输出如表 1。可见光图像和红外图像上提取的卷积特征通过特征堆叠得到融合，然后在网络后端 (在 conv5 层之后) 通过互相关直接比较样本分支和搜索分支的高级语义特征，以进行鲁棒跟踪。由于搜索分支与样本分支具有相同的大小，因此这里只提供样本分支中的参数维度。DMSiam 网络的前两个卷积层采用最大池化，在其之后立即插入批规范化，卷积层激活函数的采用线性整流单元。

为了提高跟踪的运算效率，在 DMSiam 前端网络插入可微相关滤波器层。相关滤波器层在频域中实现，可以进行端到端训练，提高了计算效率，并能在线更新适应目标的变化。

相关滤波层利用目标周围密集提取的样本去有效地学习一个相关滤波器，主要通过对搜索窗口的循环

表 1 网络参数维度

Tab.1 Dimensions of network parameters

Layer	Kernel size	Channel×Map	Stride	Size	Channel
Input	11×11			255×255	3
Conv1	3×3	16×3	2	123×123	16
Pool1	5×5		2	61×61	16
Conv2	3×3	32×16	1	57×57	32
Pool2	3×3		1	55×55	32
Conv3	3×3	64×32	1	53×53	64
Conv4	3×3	128×64	1	51×51	128
Conv5	3×3	32×128	1	49×49	32

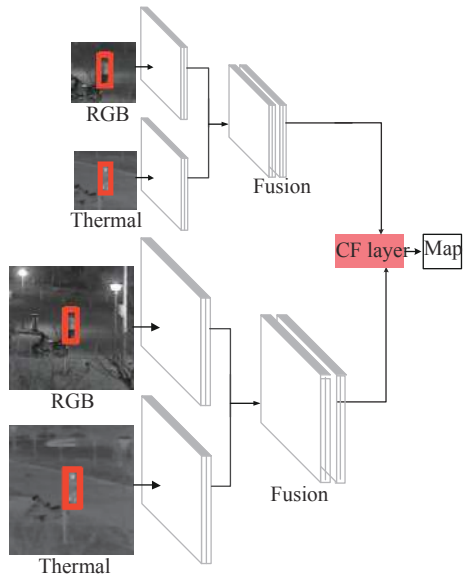


图 3 可见光图像 (上) 和红外图像 (下)

Fig.3 Visible image (up) and infrared image (down)

移位得到目标平移的所有可能性。给定一个图像标量值和相应的高斯标签，通过对所有的循环移位和对应标签进行岭回归得到相关滤波器模板，具体公式如下：

$$\arg\min_{w}\frac{1}{2n}\|w*x-y\|^2+\frac{\lambda}{2}\|w\|^2 \tag{5}$$

式中：n是样本的有效数目；*表示循环相关运算。利用循环矩阵在傅里叶域的性质，求解公式 (5) 得到相关滤波器的解为：

$$\begin{cases}\hat{k}=\dfrac{1}{n}(\hat{x}^*\cdot\hat{x})+\lambda_1\\ \hat{\alpha}=\dfrac{1}{n}\hat{k}^{-1}\cdot\hat{y}\\ \hat{w}=\hat{\alpha}^*\cdot\hat{x}\end{cases} \tag{6}$$

式中：$\hat{x}$表示变量x的傅里叶变换；$\hat{x}^*$表示$\hat{x}$的复共轭，乘积和除法是点运算。

将相关滤波器集成为 DMSiam 网络中的一个可微层，通过反向传播算法进行端到端训练。给定图像样本和相应的标签，可以求出相关滤波器的系数，完成网

络的正向传播。给定输出标量损失l和l对w上的偏导数$\nabla_w l$, 从$\nabla_w l$得到$\nabla_x l$和$\nabla_y l$的反向传播推导如公式 (7):

$$\begin{cases} \nabla_\alpha \hat{l} = \hat{x} \cdot (\nabla_w \hat{l})^* \\ \nabla_y \hat{l} = \dfrac{1}{n} \hat{k}^{-*} \cdot \nabla_\alpha \hat{l} \\ \nabla_k \hat{l} = -\hat{k}^{-*} \cdot \hat{\alpha}^* \cdot \nabla_\alpha \hat{l} \\ \nabla_x \hat{l} = \hat{\alpha} \cdot \nabla_w \hat{l} + \dfrac{2}{n} \hat{x} \cdot \mathrm{Re}\{\nabla_k \hat{l}\} \end{cases} \tag{7}$$

在推导了相关滤波器层的正向和反向传播后, 构造了损失函数L如下:

$$L_{low} = \frac{1}{|D|} \sum_{u \in D} \log(1 + \exp(-y[u] \cdot v[u])) \tag{8}$$

式中: D表示低层次特征映射图; $|D|$表示特征映射图像素的个数; v表示特征响应映射的计算值; y表示真实响应映射的标记值。

3 实 验

文中在公开数据集 RGBT234 上对提出的算法进行了验证并证明了算法的有效性。算法是基于 MatConvNet[16] 利用 Matlab 实现的, 硬件平台是 NVIDIA GeForce GTX Titan GPU 和 Intel Core i7-6700K。

3.1 评价指标

(1) 距离精确率和精确率图

平均中心位置误差 (Average Center Location Error) 是一种广泛使用的跟踪精度评价指标, 其定义为在整个视频序列中被跟踪目标的中心位置与 groundtruth 之间的平均欧式距离。然而当跟踪器丢失目标时, 由于输出位置是随机的, 所以该指标可能无法正确地评估跟踪器的性能。目前, 距离精确率 (DPR) 和精确率图 (Precision Plot) 通常被用来评估跟踪器的整体性能。距离精确率表示整个视频序列内中心误差小于给定阈值距离d的帧数占序列总帧数的百分比, 通常d取值为 20。精确率图是根据不同距离阈值下距离精确率的变化绘制出的曲线图, 可以更加全面地反映跟踪器的性能。

(2) 重叠成功率和成功率图

边界框重叠率 (Overlap Rate) 是 OTB 数据集上另一种评价目标跟踪器性能的指标。给定目标跟踪的边界框和真值, 其重叠率被定义为:

$$S = \frac{|Area(B_p \cap B_g)|}{|Area(B_p \cup B_g)|} \tag{9}$$

式中: B_p和B_g分别表示预测边界框和真值; $\cap$和$\cup$分别表示集合的交运算和并运算。重叠成功率 (OSR) 表示视频序列中重叠率大于阈值S的帧数占总帧数的百分率, 通常S取值为 0.5。成功率图是根据不同阈值下重叠成功率的变化而绘制出的曲线图, 其曲线下面积 (AUC) 通常作为跟踪算法性能排序的依据。

3.2 参数设置

在网络离线训练中, 输入的模板图像大小是 127×127×3, 搜索图像大小是 255×255×3。采用 GTOT[7] 数据集, GTOT 数据集包括 50 段人工标注的 RGBT 视频。DMSiam 网络在 GTOT 数据集上训练了 20 个 epoch, 训练数据共包含 900 对图像, 采用 SGD 梯度下降算法对网络进行离线训练。在线跟踪过程中, DMSiam 在三个尺度 $1.05^{\{01,0,1\}}$上对目标进行搜索并估计尺度。

3.3 性能对比

3.3.1 总体对比

将文中提出的 DMSiam 算法与 RGBT234 数据集上的 state-of-the-art 算法进行对比。所有对比算法的 Precision Plots 和 Success Plots 如图 4 所示。实验结果表明, 和基准算法 CFNet+RGBT 相比, 文中提出的 DMSiam 算法可以有效地实现稳定跟踪。

3.3.2 各属性性能对比

图 5 和 6 分别给出了对比算法在各种视频属性

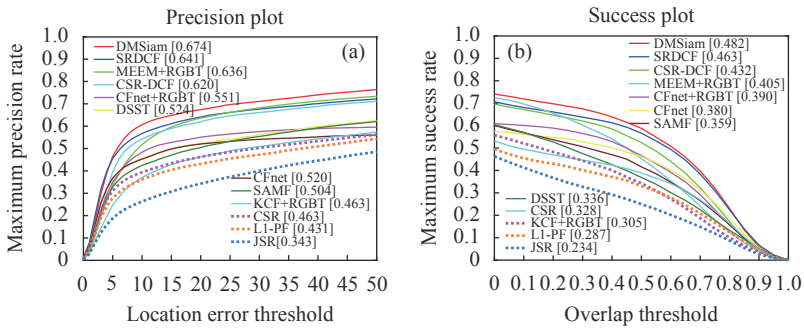


图 4 精确率图 (a) 和成功率图 (b)

Fig.4 Precision plot (a) and success plot (b)

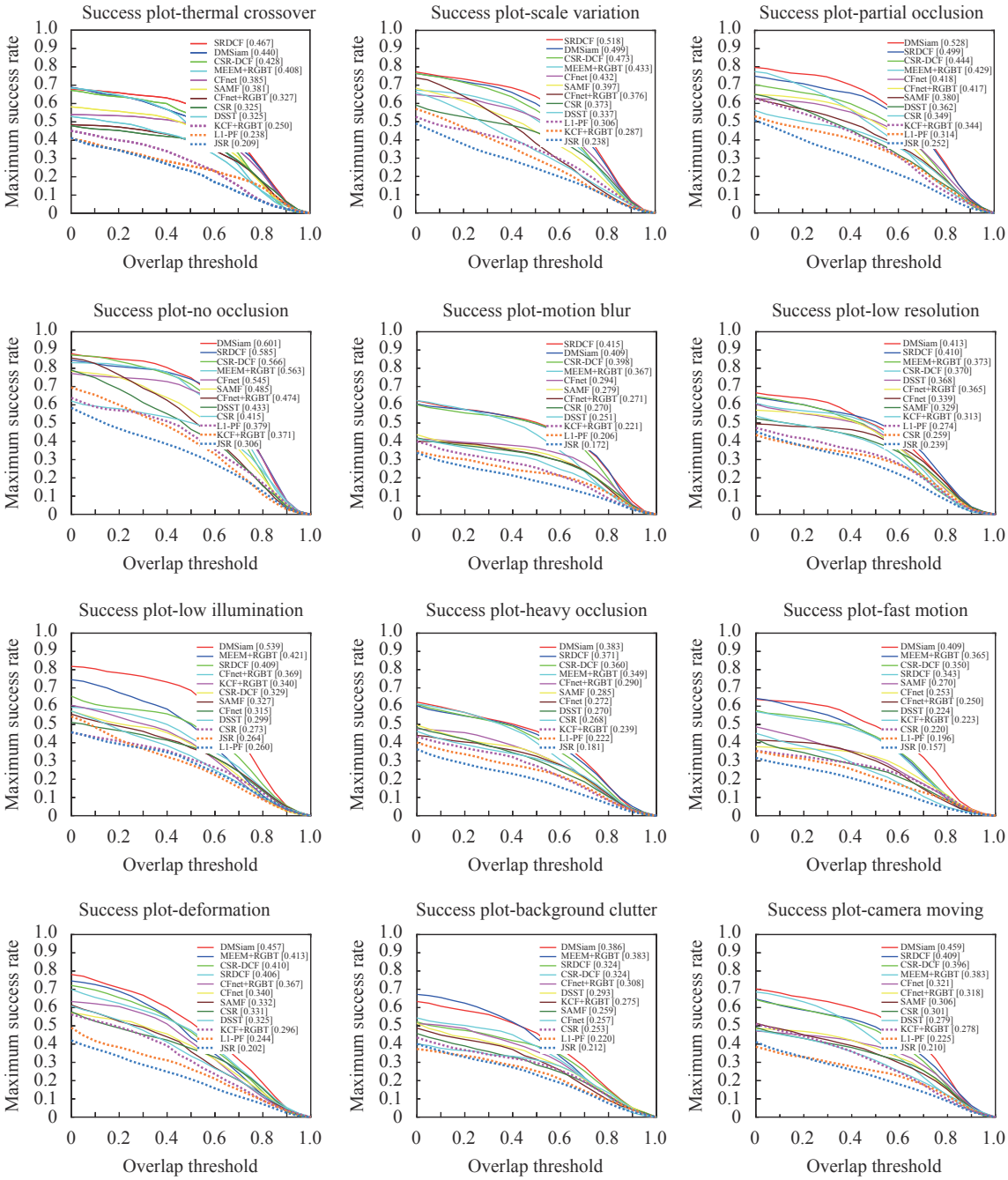
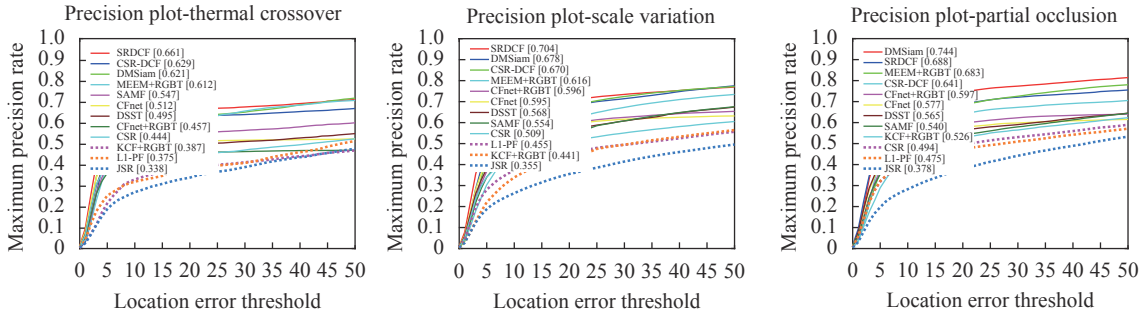


图 5 对比算法在 12 种属性下的成功率图

Fig.5 Success plot under 12 different attributes



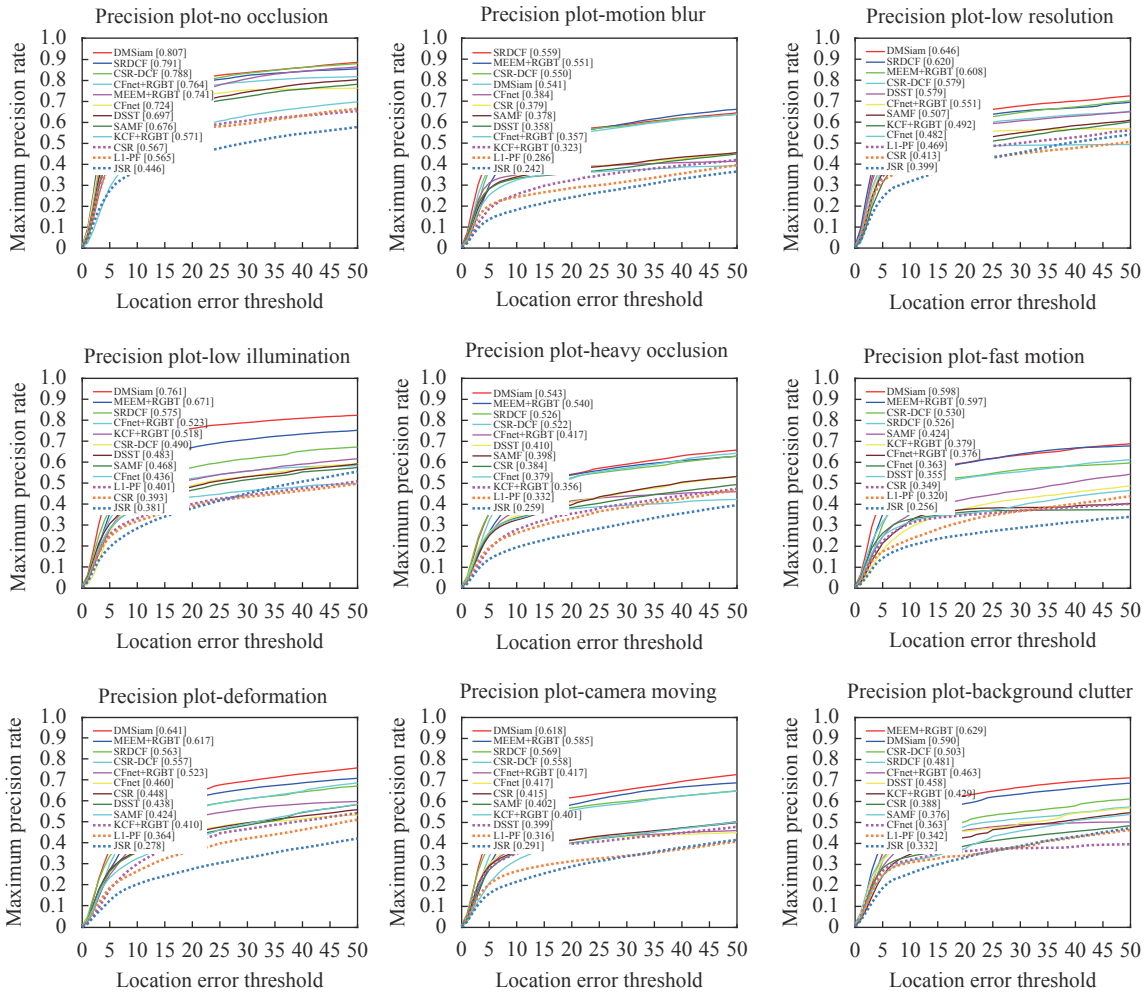


图 6 对比算法在 12 种属性下的精确率图

Fig.6 Precision plot under 12 different attributes

(无遮挡、局部遮挡、严重遮挡、低光照、低分辨率、热交叉、变形、快速运动、尺度变化、运动模糊、相机运动和背景干扰)下的最大成功率图和最大准确率图。文中算法 DMSiam 在 12 种属性下都取得了较好的跟踪性能。

4 结 论

针对可见光图像和热红外图像在视觉跟踪任务上的互补优势,文中利用特征融合提出了可见光-热红外双模态孪生跟踪网络模型。该网络首先将 RGBT 双模态图像中提取的深度特征进行堆叠从而实现特征融合,然后对网络模板分支和搜索分支上的融合特征输入相关滤波层实现快速的目标跟踪。文中提出网络对光照变化、云雾遮挡具有较强的鲁棒性,并且可以利用训练数据进行端到端的离线训练。实验表明,和基准算法 CFNet+RGBT 相比,文中提出双模态视觉跟踪网络在复杂跟踪场景中能够实现鲁棒跟踪,并具有一定的性能提升。

参考文献:

[1] Chen X J, Yang Y M. Realization of dual-band fire detector based on infrared video [J]. ***Journal of Electronic Measurement and Instrumentation***, 2016, 33(3): 473-479.

[2] Li C L, Liang X Y, Lu Y J, et al. Rgb-t object tracking: benchmark and baseline [J]. ***Pattern Recognition***, 2019, 96: 106977.

[3] Guan H, Xue X Y, An Z Y. Online single object video tracking: A survey [J]. ***Mini-Micro Systems***, 2017, 38(1): 147-153.

[4] Yilmaz A, Javed O, Shah M. Object tracking: A survey [J]. ***ACM Computing Surveys***, 2006, 38(4): 1-45.

[5] Wu Y, Blasch E, Chen G S, et al. Multiple source data fusion via sparse representation for robust visual tracking[C]//International

Conference on Information Fusion, 2011.

[6] Sun F, Liu H. Fusion tracking in color and infrared images using joint sparse representation [J]. ***Science China Information Sciences***, 2012, 55(3): 590-599.

[7] Li C, Cheng H, Hu S, et al. Learning collaborative sparse representation for grayscale-thermal tracking [J]. ***IEEE Transactions on Image Processing***, 2016, 25(12): 5743-5756.

[8] Li C, Nan Z. Lu Y, et al. Weighted sparse representation regularized graph learning for rgb-t object tracking[C]//ACM on Multimedia Conference, 2017.

[9] Li C, Wu X, Zhao N, et al. Fusing two-stream convolutional neural networks for rgb-t object tracking [J]. ***Neurocomputing***, 2018, 28(1): 78-85.

[10] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1420-1429.

[11] Held D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks[C]//European Conference on Computer Vision, 2015, 15(12): 625-637.

[12] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C]//IEEE Conference on Computer Vision, 2015: 3119-3127.

[13] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4057-4068.

[14] Wang Q, Gao J, Xing J, et al. DCFNet: Discriminant correlation filters network for visual tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3027-3038.

[15] Xiong Y J, Zhang H T, Deng X. RGBT dual-modal tracking with weighted discriminative correlation filters [J]. ***Journal of Signal Processing***, 2020, 36(9): 1590-1597. (in Chinese)

[16] Vedaldi A, Lenc K. Matconvnet: convolutional neural networks for matlab[C]//Association for Computing Machinery, 2015: 689-692.

第一作者简介： 申亚丽 (1979−), 女, 副教授, 博士, 主要从事计算机软件与理论, 应用数学方面的研究。近年来主持省科技创新项目 1 项、省十三五规划课题 1 项; 参与国家自然科学基金项目 3 项、省高等学校教学改革项目 2 项等。 在国际国内期刊上共发表学术论文 10 余篇, 其中 SCI 收录 5 篇。 Email: 272944179@qq.com