

基于高秩特征和位置注意力的RGBT目标跟踪

杨岚岚, 王文琪, 王福田

引用本文

杨岚岚, 王文琪, 王福田. [基于高秩特征和位置注意力的RGBT目标跟踪](#)[J]. 计算机科学, 2022, 49(12): 236-243.

YANG Lan-lan, WANG Wen-qi, WANG Fu-tian. [RGBT Object Tracking Based on High Rank Feature and Position Attention](#) [J]. Computer Science, 2022, 49(12): 236-243.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[自适应的集成定序算法](#)

Adaptive Ensemble Ordering Algorithm

计算机科学, 2022, 49(6A): 242-246. <https://doi.org/10.11896/jsjcx.210200108>

基于高秩特征和位置注意力的 RGBT 目标跟踪

杨岚岚 王文琪 王福田

安徽大学计算机科学与技术学院 合肥 230000

(namyeung@foxmail.com)

摘要 RGBT 目标跟踪利用可见光(RGB)与热红外(T)两种不同模态的优势来解决单一模态目标跟踪中常见的模态受限问题,以此提升复杂环境下的目标跟踪性能。在 RGBT 目标跟踪算法中,精准定位目标位置和有效融合两种模态都是非常重要的问题。为了达到精准定位目标以及有效融合两种模态的目的,提出了一种探索高秩的特征图以及引入位置注意力来进行 RGBT 目标跟踪的新方法。该方法首先根据主干网络的深层与浅层的特征,使用位置注意力来关注目标的位置信息,接着通过探索两种模态融合前的高秩特征图,关注特征的重要性,以指导模态特征融合。为了关注目标位置信息,在行和列上使用平均池化操作。对于高秩特征指导模块,文中根据特征图的秩来指导特征图的融合。并且,为了去除冗余和噪声,实现更加鲁棒的特征表达,直接删除了秩小的特征图。在两个 RGBT 跟踪基准数据集上的实验结果表明,与其他 RGBT 目标跟踪方法相比,所提方法在准确度和成功率上取得了更好的跟踪结果。

关键词: RGBT 目标跟踪;高秩特征图;目标位置信息

中图法分类号 TP18

RGBT Object Tracking Based on High Rank Feature and Position Attention

YANG Lan-lan, WANG Wen-qi and WANG Fu-tian

School of Computer Science and Technology, Anhui University, Hefei 230000, China

Abstract RGBT target tracking uses the advantages of two different modes of visible light(RGB) and thermal infrared(T) to solve the common modal limitation problem in single mode target tracking, so as to improve the performance of target tracking in complex environment. In the RGBT object tracking algorithm, the precise location of the object and the effective fusion of the two modalities are very important issues. In order to accurately locate the object and effectively fuse the two modalities, this paper proposes a new method to explore high-rank feature maps and introduce position attention for RGBT object tracking. The method first uses location attention to focus on the location information of the object according to the deep and shallow features of the backbone network, and then focuses on the importance of the features by exploring the high-rank feature maps before the fusion of the two modalities to guide the modal features fusion. In order to focus on the object location information, this paper uses the average pooling operation on the rows and columns. For the high-rank feature guidance module, this paper guides the fusion of feature maps according to the rank of the feature maps. In order to remove redundancy and noise and achieve more robust feature expression, the feature graph with small rank is deleted directly. Experimental results on two RGBT tracking benchmark data sets show that compared with other RGBT target tracking methods, the proposed method achieves better tracking results in accuracy and success rate.

Keywords RGBT object tracking, High rank feature, Object location information

视觉目标跟踪作为计算机视觉中的一个重要研究方向,在许多领域有着广泛的应用,如军事领域、智能交通和视频监控领域等。近年来,视觉目标跟踪取得了巨大的进步,尤其是 RGB 目标跟踪^[1-3]。许多研究人员主要关注 RGB 跟踪,因为有大量可用数据集^[4-5]和大量基准算法^[6-8]。然而,在恶劣环境条件下(如低光照、下雨或雾霾),RGB 目标跟踪的性能不尽如人意。热红外传感器在这些环境下可以提供更稳定的信号,并且随着热感红外相机价格的下降,热红外相机在计算机视觉任务中的应用也得到广泛推广。如显著性检测^[9]、行人

重新识别^[10]、对象分割^[11-12]和行人检测^[13]等。热红外相机基于物体表面的热辐射进行伪彩成像,它对光照变化、物体的伪装和姿态变化不敏感,并具有很强的穿透烟雾和霾的能力,因此热红外图像能够为可见光图像提供强有力的补充信息。但是与可见光图像相比,热红外图像具有高图像噪声、低分辨率^[14]、边缘和纹理信息不足等问题。因此,近年来 RGBT 目标跟踪受到了广泛的关注,并取得了巨大的进展^[15-23]。但是在 RGBT 目标跟踪中,仍然有很多方面值得探索。如有效融合两种模态和精准定位跟踪目标等。本文提出了一种新

方法,即基于高秩特征和位置注意力的 RGBT 目标跟踪算法,用于目标跟踪任务。本文使用秩排序来指导模态融合,原因是在已有的研究中我们发现,无论卷积神经网络(CNNs)接收到多少个图像批次,由单个滤波器生成的多个特征图的平均秩总是相同的^[24]。目前,有的融合方法是基于自适应重要性的,它一般需要嵌入到网络训练损失中,这将受到机器时间和繁重的人工劳动的限制。本文算法基于属性重要性,该方法不涉及训练损失,并且不需要引入附加辅助约束,从而简化了算法。具体来说,首先离线计算特征图的秩,并将其保留,因为无论 CNNs 接收到多少个图像批次,由单个滤波器生成的特征图平均秩总是相同的,并且特征图是一个中间步骤,反映了滤波器属性和输入图像^[24],即对于同一个滤波器来说,特征图的平均秩代表了滤波器提取特征的能力,因此,可以先离线计算特征图的秩,而不用在跟踪时对每一个视频序列都进行一次特征图秩的计算,离线计算平均秩对算法的速度基本没有影响。然后将特征图的秩作为权重,在融合时使网络可以侧重于秩大的特征图。最后为了减少噪声和冗余信息,将秩小于一定阈值的特征图删除。另外,目标跟踪的主要任务是找到要跟踪目标的位置和大小并将其框出。因此,准确定位目标的位置非常重要。受文献[25-27]启发,为了可以更好地关注目标的空间位置,本文设计了一种空间注意力模块。在空间注意力模块中,本文将全局平均池化改成对行和列单独的平均池化操作,这样可以更细粒度地关注到特征图行和列的信息,进而关注到目标位置。

1 相关工作

RGBT 跟踪作为视觉跟踪的一个研究分支,其原理是利用可见光信息和热红外信息的互补,在限定第一帧真值框的情况下,估计视频序列后续帧中目标框的位置和大小。目前 RGBT 跟踪算法取得了巨大的进步,例如,Li 等^[23]提出了一种新的挑战感知神经网络来处理模式共享的挑战(如快速运动、尺度变化和遮挡)和模式特定的挑战(如光照变化和热交叉),

该网络在每一层中设计了几个参数共享分支来模拟在模态共享挑战下的目标外观,以及在模态特定挑战下的几个参数独立分支。并且该网络提出了一个引导模块来将区别特征从一个模态转移到另一个模态,以此增强某些弱模态的区别能力。此外,该网络的所有分支以自适应的方式聚合在一起后,并行嵌入骨干网络,以有效地形成更有区别性的目标表示。Zhu 等^[18]提出一种密集特征聚合和剪枝的网络,该网络为了利用两种模态图像的特征的互补性,提出一种递归策略来密集地聚集这些特征,从而对每个模态都产生目标对象的健壮表示。对于不同的模态,该网络以协作的方式修建所有模态的密集聚集特征。Zhang 等^[20]提出一种基于模态感知注意网络和竞争学习的 RGBT 目标跟踪算法,该算法在特征提取时采用双流网络的形式提取每个模态的图像的特征。模态感知注意网络在原始数据上建立表示不同特征层重要性的注意模型,该网络还引导特征融合,以此来增强模态间的信息交互,然后在竞争学习的训练方式下,该算法从可见光模态和热红外模态最优融合的方向来微调网络。Zhu 等^[28]提出了一种质量感知特征聚合网络(FANet),用于鲁棒的 RGBT 跟踪。他们在每个模态中深层次聚合特征,以处理由低光照、变形、背景杂波和遮挡引发的外观显著变化的挑战。并且为了模拟可见光和热红外模态之间的相互作用,他们还精心设计了一个自适应聚合子网,根据不同模态的可靠性集成特征,从而缓解低质量源引入的噪声影响。Tang 等^[29]提出一种在线自适应决策级融合方法来挖掘隐藏的互补性,在孪生网络上扩展成双模态网络,并在决策层面上发展了一种可见光图像和热红外图像动态加权方法。

2 本文算法

本节主要介绍算法的整体网络架构、空间位置注意力模块、高秩特征指导模块、网络训练和网络跟踪等部分。本节的重点部分为位置注意力模块和秩指导模块。图 1 给出了 HRPANet(High Rank Feature and Position Attention for RGBT Object Tracking)的整体网络架构。

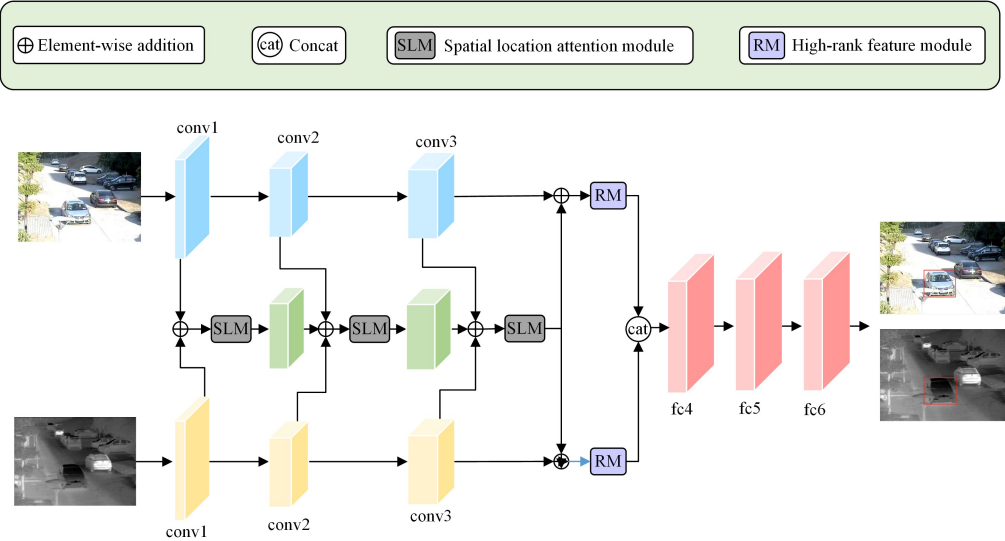


图 1 高秩特征和位置注意力网络结构图
Fig. 1 High rank feature and positional attention network structure

2.1 网络结构

如图 1 所示,本文提出的网络结构由特征提取模块、高秩特征指导模块、空间位置注意力模块和全连接层这 4 个部分组成。算法采用 VGG-M^[30] 作为主干网络,用于提取图像的特征信息。为了提取两个模态的特征,算法将主干网络的卷积层扩展成双模态。因此,主干提取网络的第一层有两个输入,分别是可见光图像和热红外图像。在图像输入网络中时将图像均裁剪为 107×107 的大小,并在每一层卷积后面加入空间位置注意力模块(SLM),最后在第三层卷积后面引入高秩指导模块(RM)来指导模态间特征的融合。将经过第一层卷积后得到的可见光特征和热红外特征直接进行相加,然后送入 SLM 模块。为了使经过第二层卷积得到的可见光特征和热红外特征能够与经过第一个 SLM 模块的特征进行相加操作(特征图要保证维度、大小一致才能相加),这里将经过第一个 SLM 模块的特征图进行卷积、池化、激活等操作。基于同样的目的,对经过第二个 SLM 模块的特征图进行同样的操作。然后将相加后的特征送入第三个 SLM 模块,并分别将经过第三层卷积获得的可见光特征和热红外特征与其相加。最后将相加后的可见光特征与热红外特征分别送入 RM 模块, RM 模块的输出是经过离线计算得到的特征图秩信息,并利用秩信息对两种模态做一个融合指导。

为了获得正负样本得分,将融合后的特征送入全连接层进行前景、背景的分类。在送入全连接层之前会对特征图的维度做一个变换,将送入全连接层的张量变成(32,9216)。这里有 3 个全连接层(fc4,fc5,fc6),fc4,fc5 只有一个分支,且保存了网络训练阶段时的参数。最后一层全连接层 fc6 在训练时有 K(视频序列数)个域,即 $fc6^1, fc6^2, \dots, fc6^K$ 。训练结束后 fc6 不保存训练时的参数,在跟踪时重新训练得到 fc6,此

时 fc6 只有一个分支。最后根据候选框的前景或背景的得分,确定跟踪目标包围框,从而实现目标跟踪任务。

2.2 空间位置注意力模块

下面将详细介绍空间位置注意力模块。如图 2 所示。在跟踪任务中,关注跟踪目标位置信息对跟踪性能有很大的影响。受文献[25-27]的启发,本文加入空间位置注意力模块,使网络可以关注到目标的空间位置信息,以提高网络的定位能力。首先采用逐像素加法操作,将来自不同分支的特征(可见光和热红外分支)相加。然后,为了关注行和列两个维度的目标位置,对相加后的特征进行两次全局平均池化。具体操作如式(1)和式(2)所示:

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \leq j < W} x_c(h, j)$$
 (1)

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w)$$
 (2)

其中, i 代表第行中第 i 个元素值, j 代表第 w 列中第 j 个元素。式(1)计算得到的结果是特征图上某一行特征的平均值,式(2)计算得到的结果是特征图上某一列特征的平均值。接着将由式(1)式和(2)计算得到的结果进行卷积和激活等操作,表达式如下:

$$f_c^h = \psi(F_{\text{conv}}(Z_c^h)) \cdot \gamma(\psi(F_{\text{conv}}(Z_c^h)) + 3) / 6$$
 (3)

$$f_c^w = \psi(F_{\text{conv}}(Z_c^w)) \cdot \gamma(\psi(F_{\text{conv}}(Z_c^w)) + 3) / 6$$
 (4)

$$a_h = \delta(F_{\text{conv}}(f_c^h))$$
 (5)

$$a_w = \delta(F_{\text{conv}}(f_c^w))$$
 (6)

式(3)–(5)中, ψ 表示 BatchNorm2d 进行数据的归一化处理操作; γ 表示 rule 函数, δ 表示 sigmoid 函数, γ 和 δ 都是激活函数。式(5)和式(6)分别计算每一行和每一列上的目标权重值。

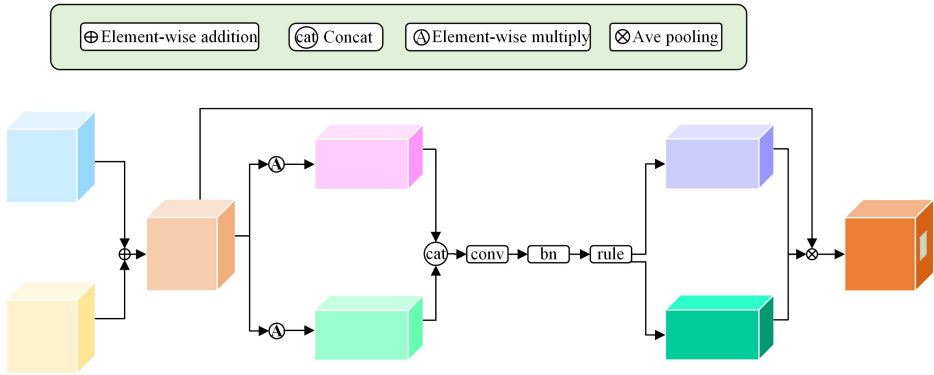


图 2 空间位置注意力模块
Fig. 2 Spatial location attention module

空间位置注意力模块首先分别对两种图像特征图的行与列做池化操作,获得目标的位置,然后经过归一化、激活等操作处理后将计算得到的目标位置信息乘到特征图上去。

2.3 高秩特征指导模块

本文提出的秩指导模块是受到文献[24]的启发,文献[24]中的研究发现,无论卷积神经网络接收到多少数据,单个滤波器生成的特征图的平均秩排名总是相同。又因为特征图

的秩能反映滤波器提取特征的能力,即特征图的平均秩越大,对应的滤波器提取特征的能力便越强。因此,本文只需要离线输入视频中的一小部分图像,即可计算特征图的平均秩(图 3 中的 X_n 对应着每一个特征图的平均秩的值)。然后根据离线计算保存的特征图平均秩信息,在跟踪时将这些平均秩作为权重来指导模态的融合。这样既不会增加跟踪时的计算负担,也可以使模态更好地融合。图 3 给出了高秩指导模块的流程图。

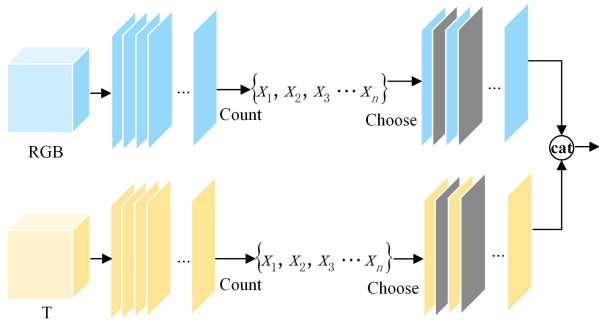


图3 高秩特征指导流程图

Fig. 3 High rank feature guide flow chart

下面详细介绍秩计算过程,为了获得特征图的秩信息,首先在跟踪时随机选择 5 个视频序列,对这 5 个视频中图像的特征图使用秩计算公式:

$$C^{rgb} = \text{matrix_rank}(f^{rgb}[i, j, :, :]) \quad (7)$$

$$C^t = \text{matrix_rank}(f^t[i, j, :, :]) \quad (8)$$

$$C_A^{rgb} = \sum_{i=0}^5 C_i^{rgb} / 5 \quad (9)$$

$$C_A^t = \sum_{i=0}^5 C_i^t / 5 \quad (10)$$

式(7)和式(8)中 matrix_rank 函数是 torch 中用来计算特征图秩的函数, i 是每帧图像的正样本区域数, j 是卷积层的通道数。式(9)和式(10)是对 5 张图片特征图的秩求平均值。因为视频中的第一幅图像被用来微调网络,所以在计算特征图的秩信息时,本文使用的是第二幅图像。由于每幅图像都采集 64 个正样本,因此本文 i 的值为 64。 j 是第三层卷积层的通道数,对应的数值是 512。为了计算特征图的平均秩,将每幅图像计算得到的秩信息保存下来,然后将保存的秩信息对应相加后计算平均值。为了后面的删除操作,将计算得到的平均值做规范化操作,使平均秩的值均在 0~1 之间(包括 0,1),计算式如下:

$$N = (C_A^{rgb} - \min(C_A^{rgb})) / (\max(C_A^{rgb}) - \min(C_A^{rgb})) \quad (11)$$

$$N = (C_A^t - \min(C_A^t)) / (\max(C_A^t) - \min(C_A^t)) \quad (12)$$

其中, $\max$ 是计算通过式(9)和式(10)得到的平均秩中的最大值, $\min$ 是计算通过式(9)和式(10)得到的平均秩中的最小值, C_A^{rgb} , C_A^t 是式(9)和式(10)计算得到的可见光特征和热红外特征的平均秩信息。为了去掉冗余信息和多余噪声,根据规范化后的平均秩信息对特征图做删除操作,即将平均秩小于 0.2 的特秩图做置零操作。将这些冗余信息和噪声去除后,网络取得了更鲁棒的性能。

根据特征图的秩信息来删除低秩特征图的操作,依据是奇异值的一些性质。奇异值矩阵是按照从大到小排列,而且奇异值减少特别快,在很多情况下,前 10% 甚至前 1% 的奇异值之和就占了全部奇异值之和的 99% 以上。换句话说,本文可以用最大的 k 个奇异值和每个奇异值所对应的左、右奇异向量来近似描述特征图。因此奇异值分解从数学上证明了高秩特征图可能包含了低秩特征图。假设本文有特征图 $F_{m \times n}$, 则证明公式如下:

$$F_{m \times n} = U_{m \times n} Z_{m \times n} V_{n \times n}^T \approx U_{m \times k} U_{k \times k} V_{k \times n}^T \quad (13)$$

高秩特征指导模块首先将经过特征提取网络得到的特征图做一个秩的计算,然后将计算得到的特征图的秩做归一化

操作,接着设置一个阈值,将低于阈值的特征图权重置零,高于阈值的特征图的权重为归一化后的秩的值。

2.4 网络训练

在网络训练这一部分,因为要离线计算秩,所以分两次训练网络。第一次训练时,首先加载预训练模型^[31]初始化主干网络的参数(卷积层和全连接层),接着使用随机梯度下降算法 SGD^[32]来优化网络,使用交叉熵损失对网络进行 100 次迭代。在每一次的迭代过程中,均从每一个视频序列中随机选择 8 帧,接着从每一帧中选取 64 个正样本和 192 个负样本。对于正样本的定义,是将与真值框的 IOU 大于 0.7 的框定义为正样本,与真值框的 IOU 小于 0.5 的框定义为负样本。训练时进行多域训练,即在 fc6 层是每个视频分别对应一个分支 $fc6^i$ 。

在第二次训练前,首先根据第一次训练得到的模型进行跟踪测试,并在跟踪时随机选择 5 个视频序列,计算这 5 个序列中第二张图片的特征图的秩并将器保存下来。之所以不选择第一张图片计算秩信息,是因为跟踪时算法用第一张图片进行了困难负样本挖掘和微调网络。接着对保存的特征图的秩信息计算平均值。在第二次训练时将特征图的平均秩作为权重乘到特征图上去,再训练网络,对于超参数的设置和第一次训练时基本一致,唯一不同的是,第二次训练的迭代次数为 500。

2.5 算法跟踪

在对目标进行跟踪时,与文献[33]一样,将域特定层(即最后一层全连接层)的多个分支用单个分支替换,并且在第一帧时更新模型的全连接层的参数。具体做法是,给定第一帧目标框,本文根据高斯分布选择 500 个正样本框和 5 000 个负样本框,然后用 50 次迭代来训练域特定层。需要注意的是,在跟踪时不使用秩特征模块,因为对于特征图的秩是离线计算并保存平均秩,所以在跟踪时只需要将保存的平均秩引入来指导模态融合即可。在跟踪第 T 帧图片时,将 $T-1$ 帧图片的跟踪结果作为真值框,然后根据高斯分布进行采样。跟踪时算法采用短时和长时两种更新方式,当跟踪失败时才进行短时更新。短时更新是选择最近的 20 帧跟踪成功时保存的正负样本集,长时更新是每 10 帧便自动进行一次,选择最近 100 帧跟踪成功时保存的正样本集和最近 20 帧跟踪成功时保存的负样本集。另外,对于跟踪成功的目标框采用回归策略微调包围框,以使目标框的大小和位置更加接近真值框。

3 本文算法

在 GTOT^[34] 和 RGBT234^[35] 两个公开的 RGBT 跟踪数据集上测评本文算法的性能,实验结果如图 4、图 5 所示。为了充分验证本文方法的有效性,本节做了大量的比较实验。

3.1 RGBT 数据集和评价指标

本文使用两个 RGBT 跟踪领域的公开数据集 RGBT234 和 GTOT。RGBT234 数据集提供了对齐的可见光和热红外图像对。它有多个属性挑战:目标未被遮挡、目标部分遮挡、目标被严重遮挡、低光照、目标与背景相似以及形变等。RGBT234 数据集有 234 个跟踪序列,每个序列的图像都被精准

标注,该数据集大约有 234 000 帧数据,其中最大的序列有 8000 帧左右。GTOT 数据集包含 50 个序列,这些序列包含了许多场景下对齐的可见光和热红外图像对,如实验室、公共道路、水池等。每个序列都被手工标注了真值框。

本文使用精确度和成功率这两个常见的评价指标来评估算法。精确度(PR)表示在视频序列中预测框的中心位置到真值框的中心位置的距离小于预定阈值的帧占总帧的百分比。对 GTOT 数据集采用的阈值为 5 像素(参考文献[32] GTOT 数据集),对 RGBT234 数据集采用的阈值为 20 像素(参考文献[33]RGBT234 数据集)。成功率(SR)表示预测输出位置与真值框重叠率大于阈值的帧所占百分比。通过改变阈值,得到 SR 曲线。

3.2 在 GTOT 上的评估

为了评估本文方法的有效性,在数据集 GTOT 上测试了结果。图 4 直观地给出了本文方法与多种跟踪器(MDNet+RGBT^[21], CAT^[23], DAT + RGBT^[36], DAFNet^[37], DAPNet^[38], RT-MDNet + RGBT^[39], SiamDW + RGBT^[40], ECO^[41],SGT^[42],MANet^[43],STRUCK^[44])的比较结果。

图 4 是本文 HRPANet 算法与其他跟踪算法在 RGBT 数据集 GTOT 上对比的曲线图。图 4 中不同颜色、不同线型的曲线代表不同算法。Precision 表示精确度 PR,Success Rate 表示成功率 SR。从图 4 可以直观地看出本文提出的方法在精确度上高于其他 RGBT 算法和其他 RGB 扩展后的算法(为了统一比较,将原本是 RGB 的算法扩展成 RGBT,如 MDNet,STRUCK 等算法)。因为精确度代表预测框中心位置和真值框中心位置的偏差,所以本文算法的结果证明了在网络中加入位置注意模块的有效性。

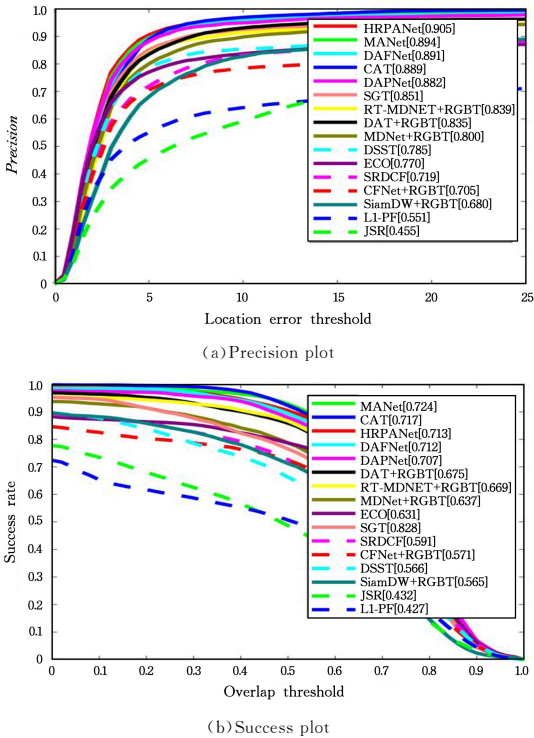


图 4 在数据集 GTOT 上的结果曲线(电子版为彩图)

Fig. 4 Curve of results on GTOT dataset

为了方便地对比网络的性能,在表 1 中列出了图 4 中

所有算法的 PR 和 SR 值。可以清楚地看到,本文提出的 HRPANet 算法的 PR 值比性能第二的 MANet 高出 1.1%,SR 略高于 DAFNet。本文方法的 PR 值较高的原因在于其主要依赖于空间位置注意力模块的有效性。PR 表示精确度,它代表预测框中心位置与真值框中心位置误差小于阈值的帧占总帧数的比值,位置注意力模块关注目标的位置信息,因此本文方法的 PR 值较基准方法有很大的提高,较其他跟踪算法相比也有一定的提高。

表 1 在数据集 GTOT 上的结果数据

跟踪器	PR	SR
HRPANet	0.905	0.713
MANet	0.894	0.724
DAFNet	0.891	0.712
CAT	0.889	0.717
DAPNet	0.882	0.707
SGT	0.851	0.628
RT-MDNet+RGBT	0.839	0.669
DAT+RGBT	0.835	0.675
MDNet+RGBT	0.800	0.637
ECO	0.770	0.631
STRUCK	0.681	0.533
SiamDW+RGBT	0.680	0.565

3.3 在 RGBT234 上的评估

为了进一步评估本文方法的有效性,在一个更大的数据集 RGBT234 上进行了测试。如图 5 所示,将本文的算法与 12 种 RGBT 算法(CAT^[23],MANet^[43],DAFNet^[37],DAPNet^[38],RT-MDNet + RGBT^[39],DAT + RGBT^[36],MDNet + RGBT^[2],SGT^[42],ECO^[41],SiamDW + RGBT^[40],CFnet + RGBT^[45],KCF+RGBT^[46],L1-PF^[47],JSR^[48])做了比较。

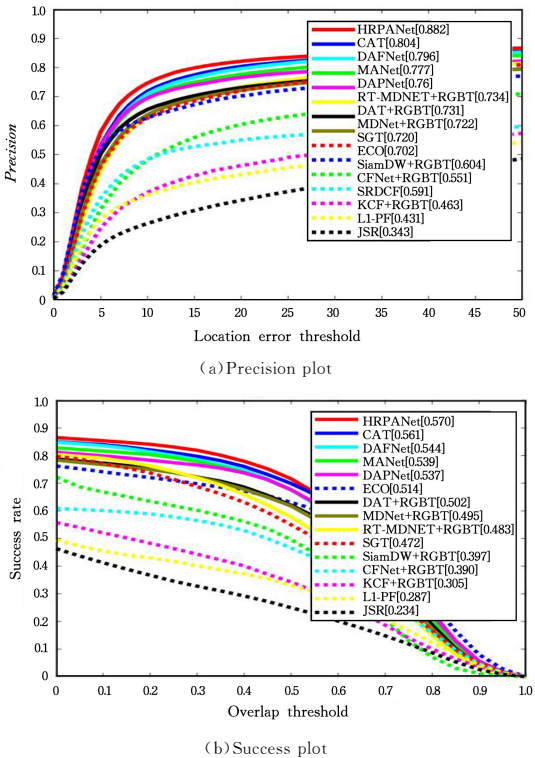


图 5 在数据集 RGBT234 上的结果曲线

Fig. 5 Curve of results on RGBT234 dataset

为了能准确地分析数据,表 2 中列出了本文算法 HRPANet 与其他目标跟踪算法的 PR 和 SR 数值。可以看到,本文提出的算法 HRPANet 的实验结果为: $PR=0.822$, $SR=0.570$,均高于其他算法。在精确度上比性能第二的挑战感知网络(CAT)高出了约 2%,在成功率上高出了约 1%。通过在以上两个 RGBT 数据集上与多种 RGBT 算法的实验结果对比可以得出,本文算法无论在准确率还是在成功率上都取得了较好的性能。

表 2 在数据集 RGBT234 上的结果数据

Table 2 Result data on RGBT234 dataset

跟踪器	准确度	成功率
HRPANET	0.822	0.570
CAT	0.804	0.561
MANet	0.799	0.567
DAFNet	0.796	0.544
DAPNet	0.766	0.537
RT-MDNet+RGBT	0.734	0.483
DAT+RGBT	0.731	0.502
MDNet+RGBT	0.722	0.495
SGT	0.720	0.472
ECO	0.702	0.514

3.4 消融实验

为了验证所提位置注意力模块和高秩特征模块的有效性,本文设计了 3 个变种跟踪器:HRPANet-RM,HRPANet-HRM 和 HRPANet-SLM。因为本文算法主要解决精准目标定位,所以主要分析 PR 值(精确度)。分别在 GTOT 和 RGBT234 两个数据集上进行测试分析。HRPANet-RM 只使用特征图的秩指导模态融合,与基准相比,在 GTOT 和 RGBT234 两个数据集上, PR 分别提高了 0.29 和 0.06,结果达到了 0.895 和 0.808;HRPANet-HRM 不仅使用特征图的秩指导模态融合还删除了低秩的特征图,从而减少了噪声,与基准相比,在 GTOT 和 RGBT234 两个数据集上, PR 分别提高了 0.36 和 0.08,结果分别达到了 0.902 和 0.810;HRPANet-SLM 只使用了位置注意力模块, PR 结果分别达到了 0.900 和 0.817。HRPANet 为本文提出的算法。结果如表 3 所列。

表 3 各个模块的效果数据

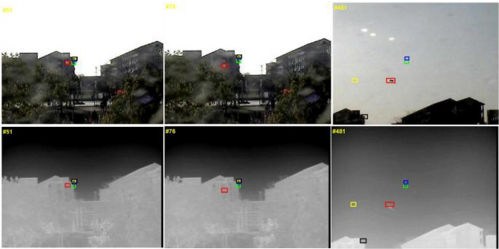
Table 3 Effect data of each module

跟踪器	RGBT234		GTOT	
	PR	SR	PR	SR
Baseline	0.784	0.554	0.866	0.693
HRPANET-RM	0.808	0.569	0.895	0.723
HRPANET-HRM	0.810	0.569	0.902	0.722
HRPANET-SLM	0.817	0.568	0.900	0.721
HRPANET	0.822	0.570	0.905	0.713

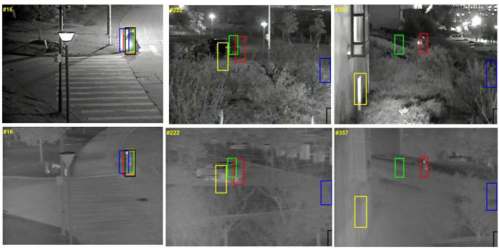
3.5 定性评估

为了更好地展示本文算法的效果,图 6 给出了算法的可视化结果图。不同的颜色框对应不同的算法,红色框为 HRPANet 即本文提出的算法结果框,绿色框为 DAPNet,蓝色框为 MDNet+RGBT,黄色框为 MANet,黑色框为 ECO。从

图 6 中的结果框可以看出,本文算法 HRPANet 与其他目标跟踪算法相比取得了较好的跟踪结果。如 Kite2 中当目标出现较大位置移动时,本文算法依然能够实现精准定位并跟踪目标;在 Elecbike10 和 Children3 中当目标被遮挡和背景杂乱时,本文算法与其他目标跟踪算法相比跟踪效果依然是最好的。



(a) Kite2



(b) Elecbike10



(c) Children3

图 6 算法可视化结果图(电子版为彩图)

Fig. 6 Graph of algorithm visualization results

3.6 在各种挑战上的评估

在数据集 RGBT234 上共有 12 种挑战,分别是:目标无遮挡(No Occlusion)、目标存在部分的遮挡(Partial Occlusion)、目标区域的 80%被遮挡(Heavy Occlusion)、目标区域光照强度较低(Low Illumination)、目标温度和背景温度相似(Thermal Crossover)、目标的分辨率偏低(Low Resolution)、目标发生形变(Deformation)、目标边界框移动的距离超过 20 个像素值(Fast Motion)、第一帧目标边界框与当前帧目标边界框的重叠率相比低 0.5(Scale Variation)、由于运动使目标模糊(Motion Blur)、相机发生移动(Camera Moving)和目标区域中背景杂乱(Background Clutter)。将本文算法与其他 RGBT 目标跟踪算法在各种挑战上进行对比,结果如表 4 所列。从表 4 中可以看出本文算法 HRPANet 在所有的挑战中的表现几乎都是最好的。HRPANet 能很好地解决所有挑战的原因在于该算法是基于目标跟踪任务设计网络的(精准定位目标),所以其对于不同视频序列以及不同挑战均能取得很好的性能。

表 4 多种算法在不同挑战上的结果

Table 4 Results of multiple algorithms on different challenges

	SOWP+ RGBT	CFNET+ RGBT	SGT	L1-PF	MDNet+ RGBT	FANet	DAPNet	ours
NO	86.8/53.7	76.4/56.3	87.7/55.5	56.5/37.9	86.2/61.1	84.7/61.1	90.0/64.4	94.1/66.8
PO	74.7/48.4	59.7/41.7	77.9/51.3	47.5/31.4	76.1/51.8	78.3/54.7	82.1/57.4	86.8/60.8
HO	57.0/37.9	41.7/29.0	59.2/39.4	33.2/22.2	61.9/42.1	70.8/48.1	66.0/45.7	72.2/48.9
LI	72.3/46.8	52.3/36.9	70.5/46.2	40.1/26.0	67.0/45.5	72.7/48.8	77.5/53.0	85.0/56.6
LR	72.5/46.2	55.1/36.5	75.1/47.6	46.9/27.4	75.9/51.5	74.5/50.8	75.0/51.0	83.7/55.6
TC	70.1/44.2	45.7/32.7	76.0/47.0	37.5/23.8	75.6/51.7	79.6/56.2	76.8/54.3	82.1/57.3
DEF	65.0/46.0	52.3/36.7	68.5/47.4	36.4/24.4	66.8/47.3	70.4/50.3	71.7/51.8	77.7/55.1
FM	63.7/38.7	37.6/25.0	67.7/40.2	32.0/19.6	58.6/36.3	63.3/41.7	67.0/44.3	81.4/50.9
SV	66.4/40.4	59.8/43.3	69.2/43.4	45.5/30.6	73.5/50.5	77.0/53.5	78.0/54.2	81.4/56.6
MB	63.9/42.1	35.7/27.1	64.7/43.6	28.6/20.6	65.4/46.3	67.4/48.0	65.3/46.7	72.6/52.4
CM	65.2/43.0	41.7/31.8	66.7/45.2	31.6/22.5	64.0/45.4	66.8/47.4	66.8/47.7	76.6/54.5
BC	64.7/41.9	46.3/30.8	65.8/41.8	34.2/22.0	64.4/43.2	71.0/47.8	71.7/48.4	83.9/55.0
ALL	69.6/45.1	55.1/39.0	72.0/47.2	43.1/28.7	72.2/49.5	76.4/53.2	76.6/53.7	82.2/57.0

结束语 本文提出了一种基于高秩特征和位置注意模块的 RGBT 目标跟踪算法。本文方法在主干网络的每一层均引入位置注意模块来关注目标的位置信息,该位置注意模块通过同时关注特征图的行与列信息来实现关注目标位置的目的,最终实现帮助网络精确地定位目标。对于每一层特征,本文方法都关注了目标位置,然后在最后一层特征融合时,通过探索高秩的特征图来关注通道重要性,以达到对两个模态特征图更有效的融合,也就是本文方法中的高秩特征模块。该模块通过离线训练计算得出平均秩信息。对于平均秩小的特征图,本文算法采用了删除操作,这样做是为了去除冗余和噪声,实现更加鲁棒的特征表达,并且计算得到的平均秩能反映网络滤波器提取特征的能力,因此该模块在跟踪时几乎不会对网络产生额外的负担。通过充分的实验证明,本文提出的方法在两个公开的 RGBT 基准数据集上均取得了较好的性能。在未来的工作中将继续探索如何更好地关注目标的边界框和探索特征图秩信息对其他任务的有效性。

参 考 文 献

[1] JI H X Y,LIANG P P,CHAI Y M,et al. Planar Object Tracking Algorithm Based on Key Points and Optical Flow[J]. Computer Engineering,2021,47(4):234-240.

[2] DANELLJAN M,BHAT G,KHAN F S,et al. Atom: Accurate tracking by overlap maximization [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4660-4669.

[3] WANG Q,ZHANG L,BERTINETTO L,et al. Fast online object tracking and segmentation: A unifying approach[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:1328-1338.

[4] WU Y,LIM J,YANG M H. Online object tracking: A benchmark[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013:2411-2418.

[5] VALMADRE J,BERTINETTO L,HENRIQUES J F,et al. Long-term tracking in the wild: A benchmark[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 670-685.

[6] MARTIN D,GUSTAV H,FAHAD K,et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C]//ICCV. 2015.

[7] MARTIN D,BHAT G,SHAHBAZ KHAN F,et al. Eco:Effi-

cient convolution operators for tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6638-6646.

[8] LI F,TIAN C,ZUO W,et al. Learning spatial-temporal regularized correlation filters for visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4904-4913.

[9] LI G,YU Y. Deep contrast learning for salient object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:478-487.

[10] WU A,ZHENG W S,YU H X,et al. RGB-infrared cross-modality person re-identification[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5380-5389.

[11] BADRINARAYANAN V,KENDALL A,SEGNET R C. A deep convolutional encoder-decoder architecture for image segmentation[J]. arXiv:1511.00561,2015.

[12] LI C,XIA W,YAN Y,et al. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,32(7):3069-3082.

[13] XU D,OUYANG W,RICCI E,et al. Learning cross-modal deep representations for robust pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:5363-5371.

[14] GADE R,MOESLUND T B. Thermal cameras and applications: a survey[J]. Machine Vision and Applications,2014,25(1):245-262.

[15] LI C,ZHAO N,LU Y,et al. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017:1856-1864.

[16] LAN X,YE M,ZHANG S,et al. Robust collaborative discriminative learning for RGB-infrared tracking[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 7008-7015.

[17] WANG Y,LI C,TANG J. Learning soft-consistent correlation filters for RGB-T object tracking[C]// Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer,2018:295-306.

[18] LI C,ZHU C,HUANG Y,et al. Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:808-823.

- [19] ZHANG L,DANELLIAN M,GONZALEZ-GARCIA A,et al. Multi-modal fusion for end-to-end rgb-t tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019:2252-2261.
- [20] ZHANG H,ZHANG L,ZHUO L,et al. Object tracking in RGB-T videos using modal-aware attention network and competitive learning[J]. Sensors,2020,20(2):393-411.
- [21] WANG C,XU C,CUI Z,et al. Cross-modal pattern-propagation for RGB-T tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7064-7073.
- [22] ZHANG P,ZHAO J,BO C,et al. Jointly modeling motion and appearance cues for robust RGB-T tracking[J]. IEEE Transactions on Image Processing,2021,30:3335-3347.
- [23] LI C,LIU L,LU A,et al. Challenge-aware RGBT tracking [C]// European Conference on Computer Vision. Cham: Springer,2020:222-237.
- [24] LIN M,JI R,WANG Y,et al. Hrank:Filter pruning using high-rank feature map[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:1529-1538.
- [25] HOU Q,ZHOU D,FENG J. Coordinate attention for efficient mobile network design [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713-13722.
- [26] WOO S,PARK J,LEE J Y,et al. Cbam:Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV). 2018:3-19.
- [27] HU J,SHEN L,ALBANIE S,et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,42(8):2011-2023.
- [28] ZHU Y,LI C,TANG J,et al. Quality-aware feature aggregation network for robust RGBT tracking[J]. IEEE Transactions on Intelligent Vehicles,2020,6(1):121-130.
- [29] TANG Z,XU T,LI H,et al. Exploring Fusion Strategies for Accurate RGBT Visual Object Tracking[J]. arXiv:2201.08673, 2022.
- [30] CHATFIELD K,SIMONYAN K,VEDALDI A,et al. Return of the devil in the details:Delving deep into convolutional nets[J]. arXiv:1405.3531,2014.
- [31] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.
- [32] BOTTOU L. Large-scale machine learning with stochastic gradient descent[C]//Proceedings of COMPSTAT. 2010:177-186.
- [33] NAM H,HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4293-4302.
- [34] 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.
- [35] LI C,LIANG X,LU Y,et al. RGB-T object tracking: Benchmark and baseline[J]. Pattern Recognition,2019,96:106977.
- [36] PU S,SONG Y,MA C,et al. Deep attentive tracking via reciprocal learning[J]. Advances in Neural Information Processing Systems,2018,31:1935-1945.
- [37] GAO Y,LI C,ZHU Y,et al. Deep adaptive fusion network for high performance RGBT tracking [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019:91-99.
- [38] ZHU Y,LI C,LUO B,et al. Dense feature aggregation and pruning for rgbt tracking[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019:465-472.
- [39] JUNG I,SON J,BAEK M,et al. Real-time mdnet[C]//Proceedings of the European Conference on Computer Vision(ECCV). 2018:83-98.
- [40] ZHANG Z,PENG H. Deeper and wider siamese networks for real-time visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4591-4600.
- [41] DANELLIAN M,BHAT G,SHAHBAZ KHAN F,et al. Eco: Efficient convolution operators for tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6638-6646.
- [42] LI C,ZHAO N,LU Y,et al. Weighted sparse representation regularized graph learning for RGB-T object tracking[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017:1856-1864.
- [43] LI C L,LU A,ZHENG A H,et al. Multi-adapter RGBT tracking [C] // 2019 IEEE/CVF International Conference on Computer Vision Workshop(ICCVW). IEEE,2019:2262-2270.
- [44] HARE S,GOLODETS S,SAFFARI A,et al. Struck:Structured output tracking with kernels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,38(10):2096-2109.
- [45] VALMADRE J,BERTINETTO L,HENRIQUES J,et al. End-to-end representation learning for correlation filter based tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2805-2813.
- [46] 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, 2014,37(3):583-596.
- [47] WU Y,BLASCH E,CHEN G,et al. Multiple source data fusion via sparse representation for robust visual tracking[C]// 14th International Conference on Information Fusion. IEEE,2011:1-8.
- [48] LIU H P,SUN F C. Fusion tracking in color and infrared images using joint sparse representation[J]. Science China Information Sciences,2012,55(3):590-599.



YANG Lan-lan, born in 1994, postgraduate. Her main research interests include RGBT object tracking and so on.



WANG Fu-tian, born in 1981, Ph. D, professor. His main research interests include image processing, computer vision and edge computing.