文章编号:1001-5078(2024)05-0774-07

·红外技术及应用·

基于区域候选孪生网络的红外目标跟踪算法

刘效伦,安成斌
(华北光电技术研究所,北京 100015)

摘　要:目标跟踪是光电设备的基本功能。为了应对跟踪任务中目标快速运动、复杂背景干扰以及遮挡的影响,不同于传统生成式方法与核相关滤波方法,本文提出一种使用深度学习的红外目标跟踪算法,使用双分支孪生特征提取网络对输入进行空间映射,经锚框划分图像区块后,分流为区域候选网络的“分类”与“回归”分支并进行特征模板匹配,对每个锚框进行分数评价后取“分类”分支中的最佳锚框,经“回归”分支进行预测边界回归后确定目标跟踪预测框,得到一种可以达到实时要求的红外单光宏观单目标跟踪算法。这种方法能够通过完全离线端到端训练整体系统参数获得,其制作过程简单,只要方法得当地进行参数精调,其性能有充分潜力可供挖掘。

关键词:信号与信息处理;跟踪算法;深度学习;红外目标;孪生网络;锚框
中图分类号:TN215;TP391.4　　**文献标识码:**A　　**DOI:**10.3969/j.issn.1001-5078.2024.05.017

Infrared target tracking algorithm based on Siamese region proposal network

LIU Xiao-lun, AN Cheng-bin
(North China Research Institute of Electro-Optics, Beijing 100015, China)

Abstract: Target tracking is a basic function of photoelectric equipment. In order to cope with the impact of fast target movement, complex background interference and occlusion in tracking tasks, an infrared target tracking algorithm using deep learning is proposed in this paper, which is different from traditional generative methods and kernel correlation filtering methods. The input is mapped using a double-branch Siamese network into a higher dimensional space of features, and the image blocks in video frames divided by anchors are sent into the "classification" and "regression" branches of the regional proposal network. Then, correlation calculations will be conducted on "classification" branch to evaluate the matching scores between features from the template image and the search image, producing a matrix of scores for every anchor generated. The best anchor is selected after the score evaluation, and the target tracking prediction box is determined after the boundary regression from that anchor with the information of "regression" branch. An infrared single-light macro single-target tracking algorithm meeting the real-time requirements is proposed. This approach can be obtained by training the overall system parameters end-to-end completely offline, is simple to produce, and has full potential for performance that can be exploited with proper parameter fine-tuning of the methodology.

Keywords: signal and information processing; tracking algorithm; deep learning; infrared targets; Siamese network; anchor box

1　引　言

目标跟踪是计算机视觉的一项基本任务,在设备自主火控、智能监控甚至系统智能控制方面起到着至关重要的作用。在可见光、红外成像设备的智

收稿日期:2023-12-07;**修订日期:**2024-01-29

能化进程中,发展出一套高性能的目标跟踪算法是极具战略意义的。

在单目标跟踪算法领域,常用的跟踪方法包括传统的人工设计信号滤波方法、核相关滤波方法、以及近八年新兴的深度学习方法。在生产与生活中,使用最为广泛的是核相关滤波方法。这种方法基于图像模板匹配,其主要思想是设定一个图像滤波核,在过程中通过相关运算估计预测位置,同时设计模板更新方法在线学习新滤波核。此类方法包括MOSSE[1]、CSK[2]、KCF[3]、DSST[4],其中的 KCF 算法已被广泛用作解决方案。但是核相关滤波算法存在诸多不足,如仅依靠输入视频的像素级特征,很难应对跟踪目标形变、运动、尺度变化、背景干扰的挑战,再如仅支持在线学习也限制了对算法的调整空间,对于不同任务无法进行针对性调整,还如初始滤波核尺寸固定,多尺度、多比例目标适应性较差等。

现有核相关滤波算法无法满足日渐增长的性能需求,对此本文提出一种以孪生网络双分支运算为核心,同时使用区域候选网络来提高性能的,基于深度学习模型训练的目标跟踪算法。这种方法继承了核相关滤波方法目标模板匹配的思想,将匹配对象转为目标的浅层与深层深度特征,这在保留了目标结构信息的同时,还增加了对目标语义信息的匹配度考量,使得这种方法能够更深刻地捕捉目标的关键信息,减少了目标形变、运动、背景干扰对跟踪过程的影响。同时,为了解决目标多尺度、多比例变化的问题,使用区域候选网络中的“多比例锚框”进行“回归”操作,在支持对目标尺度、比例自适应预测的同时,还提高了预测框的定位精度。此外,得益于深度学习方法面向实践、可离线训练的特性,其制作过程、调试过程简单,只要方法得当的精调参数,其性能有充分潜力可供挖掘。

算法整体由两大主要部分组成:使用深度特征的孪生网络结构,以及使用分类、回归分支的区域候选网络,如图 1 所示。

2 深度神经网络

基于深度学习方法与传统生成式滤波方法、核相关滤波方法最本质的不同是,其使用深层神经网络将视频与视频帧中的低层特征表示,如灰度、轮廓、形状、位置等,转化为高层特征表示,从而把初始的、与输出目标之间联系不太密切的输入表示,转化成与输出目标联系更密切的表示,使得原来仅基于最后一层输出映射难以完成的任务成为可能。

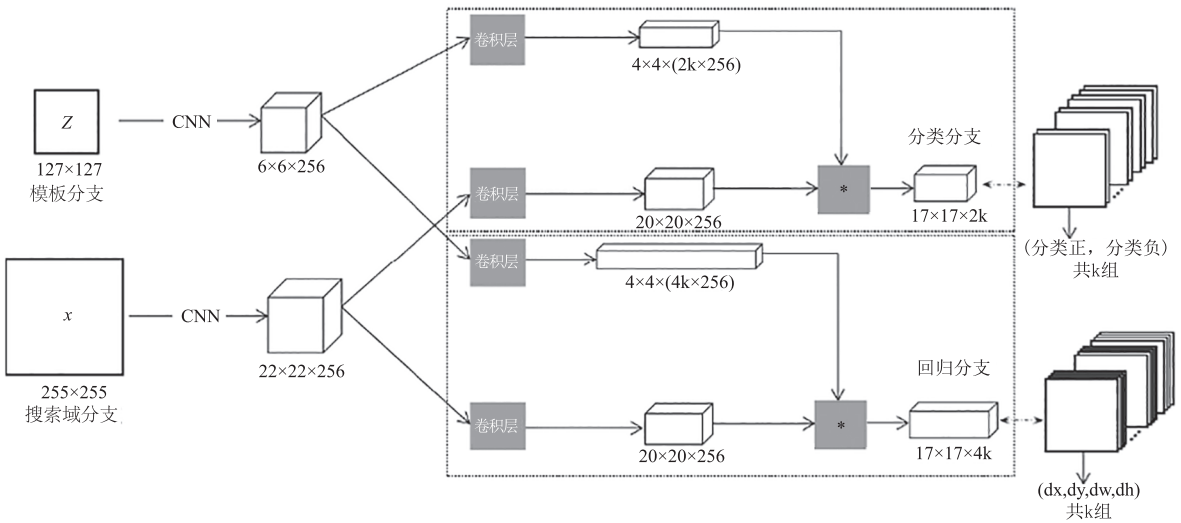

图 1 算法整体流程图

Fig. 1 Algorithm procedure

在 LeCun 首次应用卷积神经网络(Convolutional Neural Network,简称 CNN)[5-6]解决 MNIST 手写数字识别任务后,CNN 便成为使用深度学习解决视觉算法问题的重要方法,如图 2 所示。

传统生成式滤波方法中“基于特征匹配”的方法与深度学习方法思路相近,但是由于描述样本的特征通常需要由人类专家来设计,而人类专家设计出好特征并非易事,深度学习却能通过特征学习的过程自动产生好特征。核相关滤波方法同样基于模板匹配的思想,但是在特征选取方面,主要依靠人工

设计浅层特征。后续研究者发现,深度学习的 CNN 特征的浅层特征有助于进行目标定位,而深层特征有助于进行目标语义描述,便尝试加入 CNN 特征。但是由于其只能支持在线学习,而在线学习中循环移位产生的样本量,远远不能达到训练深度 CNN 特征的要求,因此不能充分发挥 CNN 特征的所有优势。而深度学习方法通过使用下一节将要介绍的孪生网络(Siamese Network)结构,能够实现使用层数更深、层次更加丰富的神经网络,通过合理的算法结构设计,可以更加充分地发挥深度特征的潜力。

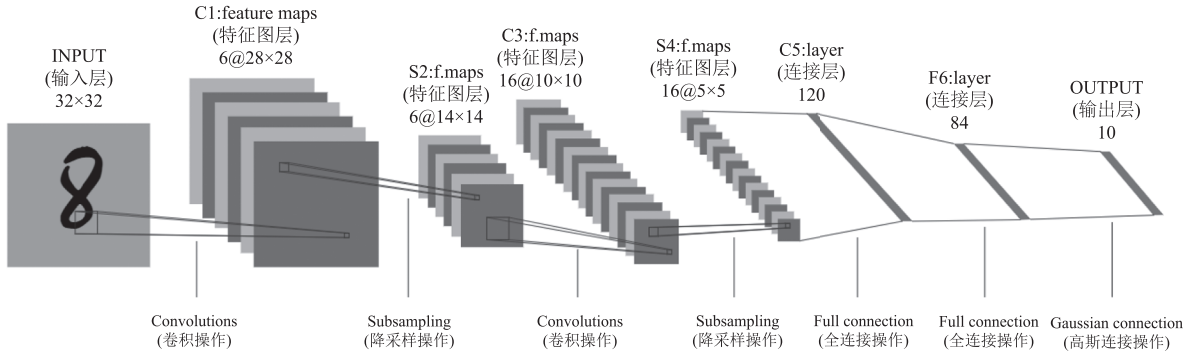

图 2 MNIST 手写数字识别卷积神经网络

Fig. 2 CNN for MNIST handwritten number recognition

3 孪生网络

孪生网络理论[7]是在解决核相关滤波问题的过程中发展而来,其成功另立新枝的关键在于,它通过使用深度学习理论,实现了在大量数据上离线进行特征学习,这使得深浅层特征与模式之间形成了复杂而丰富的表示关系。其网络结构如图 3 所示。

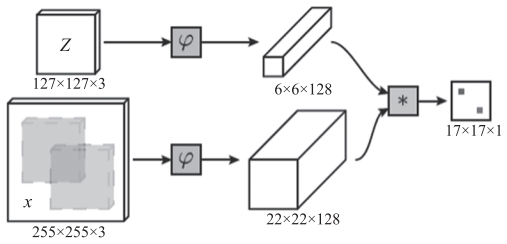

图 3 孪生网络结构图

Fig. 3 Siamese network framework

孪生网络将目标跟踪看成一个相似性学习问题。设相似度函数 $f(z,x)$,其中 z 表示目标图像,x 表示搜索区域图像,当两幅图像中的目标是同一目标时,则返回高分,反之低分。将划定待跟踪目标的第一帧图像设为 z,根据模板匹配的思想,截取目标周围一定范围内的区域作为搜索区域图像 x,在 x 内能够使两者在特征空间内距离最短的预测跟踪框,便是最佳跟踪结果。

设对于两分支的输入数据 (z,x),施加相同的特征空间变换 ϕ,并令相似度函数 $g = f(\phi(z),\phi(x))$。当函数 g 是一个简单距离度量或简单相似度度量时,变换 ϕ 可以被看作一个嵌入层变换,它将一张视频帧图像映射为特征空间中的向量或向量组。

现设计一种关于搜索图像 x 全卷积的孪生结构,也即对于 x,整体网络结构对平移运算具有互换性。更加具体的表述,引入平移操作算子 L_τ,根据定义有 $(L_\tau x)[u] = x[u-\tau]$。定义函数 h 是关于步长 k 全卷积的、信号对信号的映射,即对于任意平移 τ 具有:

$$h(L_{k\tau}x) = L_\tau h(x) \tag{1}$$

为了在整幅图像内遍历搜索相似度评分最高的位置,一般需要按步长对整幅视频帧进行平移遍历,截取与模板图像同样大小的搜索区域小窗,之后对每个小窗都使用相似度函数进行评分计算,排列为得分矩阵图。但是由于使用了全卷积网络,其优势是在为网络提供输入时,所提供的候选搜索区域图像输入可以是任意大小的,而不是像以往一样只能与模板图像相同大小的搜索区域小窗,这表示可以通过搭建一个密集的网格图来代表所有平移窗口。同时在进行相似度评分计算时,使用小窗进行逐次计算,其效果在数学上等价于,使用整体搜索域图像过系统后的特征图进行仅一次互相关运算。由于在既有的卷积运算理论与计算机库函数下,互相关运算可以极简单地通过卷积完成操作,所以采用一次互相关运算不仅不会影响计算效率,反而会极大简

便运算与程序设计过程。

此外，这样不但能极大简便推理过程的运算，而且可以同样应用于训练过程中。利用网络的全卷积特性，训练对的组成将使用一个模板图像与一个相应的整体搜索域图像，并作为输入传入网络。网络的输出将以实值互相关得分图的形式呈现，其定义域 $D \subset Z^2$，映射法则 $v: D \to R$，对于定义域内每一个位置 u，均有一个真值标签 $y[u] \in \{+1, -1\}$ 指导监督学习，因此可以非常高效的以“组”的形式从每个训练对上生成正负样本训练样例。对于整幅特征图的损失函数，定义其为每个个体样例损失的平均，即：

$$\mathrm{Loss}(y, v) = \frac{1}{|D|}\sum_{u \in D} l(y[u], v[u]) \tag{2}$$

嵌入层系统参数 θ 通过对下式进行随机梯度下降获得：

$$arg\min_{\theta} E_{(z,x,y)} \mathrm{Loss}(y, f(z, x; \theta)) \tag{3}$$

训练用样本对的获得，从已标记的数据集中选取视频片段，在视频中标记目标的中心，以设定好的边长截取模板或搜索域图像。之后，在同视频中取 T 帧为间隔的另一帧，从其标记中心外截取训练对的另一种图像。目标的物体种类在训练中予以忽略。而每幅图像中目标的尺度，在不破坏图像长宽比的情况下予以标准化。

在实验用特征变换网络 ϕ 方面，选取结构简单、参数轻量的 AlexNet[8] 进行参数微调，同时在每卷积层后进行一次批归一化操作，每卷积层前一次过激活函数层。但是值得注意的是，为了不打破网络的全卷积性，在特征提取网络内各层均禁止了边界填充，这是与图像分类任务中所使用特征提取网络的显著不同。

在原始的孪生网络结构中没有进行边界框回归操作，同时在目标多尺度变化问题上，使用的是繁复且冗余的多尺度遍历方法。孪生网络结构支持任意无填充神经网络进行离线端到端学习，可以基于这个基本结构对算法进行合理扩展。为了同时解决边界框回归以及多尺度搜索问题，将引入下文的区域候选网络，改善孪生网络计算相似度的过程。

4　区域候选网络

区域候选网络（Region Proposal Network，简称 RPN）最初提出于 Faster-RCNN[9]，是两阶段目标检测算法系列中用于生成候选区域，确定目标位置、尺度与比例，以进行后续检测运算的关键方法。在 RPN 以前，经典的候选区域提取方法一般都很耗时。RPN 则通过训练神经网络来生成候选区域，其精准度更高、速度更快。

RPN 的主要思想是，以约定的相等间隔在整体图像中确定采样基准点，称为“锚点”。之后，以每个锚点为中心进行数次采样，每次采样为一个不同尺度、不同比例的矩形区域。通过采样，由整幅图像将生成一系列矩形小窗，称这种矩形小窗为“锚框”。可以肯定的是，目标整体或部分必将位于其中一个锚框内。通过计算，确定一组锚框中与目标重合程度最好的一个，并将该框的边界通过回归运算还原至目标边界位置，便可完成对区域候选的精准提取与定位。上述运算都可通过离线端到端学习完成。其大致效果如图 4 所示。

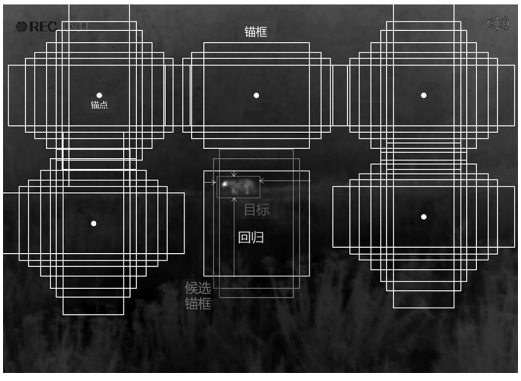

图 4　RPN 使用锚框生成目标区域候选效果图

Fig. 4 Illustration of RPN generating region proposals with anchors

RPN 计算过程插入在孪生网络特征提取之后，将孪生网络每分支的特征输出图，分别传入“分类分支”与“回归分支”两个不同分支中。“分类分支”用于区分搜索域特征图中一点是否属于目标，按“属于前景”与“属于背景”两类别进行训练，将提取区域候选问题转变成关于锚框的正负二分类模式识别问题。“回归分支”用于对锚框进行回归运算，在已知模板信息的辅助下，让每锚框经四维值修正后趋于目标实际位置。

每锚点生成的所有锚框，在经过整体系统后，其在相关度得分图中将呈现为一个多通道的点，这意味着相关度得分图与锚点之间将形成一种一一对应的关系。对于每个锚点，根据超参数设定的锚框边

长、尺度因子、比例因子生成 k 个锚框，其生成策略是，设尺度因子 Scale = $\{s_1, s_2, \cdots\cdots s_m\}$ 共 m 个元素，比例因子 Ratio = $\{r_1, r_2, \cdots\cdots r_n\}$ 共 n 个元素，锚框基础边长为 l，则每个锚点的生成锚框数 $k = m \times n$，设 $i \in N^*$ 且 $i < k$，其中第 i 个矩形锚框的宽与高分别为：

$$w_i = \left[\sqrt{\frac{l^2}{r_{i \bmod n}}}\right] \times s_{i \bmod m} \tag{4}$$

$$h_i = [w_i \times r_{i \bmod n}] \times s_{i \bmod m} \tag{5}$$

其中，[·]为取整运算，mod 为取余运算。由此，则一幅图像总计可以生成得分图边长平方个锚点，每个锚点生成 k 个锚框。模板图像分支的输入，经过特征提取网络 $\phi(z)$ 后，送入 RPN 并增加至 $[\phi(z)]_{cls}$ 和 $[\phi(z)]_{reg}$ 两个分支，通过卷积层将通道数分别提升至 2k 倍与 4k 倍于 $\phi(z)$。搜索域图像分支经 $\phi(x)$ 后，也分流为 $[\phi(x)]_{cls}$ 和 $[\phi(x)]_{reg}$ 两个分支，但是保留通道数不变。之后，在分类与回归分支上分别进行相关运算，生成每分支的相关性得分图：

$$A^{cls}_{w \times h \times 2k} = [\phi(x)]_{cls} \cdot [\phi(z)]_{cls} \tag{6}$$

$$A^{reg}_{w \times h \times 4k} = [\phi(x)]_{reg} \cdot [\phi(z)]_{reg} \tag{7}$$

根据孪生网络理论，设数据集共 n 个“输入数据－标签对”样例 (z_i, x_i, y_i)，系统全体参数集为 θ，则优化过程可表示为：

$$\arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}(y_i, f(z_i, x_i; \theta)) \tag{8}$$

整体参数集 θ 由特征提取网络参数 W_ϕ、RPN 两分支参数 (W_{cls}, W_{reg}) 组成。设 RPN 过程函数 ζ，则上式可改写为：

$$\mathrm{Loss}_{cls}(i) = \mathrm{Loss}_{cls}(y_i, f(\zeta(\phi(x_i; W_\phi), \phi(z_i; W_\phi); W_{cls}))) \tag{9}$$

$$\mathrm{Loss}_{reg}(i) = \mathrm{Loss}_{reg}(y_i, f(\zeta(\phi(x_i; W_\phi), \phi(z_i; W_\phi); W_{reg}))) \tag{10}$$

$$\arg\min_{W_\phi W_{cls} W_{reg}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{Loss}_{cls}(i) + \lambda\, \mathrm{Loss}_{reg}(i) \tag{11}$$

分类损失 $\mathrm{Loss}_{cls}(i)$ 使用交叉熵损失，回归损失 $\mathrm{Loss}_{reg}(i)$ 使用 smooth L_1 损失，λ 为平衡超参数。

推理阶段，在正分类得分图中取 K 个得分最高的点，作为初始预测点。之后，依照以下策略对 K 个点进行重新排列：

(1) 以 g 为半径，丢弃得分图中距离中心过远的点，以及这些点表示的所有锚框，因为孪生网络结构截取输入图的过程，跟踪目标会定位在图中心；

(2)施加余弦窗，以抑制较大位移位置的得分；

(3)全体得分乘以惩罚因子，以抑制尺寸与比例变化较大位置的得分，惩罚因子：

$$penalty = e^{\alpha \cdot \max\left(\frac{r}{r'}, \frac{r'}{r}\right) \cdot \max\left(\frac{s}{s'}, \frac{s'}{s}\right)} \tag{12}$$

其中，α 为超参数系数；r 为当前位置的高宽比；r' 表示前一帧该位置的高宽比，s 与 s' 表示该位置在当前与前一帧的尺度。选取最佳分数预测点，之后使用该点锚框集，调用每锚框对应的回归数据，通过变换反推回在原视频帧中跟踪目标的跟踪预测框，至此得到最终预测输出。

5 实 验

实验在红外数据集 LSOTB-TIR[10] 上进行。该数据集由一个训练集与一个评估集组成，总共有 1400 个热红外图像视频序列与超过 600 k 帧视频帧。为了评估不同属性的跟踪器，该数据集定义了 4 个场景属性与 12 个挑战属性。训练过程遵循孪生网络同序列隔帧取样的规则，而测试过程使用 OPE 评估方法。

5.1 实施细节

使用在 ImageNet 上预训练且经调整的 AlexNet 作为特征提取网络的算子 $\phi(\cdot)$，并且前三层的卷积层锁定，仅精调最后两层卷积层。参数的获取通过对式(11)进行 SGD 优化获得。训练迭代层数共 50 层，学习率从 10^{-2} 下降至 10^{-6}。训练样本从 LSOTB-TIB 数据集中，以 T < 100 的间隔提取训练样本对。模板图像裁剪尺寸 127 × 127，搜索域图像裁剪尺寸两倍于模板，为 255 × 255。锚框尺度因子 $S = \{2, 4, 8, 12, 16\}$，比例因子 $R = \{0.33, 0.66, 1, 1.5, 3\}$，锚框基础边长 4，网络步长 8。推理过程中，算法全程没有使用在线模板调整。实验全程使用 PyTorch2.0.1 + CUDA11.7 框架，在 Intet I9 CPU，16G RAM，Nvidia RTX 4060 上进行。

5.2 性能指标

5.2.1 归一化的精度

精度表示预测框中心点与真值框中心点的欧氏距离，通常阈值为 20 像素，即预测值与真值的欧氏距离在 20 像素之内便视为追踪成功。但是精度没有考虑到目标的大小，导致对于小目标，即使预测框与真值框相距较远，欧式距离仍会在 20 像素内。为

了解决这种问题,考虑到真值框的尺度大小,将精度进行归一化,其取值在[0,0.5]之间,于是引入这种“归一化精度”作为评价方法。这种方法实质上,即判断预测框与真值框中心点的欧氏距离与真值框斜边的比例。

5.2.2 成功率

成功率指标是计算预测框与真值框的区域内像素的交并比,即两框交叠区域与两框所有覆盖区域的比值。该指标可以反应跟踪预测位置与实际位置的贴合程度,也同时反应了跟踪框的平稳度。通常使用的 AUC 分数,是积分下成功率曲线围成的面积,这种计算方法考虑到了不同阈值下的成功率分数,对算法性能的评估更具整体性。在一些研究中,性能评估会直接指定交并比阈值(如 0.5),这是因为当成功率曲线足够光滑,取 0.5 对应的成功率分数和计算成功率的 AUC 分数其意义是一样的。

5.3 实验结果

实验跟踪效果如图 5、6 所示,不同跟踪器间性能比较如表 1 所示。

表 1 不同跟踪器性能比较表

Tab. 1 Performace comparision of different trackers

跟踪器	归一化精度	成功率	成功率分数
Struck	0.474	0.398	0.385
KCF_ HOG	0.415	0.341	0.323
DSST	0.553	0.554	0.479
SRDCF	0.641	0.636	0.532
ECO_tir	0.749	0.747	0.619
CFNet	0.517	0.477	0.418
Siamese_ FC	0.649	0.638	0.519
MDNet	0.748	0.749	0.604
Ours	0.669	0.686	0.577

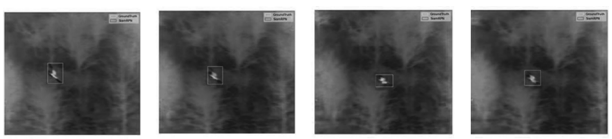

图 5 复杂背景、快速运动目标跟踪

Fig. 5 Tracking target of complex background and fast motion

本次实验着重于比较三种不同方法系列随研究发展的性能表现,传统滤波方法包括 Struck,核相关滤波方法包括 KCF-HOG、DSST、SRDCF、ECO,深度学习方法包括 CFNet、SiamFC、MDNet 以及本文的方法。由表 1 中各跟踪器的性能表现可以发现,传统滤波方法的性能表现显著低于核相关滤波与深度学习方法。核相关滤波方法中,早期仅使用混合人工设计像素级特征的方法,如 KCF-HOG、DSST 等,其性能相较后续加入了空间注意力机制的 SRDCF、加入了深度卷积特征的 ECO 等,存在明显的差距。新兴的深度学习方法方面,如使用双分支特征提取后进行模式匹配的孪生网络 Siam 系列,以及以 MDNet 为代表的为跟踪任务特制深度网络的系列方法,其性能与核相关滤波方法不分上下,两方都有较为优秀的表现。

图 6 多尺度变比例、部分遮挡、部分形变目标跟踪

Fig. 6 Tracking target of various scales and ratios, partial occlusion and deformation

本文提出的深度学习方法为一种孪生网络系列方法,其跟踪速度 113 f/s,达到了实时跟踪的要求。对于数据集中诸多的挑战情景,本文提出的跟踪器在应对快速运动目标、遮挡后重现的目标、多尺度变比例目标、复杂背景干扰的目标具有更好的跟踪效果。归功于深度网络强大的特征提取能力,在孪生结构的加持下,能够同时兼顾表征目标结构信息的浅层特征,以及表征目标语义信息的深层特征,这极大削弱了目标变化与目标背景干扰的影响。而在目标受到遮挡后重现的情景下,从目标中提取的深层特征语义信息,有助于在目标重现后进行更高效的模板匹配,提高了重跟踪成功的几率。

6 结 语

在目标跟踪领域,核相关滤波算法与以孪生网络结构为基础的深度学习方法,凭借优异的性能受到了学术界与工业界的广泛关注。由于两者在核心原理上的相似之处,两者在性能上也颇为接近,两者最新的研究进展在跟踪任务中也分别能够达到优秀的表现。虽然如此,但是两者在擅

长应对的场景与各自的缺陷与问题上,却有极大的差异。未来跟踪算法的发展,两种方法间的互相借鉴、互相融合将会是大势所趋。同时,基于运动描述进行跟踪的系列方法,也将会是未来跟踪算法的研究重点,以独立算法或辅助数据的形式出现在研究与应用中。

本文提出了一种结合孪生结构特征提取网络与区域候选网络的红外单光目标跟踪算法,在跟踪宏观单目标,尤其是针对快速运动目标、受遮挡目标、变尺度多比例目标、复杂背景目标时,其跟踪表现优异,在稳定精准跟踪目标的情况下,达到实时要求。但是,在跟踪较近距离内存在极度相似物的目标,以及大幅度形变导致形态完全改变的目标时,其性能表现有待进一步提升。这可能可以通过引入目标模板更新策略,或能够描述时空运动轨迹的注意力机制进行解决。

参考文献:

[1] M Danelljan, G Häger, F S Khan, et al. Discriminative scale space tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1561-1575.

[2] David S Bolme, J Ross Beveridge, Bruce A Draper, et al. Visual object tracking using adaptive correlation filters [C]//2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010: 2544-2550.

[3] F Henriques, Rui Caseiro, Pedro Martins, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//2012 European Conference on Computer Vision Part IV. Florence, Italy: Springer, 2012: 702-715.

[4] Henriques, Joao F, Caseiro, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[5] Lecun Y, Bengio Y. Convolutional networks for images, speech, and time-series[M]. The Handbook of Brain Theory and Neural Networks. Cambridge: MIT Press, 1995.

[6] Y Lecun, L Bottou, Y Bengio, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] Luca Bertinetto, Jack Valmadre, Joao F Henriques, et al. Fully-convolutional siamese networks for object tracking [C]//Computer vision-ECCV 2016 workshops, part 2: 14th European Conference on Computer Vision (ECCV 2016), October 8-10 2016, Amsterdam, The Netherland: Springer International Publishing, 2016: 850-865.

[8] Krizhevsky, Alex, Sutskever, et al. Image net classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[9] Ren Shaoqing, He Kaiming, Girshick Ross, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[10] Liu Q, Li X, He Z, et al. LSOTB-TIR: a large-scale high-diversity thermal infrared object tracking benchmark [C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020.