计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800164

基于时空注意力机制的目标跟踪算法

程　旭[1,2]　崔一平[1,2]　宋　晨[1,2]　陈北京[1,2]　郑钰辉[1,2]　史金钢[3]
1 南京信息工程大学计算机与软件学院　南京 210044
2 数字取证教育部工程研究中心南京信息工程大学　南京 210044
3 西安交通大学软件学院　西安 710049

摘　要　目标跟踪技术在智能监控、人机交互、无人驾驶等诸多领域得到了广泛的应用。近年来,学者们提出了许多高效的算法。然而,随着跟踪环境越来越复杂,目标跟踪算法在遮挡、光照变化、背景干扰等复杂环境下仍然面临着巨大的挑战,从而导致目标跟踪失败。针对上述问题,提出了一种基于时空注意力机制的目标跟踪算法。首先,采用孪生网络架构来提高对特征的判别能力;然后,引入改进的通道注意力机制和空间注意力机制,对不同通道和空间位置的特征施加不同的权重,并着重关注空间位置和通道位置上对目标跟踪有利的特征。此外,还提出了一种高效的目标模板在线更新机制,将第一帧图像特征与后续跟踪图像帧中置信度较高的图像特征进行融合,以降低发生目标漂移的风险。最后,在 OTB2013 和 OTB2015 数据集上对所提跟踪算法进行了测试。实验结果表明,所提算法的性能相比当前主流的跟踪算法提高了 6.3%。

关键词:深度学习;目标跟踪;孪生网络;注意力机制;模板更新
中图法分类号　TP301.6

Object Tracking Algorithm Based on Temporal-Spatial Attention Mechanism

CHENG Xu[1,2],CUI Yi-ping[1,2],SONG Chen[1,2],CHEN Bei-jing[1,2],ZHENG Yu-hui[1,2] and SHI Jin-gang[3]
1 School of Computer and Software,Nanjing University of Information Science and Technology,Nanjing 210044,China
2 Engineering Research Center of Digital Forensics,Ministry of Education,Nanjing University of Information Science and Technology,Nanjing 210044,China
3 School of Software Engineering,Xi'an Jiaotong University,Xi'an 710049,China

Abstract　Object tracking technology is widely used in intelligent monitoring,human-computer interaction,unmanned driving and many other fields. In recent years,many efficient tracking methods are proposed. However,object tracking methods still face great challenges in the complex scenario such as occlusion,illumination variations,background clutter,which leads to tracking failure. To solve the above mentioned problems,in this paper,an effective object tracking algorithm is proposed based on temporal-spatial attention mechanism. Firstly,we utilize the Siamese network architecture to improve the discriminative ability of object features. Then,the improved channel attention module and spatial attention module are introduced into the Siamese network,which imposes different weights on the features of different channels and spatial positions and focuses on the features that are beneficial to object tracking in spatial and channel positions. In addition,an efficient online object template updating mechanism is developed,which combines the features of the first frame and the features of the following frames with high confidence to reduce the risk of the object drift. Finally,the proposed tracking algorithm is tested on OTB2013 and OTB2015 benchmarks. Experimental results show that the performance of the proposed algorithm improves by 6.3% compared with the current mainstream tracking algorithms.

Keywords　Deep learning,Object tracking,Siamese network,Attention mechanism,Template update

到稿日期:2020-08-26　返修日期:2020-10-15
基金项目:国家自然科学基金(61802058,61911530397,62072251);中国博士后科学基金项目(2019M651650);南京信息工程大学人才启动经费(2018r057)
This work was supported by the National Natural Science Foundation of China(61802058,61911530397,62072251), Postdoctoral Research Foundation of China(2019M651650) and Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology(2018r057).
通信作者:程旭(xcheng@nuist.edu.cn)

1 引言

目标跟踪技术是计算机视觉领域的热点研究方向之一，在智能监控、人机交互、无人驾驶等诸多领域有着广泛的应用。目标跟踪指在连续的视频序列中建立所要跟踪物体的位置关系，得到物体完整的运动轨迹，通过给定图像第一帧的目标坐标位置，计算在之后序列图像中目标的位置和大小。目标跟踪技术的出现能够为行为理解、推理、决策等提供依据，是后续目标识别、行为分析、视频压缩编码和视频理解等高级视频处理任务的基础，是执行高层次智能行为的必要前提。尽管近年来目标跟踪技术取得了长足的进步，并且许多高效算法被提出，用于解决特定场景下的挑战性问题，但其仍然存在遮挡、照明变化、尺度变换、背景干扰等问题，因此目标跟踪技术的研究仍然是一项艰巨的任务。全卷积孪生网络通过使用孪生网络，将视觉目标跟踪问题转换为在已知目标和搜索区域的特征空间中求解相似性的问题。预先离线训练深度卷积网络，以此解决相似性学习的问题，使其可以在线跟踪任意目标。这种方法的优点是待搜索图像不需要与样本图像具有相同的尺寸，可以为网络提供更大的搜索图像作为输入，然后在密集网格上计算所有平移窗口的相似度。然而，这种通过模板匹配策略进行目标跟踪的方式的特征判别力受到了极大的限制，在跟踪过程中仅使用第一帧图像的目标信息，导致目标在遭到形变、遮挡等挑战时性能会下降。因此，本文引入通道注意力机制和空间注意力机制，使算法更多地关注空间位置和通道位置上对目标跟踪有利的特征。此外，孪生网络只保留第一帧图像特征，以避免目标特征被污染，但是该方法无法捕捉到目标在后续帧的变化。因此，当目标发生较大形变时，目标真实位置对应的响应值得分可能会变低，增加了目标丢失的风险。

为了应对上述问题，本文的贡献如下：1)提出了一种基于时空注意力机制的目标跟踪算法，采用孪生网络架构来提高对特征的判别能力；2)引入改进的通道注意力机制和空间注意力机制，对不同通道和空间位置的特征施加不同的权重，从而获得更具有判别能力的特征通道；3)提出了一种高效的在线更新机制，将第一帧图像特征与后续跟踪图像帧中置信度较高的图像特征进行融合，降低了目标在遭受到遮挡、形变等挑战时跟踪失败的风险。实验结果表明，本文算法在OTB2013[1]和OTB2015[2]数据集上具有更高的精度。

2 相关工作

目标跟踪是计算机视觉领域中的难点问题之一，在过去10年中许多学者对此进行了深入的研究。文献[3-4]对目标跟踪的研究现状进行了广泛的调查。本节主要讨论基于相关滤波和深度学习的目标跟踪两种算法。

基于相关滤波的目标跟踪算法[5-13]设计了一个滤波模板，利用该模板与目标候选区域做相关运算，输出最大响应的位置即为当前帧的目标位置。Bolme 等[5]提出的 MOSSE 算法是相关滤波跟踪的开篇之作，该跟踪器通过优化相关误差平方和的输出，利用自适应相关滤波器对目标外观进行编码，该算法每秒可以运行数百帧。2012 年，Heriques 等[6]提出了基于 MOOSE 的循环跟踪检测核结构 CSK(Exploiting the Circulant Structure of Tracking-by-detection with Kernels)算法。该算法通过循环矩阵解决了目标跟踪过程中训练样本太少的问题，并通过核技巧进一步提高了算法的准确性。相关滤波跟踪算法中最具有代表性的算法是 KCF[7]，其使用多通道的 HOG 特征来取代 CSK 中的单通道特征，同时利用核函数对多通道 HOG 特征进行融合，降低了计算量，当时该算法无论是在精度方面还是在速度方面都处于领先地位。具有代表性的相关滤波算法还有 SAMF[8]和 DSST[9]等。但是，这类方法在面对复杂场景时容易受到外界环境的干扰，导致跟踪结果不准确。

相比相关滤波的目标跟踪方法，基于深度学习的目标跟踪[14-32]在跟踪性能上有了很大的提高，提取的特征具有更强的表征能力，能够提升算法的跟踪性能，但是算法的运行速度一直不太理想。Wang 等[14]于 2013 年提出了首个将深度学习应用于目标跟踪领域的算法 DLT(Deep Learning Tracker)，该算法首先通过离线预训练来获得通用的目标表征能力，然后模型参数会在跟踪时利用第一帧图像信息进行进一步微调。该方法在一定程度上解决了目标跟踪时的样本匮乏问题，但其跟踪速度平均只有 15FPS。Wang 等[17]对类似的 FCNT(Tracking with Fully Convolutional Networks)算法在目标跟踪任务中的性能做了深入的分析之后，将该网络应用到了目标跟踪领域并达到了理想的跟踪效果，但是该跟踪算法的跟踪速度只有几帧每秒，远达不到实时跟踪的要求。而在深度学习领域中，基于孪生网络的目标跟踪算法[19-23,30]在跟踪精度和跟踪速度上取得了很好的平衡，在大量数据集上都取得了优异的性能，为后来的研究者开辟了新的研究方向。基于孪生网络的跟踪算法将跟踪视为一个相似度匹配任务，在超大规模的数据集上离线学习关于目标的通用特征，在线跟踪时将初始帧作为模板，在后续帧中寻找与初始帧最相似的区域作为目标的当前位置。文献[33]把视觉目标跟踪问题看作一类特殊的目标检测问题，通过元学习策略初始化网络参数，并离线训练网络模型，从单个图像中学习新实例，预测目标位置，完成跟踪。Siam R-CNN 算法[34]从 Tracking by Re-Detection 的角度提出了一种新颖的 Re-detector，将 Faster-RCNN 融入到孪生网络结构中，通过在一个图像中的任何位置对目标进行重新检测，来判断给定的建议性目标区域是否属于同一个物体，然后对该物体进行位置的回归。该类方法相比其他基于深度学习的目标跟踪算法速度较快，但是由于算法没有在线更新机制，因此当目标表观发生变化时容易发生目标漂移和丢失。

Danelljan 等[35]提出了 ATOM(Accurate Tracking by Overlap Maximization)算法，将目标跟踪分为目标分类和目标估计两个网络部分。目标估计网络离线训练 IoUNet 网络，用于粗定位；目标分类网络使用深度回归网络结构，用于精细定位。文献[36]提出了一种由源域网络和目标域网络组成的视觉跟踪自编码对模型，通过源域中的少量样本微调网络参数，再跟踪目标域的数据，该方法能够应对目标遮挡、光照变化等挑战，从而稳定地跟踪目标。

3 本文提出的目标跟踪算法

经典的全卷积孪生网络目标跟踪算法，如 SiamFC，由两

个分支网络组成，分别输入模板图像和搜索图像。该算法采用全卷积网络提取两个分支输入图像的特征，然后利用相似性度量函数进行相似度匹配，得到得分响应图，再通过线性插值的方式将得分响应图恢复到原来的大小，最后根据得分值最大的位置确定跟踪目标位置。该方法原理简单，速度快，但缺点在于其是同等地对待图像不同维度的特征。然而，目标跟踪面对的是不同的跟踪对象，图像不同维度的深度特征对于目标跟踪的重要性是不一样，该方法使得算法缺少对每个特定目标的关注。

为了弥补上述算法的不足，本文将时空注意力机制引入到孪生网络中，利用空间注意力机制和通道注意力机制来关注特定的目标特征，针对不同的跟踪对象，对不同的特征赋予不同的权重来表示其重要性。此外，基于相似度匹配策略的全卷积孪生网络目标跟踪算法在跟踪时仅仅使用第一帧的图像数据作为先验知识，这种简单的目标表观特征不足以持续保证跟踪器的性能，在跟踪过程中当目标遇到形变、遮挡等挑战性因素时，算法的跟踪性能会显著下降。另外，本文算法还加入了自适应更新机制，通过选择跟踪过程中置信度较高的图像帧，来弥补随着跟踪的进行第一帧图像特征判别力不足的缺点。

3.1　算法框架

本文提出的目标跟踪算法框架如图 1 所示。该网络由模板分支和搜索分支组成，两者分别接收第一帧的模板图像和待跟踪的第 T 帧搜索图像。我们将模板分支接收的第一帧模板图像记为 z，搜索分支接收的第 T 帧搜索图像记为 x，特征提取网络记为 φ。图 1 中，第 2 帧和第 n 帧表示在算法更新过程中用于弥补第一帧图像特征的高置信度跟踪图像，框架图中×表示点乘操作，＋表示特征相加，* 表示卷积操作。

在跟踪过程中，首先将第一帧模板图像送入特征提取网络和注意力机制网络进行特征提取，后续每一帧搜索图像经过特征提取网络得到的特征图都与模板图像利用卷积操作进行相似性计算，根据得到的响应图可以确定最终的目标位置。第一帧模板图像的特征提取操作在一段视频跟踪中只需要进行一次。

本节将依次介绍通道注意力机制和空间注意力机制、算法的训练跟踪过程以及自适应更新方法，最后介绍算法的实验对比结果和可视化结果。

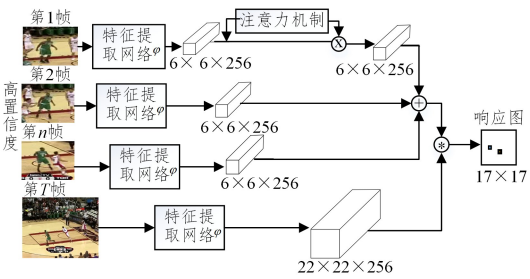


图 1　基于时空注意力机制的目标跟踪算法架构图

Fig. 1　Framework of object tracking algorithm based on temporal-spatial attention mechanism

3.2　通道注意力机制

传统的卷积神经网络提取图像特征进行目标跟踪的算法是同等地对待每一个通道内的图像特征。然而，实际上特征图的不同通道蕴含着不同的信息，对于目标跟踪算法的重要性也各不相同。本文将卷积神经网络与通道注意力机制相结合，对不同的通道赋予不同的权重，使得网络能够关注有利于目标跟踪的通道内特征，忽略其他无用的信息，对不同的目标获得自适应特征。通道注意力机制的结构如图 2 所示。

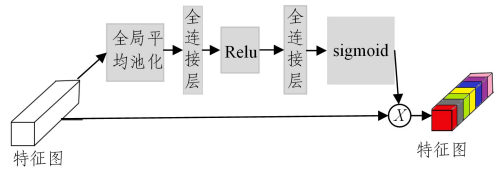


图 2　通道注意力机制结构图

Fig. 2　Structure of channel attention mechanism

通道注意力机制网络接受模板分支图像 z 经过特征提取网络 φ 提取到的图像特征 $\varphi(z)$。假定其特征大小为 $W\times H\times C$，首先特征经过全局平均池化层（Global Average Pooling）进行特征压缩，输出大小为 $1\times1\times C$ 的特征向量；然后，经过两层全连接层以及相应的激活函数分别进行降维和升维操作，使用两层全连接层可以使得网络具有更好的非线性，能够更好地拟合数据，同时可以减少网络参数量并降低计算复杂度，全连接层后得到大小为 $1\times1\times C$ 的特征向量，再利用 Sigmoid 函数来生成每个通道的权重 α；最后，通道注意力机制网络的输入图像特征 $\varphi(z)$ 的每个通道分别乘以权重 α 对应通道的权重值。通过该操作，图像特征的不同通道被赋予不同的权重，用于表示该通道对目标跟踪效果的重要性。通道注意力机制的使用，使得有利的通道特征得到增强，而没有太大作用的通道特征则被抑制。此外，注意力机制网络参数是利用图像数据离线训练得到的，跟踪时参数固定，每一次视频跟踪仅将模板分支的第一帧图像送入通道注意力机制网络以计算各通道的权重，之后第一帧图像特征与通道权重通过点乘操作得到经过注意力机制选择后的图像特征，利用选择后的图像特征与待跟踪图像进行模板匹配，从而确定待跟踪目标的位置。由于每一次跟踪只需将跟踪视频第一帧图像送入注意力机制网络以计算各通道权重，所需要的计算量很小，因此，该方法不会降低算法的运行速度。

3.3　空间注意力机制

除了通道注意力机制，本文还引入了空间注意力机制。一幅图像中不仅包含前景目标，还存在着大量的背景信息，但图像中的所有区域对跟踪任务的贡献并非是同样重要的，只有任务相关的区域才是最需要关注的。

空间注意力机制的原理正是寻找图像中对特定目标跟踪任务最重要的部分进行重点关注，通过将原始图像的特征转换到另一个空间来保留关键信息，使得特征图中不同的空间位置具有不同的权重。空间注意力机制的结构如图 3 所示。

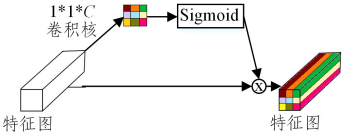


图 3　空间注意力机制结构图

Fig. 3　Structure of spatial attention mechanism

空间注意力机制网络同样接受特征提取网络 φ 提取到的

图像特征 $\varphi(\boldsymbol{z})$，特征的大小为 $W\times H\times C$。首先，利用 $1\times 1\times C$ 的卷积核对特征进行降维，得到大小为 $W\times H\times 1$ 的特征图，本文中 C 的大小为 256。然后，将得到的特征图通过 Sigmoid 函数进行归一化，得到特征图中每一个空间位置的权重 β，β 表示特征图中每一个位置对于目标跟踪的重要性。最后，将每个位置的权重 β 与原始的图像特征 $\varphi(\boldsymbol{z})$ 相乘，得到目标的响应图。

与通道注意力机制类似，空间注意力机制的网络参数也是通过离线训练得到的。跟踪时参数固定，将第一帧图像送入空间注意力机制网络以获得特征图每个位置的权重，再通过加权的第一帧图像特征与待跟踪图像进行模板匹配，最大响应值的位置即为跟踪目标的位置。

3.4 基于时空注意力机制的目标跟踪算法

3.4.1 网络离线训练阶段

基于时空注意力机制的目标跟踪算法在离线训练过程中接受一对图像 $\boldsymbol{z}$ 和 $\boldsymbol{x}$ 作为输入，分别代表模板图像和搜索图像，图像 $\boldsymbol{x}$ 的大小为 $127\times 127\times 3$，图像 $\boldsymbol{z}$ 的大小为 $255\times 255\times 3$。其中，3 表示图像的通道数，两者经过特征提取网络 φ 后分别得到大小为 $6\times 6\times 256$ 和 $22\times 22\times 256$ 的特征图 $\varphi(\boldsymbol{z})$，$\varphi(\boldsymbol{x})$，接着分别将模板图像 $\varphi(z)$ 送入通道注意力机制网络和空间注意力机制网络进行特征选择，两个网络会根据输入的图像特征产生相应的权重 α 和 β，该权重用于决定各个特征对目标跟踪的重要性，最终得到经过加权的特征图 $h(\boldsymbol{z})$，该过程可以表示为：

$$h(z)=\varphi(z)\cdot\alpha\cdot\beta \tag{1}$$

然后，网络将 $h(\boldsymbol{z})$ 作为卷积核对 $\varphi(x)$ 进行滑动卷积操作，此处的卷积操作相当于利用相似度函数对 $h(\boldsymbol{z})$ 和 $\varphi(\boldsymbol{x})$ 进行相似度匹配，网络最终输出得到大小为 17×17 的得分响应图 $f(\boldsymbol{z},\boldsymbol{x})$。整个过程可以表示为：

$$f(z,x)=h(z)*\varphi(x) \tag{2}$$

其中，$*$ 代表卷积操作，$f(\boldsymbol{z},\boldsymbol{x})$ 表示响应图得分。

本文在网络的整个训练过程中定义的损失函数如下：

$$l(y,v)=\log(1+\exp(-yv)) \tag{3}$$

其中，v 表示网络得到的响应图 $f(\boldsymbol{z},\boldsymbol{x})$ 上每个点的值，$y\in\{+1,-1\}$ 表示响应图上每个点的真实标签值，$+1$ 代表正样本，-1 代表负样本。通过对损失函数的不断优化，可以得到最终的网络模型。

3.4.2 目标跟踪阶段

将待跟踪视频序列的第一帧模板图像经过模板分支特征提取网络 φ 进行特征提取，随后送入注意力机制网络中进行特征权重的计算，从而获得不同位置和通道特征的重要性，将通道注意力机制网络和空间注意力机制网络生成的权重与原始特征相加，得到模板分支最终的图像特征，然后与后续帧的搜索图像进行卷积操作，得到 17×17 像素的响应图得分，再将响应图通过双三次线性插值方法采样得到大小为 272×272 的图像，响应图中最大响应值为目标位置。

为了适应目标尺度的变化，本文设置了 3 个搜索尺度，分别为 $\{1.025^{-1},1,1.025\}$。跟踪时，将模板图像和 3 个不同尺度的搜索图像分别送到网络模板分支和搜索分支，从而得到 3 种不同尺度对应的响应图，最大值所在的位置以及对应的响应图的尺度大小即为新一帧中目标的位置和尺度的大小。

3.4.3 在线更新

传统的全卷积孪生网络目标跟踪算法仅仅使用第一帧图像提取的特征与后续待跟踪图像帧进行相似度匹配。由于目标在跟踪过程中可能会遭受形变、遮挡、相似背景干扰等挑战，仅使用第一帧的特征难以解决上述问题，甚至会出现目标漂移现象。为此，本文引入了在线更新机制。通过 *PSR* 方法计算跟踪目标的图像置信度，再利用置信度对已获得的图像帧进行选择，选择出置信度较高的图像来更新目标模板。*PSR* 的计算式如下：

$$PSR=\frac{g_{\max}-\mu_{s1}}{\sigma_{s1}} \tag{4}$$

其中，$g_{\max}$ 表示响应图得分的最大值；μ_{s1} 是旁瓣的均值；σ_{s1} 表示旁瓣的标准差。旁瓣指响应图最大值周围大小为 11×11 的窗口。正常情况下，*PSR* 值在 20～60 之间表示高置信度，当 *PSR* 降到 7 时，意味着目标被遮挡或者目标跟丢了。本文采用线性加权更新方法来完成目标模板的更新，如式(5)所示：

$$\varphi(z)=\varphi(z)\eta+(1-\eta)\varphi(x) \tag{5}$$

其中，η 表示更新比例，本文设置为 0.7；$\varphi(\boldsymbol{z})$ 表示模板图像特征；$\varphi(\boldsymbol{x})$ 表示高置信度的搜索图像特征。

4 实验结果

4.1 参数设置和数据集

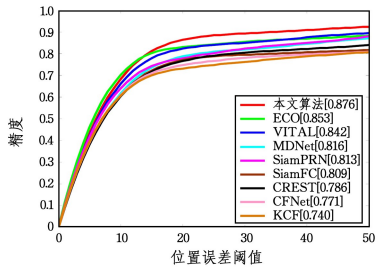
本节将介绍算法的实现细节，并将所提跟踪算法与基准数据集上的最新跟踪算法进行比较以进行性能评估。

本文采用 2015 年 ILSVRC 竞赛中用于视频目标检测的视频数据集作为训练数据，该数据集共包含 4 500 段视频，超过 100 万张标注图像。网络训练的每一个图像对都是从同一段视频中间隔不超过 10 帧的图像中提取的，每一个图像对都是以目标为中心进行提取。

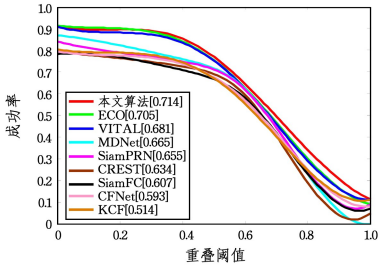
本文采用 AlexNet 网络来进行特征提取，网络由 5 个卷积层组成，卷积层的大小分别为 11×11，5×5，3×3，3×3，3×3，卷积核的个数分别为 96，256，384，384，256。在第一个卷积层和第二个卷积层以及第二个卷积层和第三个卷积层之间加入了大小为 3×3、步长为 2 的最大值池化层，每个卷积层后使用 relu 激活函数来增加网络的非线性。此外，通道注意力机制网络中两个全连接层中分别具有 64 个神经元和 256 个神经元。网络通过随机梯度下降的方式进行训练，网络的初始参数符合高斯分布，使用 50 个 epoch 进行训练，每个 epoch 的包含 50 000 个样本，每个 batch 的大小为 8，网络学习率为 0.01。本文算法在 Quadro K2200 GPU，E5 2.4 GHz CPU 和 MatConvNet 工具箱的开发环境下使用 MATLAB 实现。

4.2 定量评估

为了验证本文算法的性能，在 OTB2013 数据集和 OTB2015 数据集上与当前的主流算法进行了对比：KCF[7]，SAMF[8]，CREST[15]，CFNet[12]，ECO[16]，SiamFC[20]，SiamRPN[23]，MDNet[25]，VITAL[26] 和 Staple[24]。评价指标为精度图和成功率图，图 4 和图 5 中精度图和成功率图分别表示当中心误差和重叠率等于 20 和 0.6 时各种主流算法的得分值。



(a)OPE 精度图



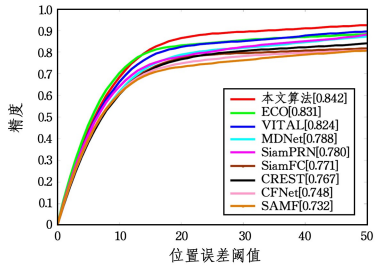
(b)OPE 成功率图

图 4　在 OTB2013 数据集上各跟踪算法的性能比较

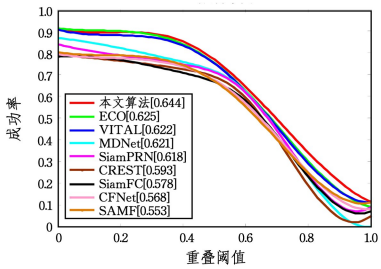
Fig. 4　Performance comparison of different tracking algorithms on OTB2013 dataset

从图 4 可以看出，本文算法在 OTB2013 数据集上的精度和成功率分别为 0.876 和 0.714，比 SiamPRN 算法的 0.813 和 0.655 分别高了 0.063 和 0.059。在 OTB2015 数据集上(见图 5)，本文算法的精度和成功率分别为 0.842 和 0.644，比 SiamPRN 算法的 0.780 和 0.618 分别高了 0.062 和 0.026。结果表明，基于时空注意力机制的目标跟踪算法的性能优于其他主流的方法。

另外，与其他主流的目标跟踪算法如 Staple 和 KCF 相比，本文算法的性能更优。由于上述这两种算法使用的是 HOG 和 CN 等手工设计的特征，不能很好地描述目标的特点，而本文提出的目标跟踪算法提取的是经过卷积神经网络和注意力机制的重选择特征。



(a)OPE 精度图



(b)OPE 成功率图

图 5　在 OTB2015 数据集上不同跟踪算法的性能比较

Fig. 5　Performance comparison of different tracking algorithms on OTB2015 dataset

从图 4 和图 5 也可以发现，由于 SiamRPN 算法引入了区域候选网络模块，让跟踪器直接回归位置、形状，省掉了多尺度测试。ECO 算法利用因式分解卷积操作对特征提取进行了简化，使用了更优的多分辨率特征融合方法。与主流的 ECO 和 SiamRPN 目标跟踪算法相比，本文提出的基于时空注意力机制的目标跟踪算法的性能仍有较大的提高。

另外，在跟踪速度方面，当目标跟踪速度大于 25 帧每秒时即可称为实时跟踪。SiamFC 算法的跟踪速度为 86 fps，而所提算法的跟踪速度为 70 fps。

4.3　定性评估

为了能够直观地感受目标跟踪算法的实际跟踪效果，我们从 OTB2015 数据集的 100 段视频中选择了 6 个具有代表性的视频帧进行可视化展示，这些视频中包含了绝大多数目标跟踪场景中遇到的挑战，如尺度变化、光照变化、目标遮挡、背景干扰等，通过利用不同颜色的目标跟踪框将本文算法与其他参与比较的主流目标跟踪算法的跟踪效果在同一幅图像中表示出来，能够更直观地分析这些算法的跟踪性能。

图 6 给出了本文算法与 SiamFC，SiamPRN，Staple 算法在视频帧中的实际跟踪效果。

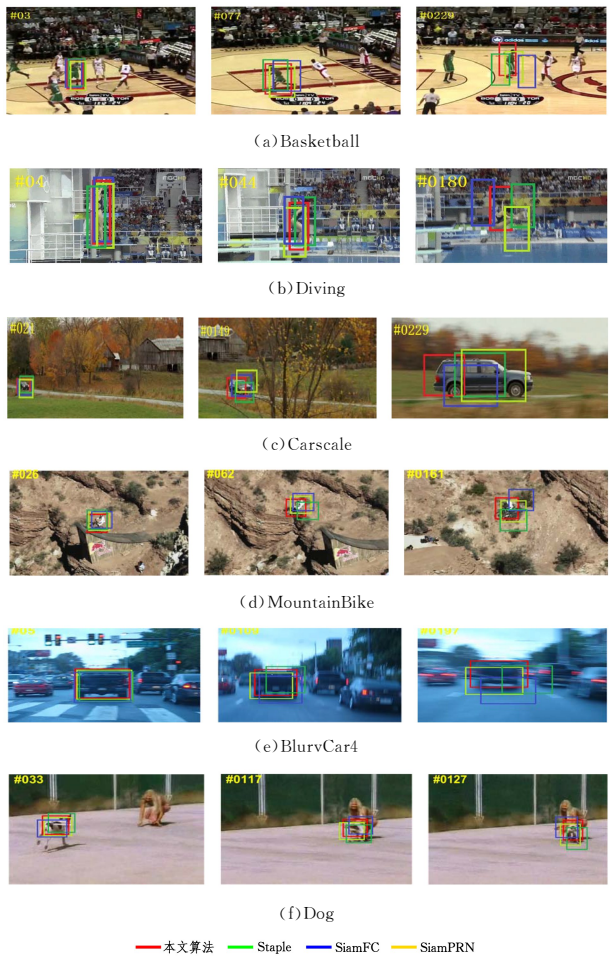


图 6　基于时空注意力机制的目标跟踪算法与 SiamFC，SRDCF 和 Staple 算法的跟踪结果

Fig. 6　Tracking results of object tracking algorithm based on temporal-spatial attention mechanism，SiamFC，SRDCF and Staple

从图 6 可以看出，由于本文算法在传统的孪生网络目标

跟踪算法的基础上引入了时空注意力机制和模型更新等技术，因此整体性能优于当前主流的目标跟踪算法。特别是对于 MotorRolling 视频序列，尽管视频中存在着平面内旋转、背景干扰等挑战，在其他几种算法都跟踪失败的情况下，本文算法仍然能够稳定地跟踪目标。在 Basketball 和 BlurCar4 视频序列中，由于目标存在遮挡、运动模糊等挑战，只有本文算法和 SiamPRN 算法的跟踪效果较好，而其他算法相继丢失目标。这主要归功于所提算法融入了时空注意力机制，使得对目标跟踪有利的特征得到了更多的关注，从而增强了图像特征的判别力，不会在目标外观发生变化时跟踪失败。在 Carscale 和 Dog 视频序列中，目标车辆和宠物狗都经历了由小到大或由大到小的尺度变化，几种跟踪算法在效果上的差异不大，这主要是由于本文算法加入了 3 种不同的搜索尺度，使得算法对于目标的尺度变化具有鲁棒性。在视频 Diving 和 MountainBike 中，由于目标经历了平面内旋转，其他算法都出现了目标漂移的现象，只有本文算法依然表现稳定。在视频序列 Human3 中，所有目标跟踪算法在第 83 帧时都能够跟踪到目标，并能保持一定的精度，但是当跟踪到第 108 帧时，目标经历了背景杂斑挑战（即目标附近的背景具有与目标相似的颜色或纹理），导致 SiamFC 和 STAPLE 算法漂移到了另一个相似的目标上，此时本文算法仍然能够有效地对目标进行跟踪。定性的评估表明了本文提出的基于时空注意力机制的目标跟踪算法具有很好的鲁棒性。

4.4 消融实验

从表 1 可以观察到，与 SIT 模型相比，在最初的孪生网络跟踪 SIT 模型[19]的基础上添加时空注意力模块，本文算法获得的跟踪效果大大提升。这说明使用时空注意力机制有助于改善跟踪器的性能，同时会产生更加平滑的跟踪结果。与 SIT 模型[19]相比，目标表观模板的在线更新机制可显著改善最终结果。在跟踪过程中使用在线更新机制，可以及时捕获到目标表观的变化，从而降低目标发生漂移的风险，在大多数情况下会产生更好的结果。然而，目标模板更新的频率不能过于频繁，频繁地更新目标模型会引入噪声，导致模板中混入背景信息，进而造成跟踪失败。

表 1 OTB2015 数据集上的消融实验结果

Table 1 Ablation experimental results on OTB2015 dataset

跟踪模型	精度	成功率
SINT[19]	0.607	0.588
SINT+在线更新（每帧更新一次）	0.691	0.594
SINT+在线更新（每 5 帧更新一次）	0.693	0.605
SINT+在线更新（每 10 帧更新一次）	0.691	0.614
SINT+在线更新（每 20 帧更新一次）	0.785	0.608
SINT+时空注意力	0.794	0.623
本文算法	0.842	0.644

结束语 本文提出了一种基于时空注意力机制的目标跟踪算法，并在现有孪生网络的基础上，提出了改进的时空注意力机制，用于提高目标特征的判别力。此外，利用在线更新机制提高算法的跟踪效果。最后，在 OTB2013 和 OTB2015 数据集上对本文算法和其他主流的目标跟踪算法进行了对比分析，实验结果验证了本文算法的高效性。

参 考 文 献

[1] YI W, LIM J, YANG M H. Online object tracking: A benchmark [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 2411-2418.

[2] WU Y, LIM J, YANG M H. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[3] FIAZ M, MAHMOOD A, JAVED S, et al. Handcrafted and deep trackers: Recent visual object tracking approaches and trends [J]. ACM Computing Surveys (CSUR), 2019, 52(2): 1-44.

[4] LI P X, WANG D, WANG L J, et al. Deep visual tracking: Review and experimental comparison [J]. Pattern Recognition, 2018, 76: 323-338.

[5] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters [C]// The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 2544-2550.

[6] 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 (ECCV). 2012: 702-715.

[7] HENRIQUES J F, CASEIRO R, BATISTA J. High speed tracking with kernelized correlation filters [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[8] LI Y, ZHU J. A scale adaptive kernel correlation filter tracker with feature integration [C]// European Conference on Computer Vision (ECCV). 2014: 254-265.

[9] DANELLJAN M, HÄGER G, KHAN F, et al. Accurate scale estimation for robust visual tracking [C]// Proceeding of the British Machine Vision Conference (BMVC). 2014.

[10] DANELLJAN M, ROBINSON A, KHAN F S, et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking [C]// European Conference on Computer Vision (ECCV). 2016: 472-488.

[11] HUANG B, XU T, LI J, et al. Transfer learning-based discriminative correlation filter for visual tracking [J]. Pattern Recognition, 2019, 100: 107157.

[12] 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 (CVPR). 2017: 2805-2813.

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

[14] WANG N, YEUNG D Y. Learning a deep compact image representation for visual tracking [C]// The Annual Conference on Neural Information Processing Systems (NIPS). 2013: 809-817.

[15] SONG Y, MA C, GONG L, et al. CREST: Convolutional residual learning for visual tracking [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2555-2563.

[16] DANELLJAN M，BHAT G，KHAN F S，et al. ECO：Efficient Convolution Operators for Tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017：6638-6646.

[17] WANG L，OU Y W，WANG X，et al. Visual tracking with fully convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015：3119-3127.

[18] MA C，HUANG J，YANG X，et al. Hierarchical Convolutional Features for Visual Tracking[C]// Proceedings of the IEEE International Conference on Computer Vision(ICCV). 2015：3074-3082.

[19] TAO R，GAVVES E，SMEULDERS A W，et al. Siamese Instance Search for Tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016：1420-1429.

[20] LUCA B. Fully-convolutional Siamese networks for object tracking[C]// European Conference on Computer Vision(ECCV). 2016：850-865.

[21] 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 (CVPR). 2019：4591-4600.

[22] WANG X，LI C，LUO B，et al. SINT++：Robust visual tracking via adversarial positive instance generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2018：4864-4873.

[23] 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 (CVPR). 2018：8971-8980.

[24] 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(CVPR). 2016：1401-1409.

[25] 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(CVPR). 2016：4293-4302.

[26] SONG Y，MA C，WU X，et al. VITAL：Visual tracking via adversarial learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2018：8990-8999.

[27] CHENG X，ZHANG Y，ZHOU L，et al. Visual Tracking via Auto-Encoder Pair Correlation Filter[J]. IEEE Transactions on Industrial Electronics，2020，67(4)：3288-3297.

[28] PARK E，BERG A C. Meta-tracker：Fast and Robust Online Adaptation for Visual Object Trackers[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：587-604.

[29] FAN H，LING H. Parallel Tracking and Verifying[J]. IEEE Transactions on Image Processing，2019，28(8)：4130-4144.

[30] LI B，WU W，WANG Q，et al. SiamRPN++：Evolution of Siamese Visual Tracking with Very Deep Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019：4282-4291.

[31] WANG Q，ZHANG M，XING J，et al. Do not Lose the Details：Reinforced Representation Learning for High Performance Visual Tracking[C]// International Joint Conference on Artificial Intelligence(IJCAI). 2018：985-991.

[32] LUKEZIC A，MATAS J，KRIATAN M. D3S - A Discriminative Single Shot Segmentation Tracker[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020：7133-7142.

[33] WANG G，LUO C，SUN X，et al. Tracking by Instance Detection：A Meta-Learning Approach[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020：6288-6297.

[34] VOIGTLAENDER P，LUITEN J，PHILIP T H S，et al. Siam R-CNN：Visual Tracking by Re-Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020：6578-6588.

[35] 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(CVPR). 2019：4660-4669.

[36] CHENG X，ZHANG Y，ZHOU L，et al. Visual Tracking via Auto-Encoder Pair Correlation Filter[J]. IEEE Transactions on Industrial Electronics，2020，67(4)：3288-3297.

CHENG Xu，born in 1983，Ph.D，associate professor. His main research interests include computer vision，object tracking and pattern recognition.