Computer Science and Application 计算机科学与应用, 2021, 11(5), 1468-1473
Published Online May 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.115150

Hans汉斯

基于Siamese网络的目标跟踪研究

方长江，林俊杰

中南民族大学电子信息工程学院，湖北 武汉
Email: 1354406850@qq.com

收稿日期：2021年4月25日；录用日期：2021年5月20日；发布日期：2021年5月27日

摘 要

视觉目标跟踪是计算机视觉领域一个重要研究方向，在自动驾驶、视频监控、人机交互、医疗诊断等众多领域有着广泛的应用。随着深度学习的崛起，基于神经网络的视觉目标跟踪已成为主流研究方向，其中基于双胞胎(Siamese)网络模型的方法在目标跟踪领域表现出了优异的性能。本文将基于Siamese网络，探究不同干扰和网络结构对目标跟踪性能的影响。

关键词

视觉目标跟踪，深度学习，双胞胎网络

Research on Object Tracking Based on Siamese Network

Changjiang Fang, Junjie Lin

College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan Hubei
Email: 1354406850@qq.com

Received: Apr. 25th, 2021; accepted: May 20th, 2021; published: May 27th, 2021

Abstract

Visual object tracking is an important research in computer vision field, with a range of applications such as autonomous driving, video surveillance, human-computer interaction, medical diagnosis, etc. With the rise of deep learning, neural networks have been employed in the mainstream frameworks for visual object tracking. Among them, the methods built on architecture of Siamese networks have shown excellent tracking performance. In this paper, we will investigate the effects of different interference and network structures on target tracking performance based on Siamese networks.

文章引用: 方长江, 林俊杰. 基于 Siamese 网络的目标跟踪研究[J]. 计算机科学与应用, 2021, 11(5): 1468-1473.
DOI: 10.12677/csa.2021.115150

Keywords

Visual Object Tracking, Deep Learning, Siamese Networks

1. 引言

视觉目标跟踪(Visual Object Track, VOT)是计算机视觉领域中一个重要的研究方向，一般在初始帧中给定感兴趣目标的初始状态信息，如目标边界框、质心等，通过设计算法逐帧标定出感兴趣目标区域[1]。视觉目标跟踪在自动化军事装备、汽车自动驾驶、视频监控、人机交互和医疗诊断系统等相关领域具有广泛的应用[2] [3] [4]。

2016 年，Bertinetto [5]等人提出了一个全卷积双胞胎(Siamese FC)网络，该网络采用离线训练的方式在大规模数据集上训练了一个共享权重的神经网络，在跟踪阶段，对输入的模板图像和待搜索图像进行裁剪等预处理后，利用离线训练的网络分别提取特征，然后用一个全连接层计算网络提取特征的相似得分，最后根据得分计算目标位置。在跟踪过程中，预训练的跟踪模型需要应对实际场景中诸多问题，如目标的遮挡、尺度变化、运动、光照变化、背景杂乱等挑战[1]。为了研究在复杂场景下各种因素的具体影响和网络结构对跟踪性能的影响，我们以 Siamese 网络为基础，进行了一系列实验。

2. Siamese 网络结构框架

Siamese 网络如图 1 所示，

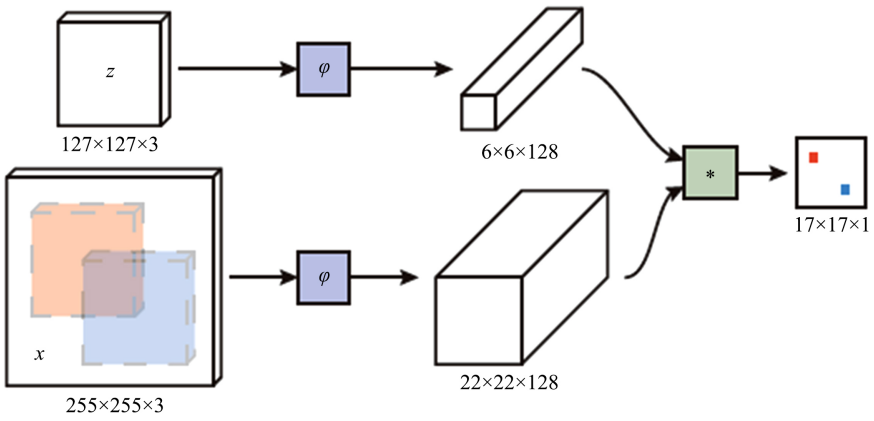

Figure 1. Siamese network framework diagram
图 1. Siamese 网络框架图

设模板图像样本为 z_k，候选区域样本实例为 x_k，相似度函数为 $f(z_k, x_k)$，定义为：如图 1 所示，Siamese 网络有两个输入，一个是模板 z，一个是搜索区域 x。具体来说，z 代表模板图像，以图像中的目标为中心裁剪得到，x 代表后续帧中的候选框搜索区域，φ 代表将原始图像映射到特定空间的特征映射操作，特征映射操作采用的网络以 AlexNet [6]为基础，并去掉了全连接层，采用全卷积结构。选用 AlexNet 作为

Siamese 网络的骨干网络，是因为 AlexNet 网络较 VggNet [7]浅层一些，跟踪任务需要浅层的外观特征，并且在实时性上有要求，较浅层的网络在跟踪速度上相对来说会快一些。将模板图像和搜索区域通过特征映射网络，得到新的空间表示，然后通过学习一个相似度函数，计算模板与搜索区域中相同大小区域的相似度，最终得到一张 17 × 17 大小的得分图，得分图上数值最大的点就代表着目标的预测位置，将其映射回原图就得到了目标的估计位置。设模板图像样本为 z_k，候选区域样本实例为 x_k，相似度函数为 $f(z_k, x_k)$，定义为：

$$f(z_k, x_k) = \varphi(z_k) * \varphi(x_k) + b1 \tag{1}$$

其中，φ 表示特征映射，*表示互相关运算，$b1$ 是偏置项。

3. 基于 Siamese 网络跟踪算法实现

本章采用 ILSVRC2015-VID 数据集离线训练跟踪模型，该数据集包含了 30 个不同类别的动物、车辆和人，训练集和验证集有 4500 个视频，总共有超过 100 万个注释帧。

在训练和跟踪阶段，都采用 127 × 127 大小的模板图像和 255 × 255 大小的搜索图像，不同的是，训练阶段模板图像是在序列中随机挑选的一帧，搜索图像是距离该帧 50 帧以内的某一帧，然后根据 ground-truth 以目标为中心进行裁剪。跟踪阶段模板图像是初始帧，并手动框出目标框，再以整个目标框进行裁剪，搜索图像是当前帧，并以上一帧的跟踪结果即目标的估计位置为中心进行裁剪。由于目标可能位于图像边缘，所以对图像进行了填充，并设置了一个缩放因子，使边界框加上填充余量后具有固定的面积。假设 Bounding box 的大小为 (w, h)，填充余量为 p，缩放因子为 s，对图像进行填充缩放后的面积等于一个常数 A，它们满足关系：

$$s(w + 2p) \times s(h + 2p) = A \tag{2}$$

在实验中，设定模板图片的面积 $A = 127^2$，并将填充余量设置为 $p = \dfrac{(w+h)}{4}$。

在训练时，采用判别式方法，用搜索图像的正负样本对网络进行训练，并采用逻辑损失：

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

其中，v 是单个候选区域的实际得分，$y \in \{+1, -1\}$ 是对应区域的真值。正负样本的判定是根据得分图中样本点与中心点的距离来判定的。具体来说，若样本点与得分图中心点的距离在一个半径为 R 的圆内，则该样本点划分为正样本，否则，划分为负样本。

$$y[u] = \begin{cases} +1 & \text{if } k\|u - c\| \le R \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

通过利用全卷积 Siamese 网络比较模板图像和一个更大的搜索图像，生成一个得分图 $v: D \to R$，得分图的损失定义为得分图所有位置的损失的平均值：

$$L(y, v) = \frac{1}{|D|} \sum_{u \in D} \ell(y[u], v[u]) \tag{5}$$

得分图中每个位置 $u \in D$ 的真实标签 $y[u] \in \{+1, -1\}$，采用随机梯度下降(SGD)对网络中的参数进行优化，最小化损失函数以得到，即：

$$\arg\min_{\theta} \boldsymbol{E}_{(z,x,y)} L(y, f(z, x; \theta)) \tag{6}$$

训练轮数设置为 50，每轮采用 10,000 个图像对，大小为 8 的批次进行迭代训练，学习率从 10^{-5} 到 10^{-2}。

4. 实验结果分析

当使用 Siamese 网络进行跟踪时，从第一帧得到目标的位置框，初始化跟踪模板，且在之后的帧中保持固定，利用得到的模板对后续帧中的目标进行跟踪。我们采用 OTB-2013 [8]验证集进行测试。实验结果如图 2 所示。

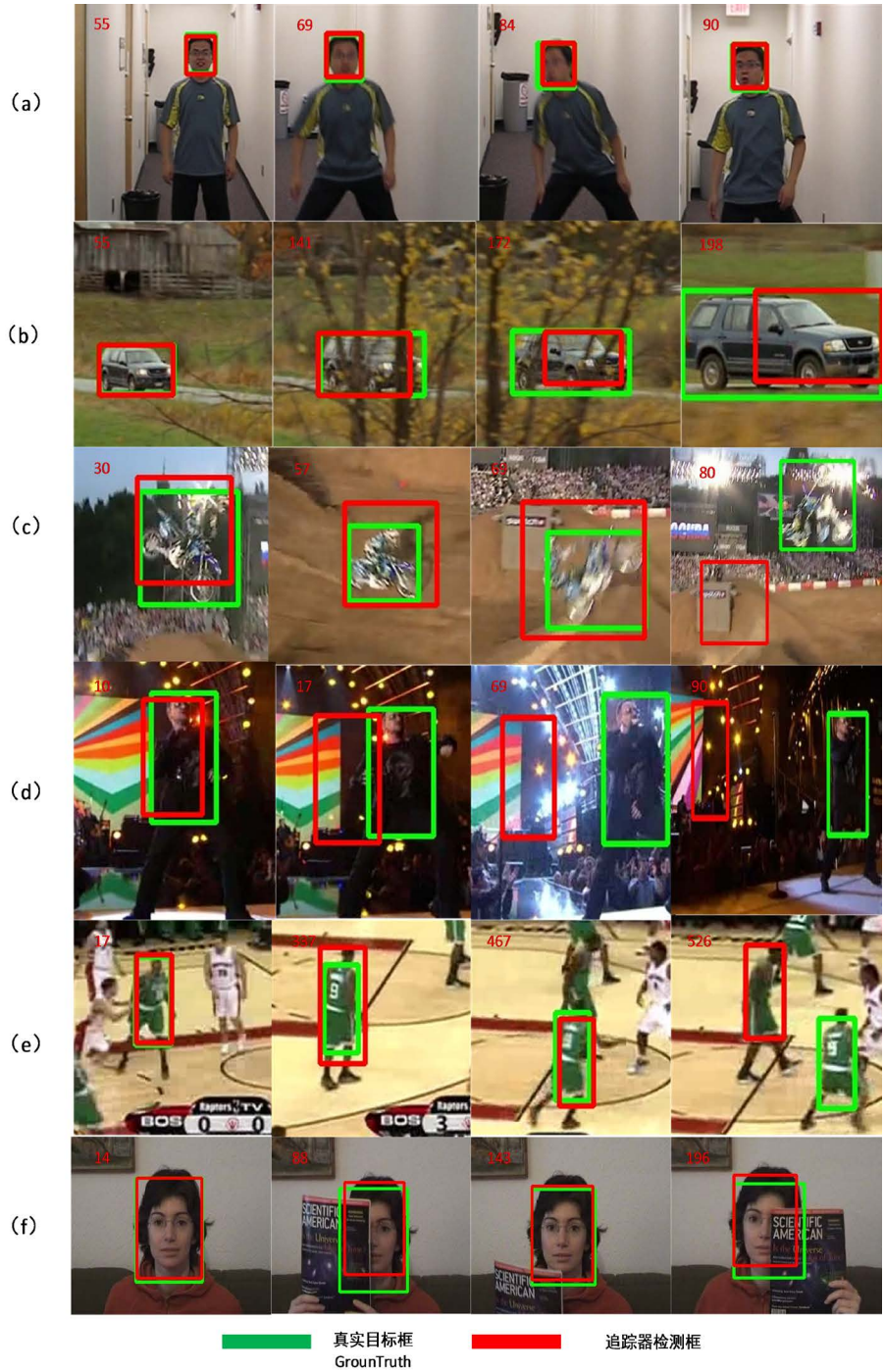

Figure 2. Siamese network tracking experiment results
图 2. Siamese 网络跟踪实验结果

图 2 中，红色表示跟踪器得到的跟踪框，绿色的为真实目标框，即 ground truth。在图 2(a)中视频的主要干扰因素为目标小范围的晃动，背景信息不变；在图 2(b)中视频的主要干扰因素是短暂性的局部遮挡；在图 2(c)中视频主要干扰因素为运动模糊；在图 2(d)中视频的主要干扰因素是光照差异且背景复杂，对比度不明显；在图 2(e)中视频的主要干扰因素是相似的外观；在图 2(f)中视频的主要干扰因素是遮挡。

在跟踪过程中，对于图 2 中(a)、(b)、(f)三种干扰情况，跟踪器几乎能做到非常精准地跟踪；面对图 2 中(c)、(d)、(e)等几种干扰情况下时，跟踪器很容易丢失目标，从而导致整个跟踪过程丢失。

由于 Siamese 网络采取第一帧的目标跟踪框作为整个视频的跟踪模板，并在整个跟踪过程中不更新目标模板与网络的权值，因此，当受干扰帧的目标候选框和跟踪模板相差较大时将导致跟踪失败，而网络权值不更新导致难以适应所有场景，所以当目标受到小幅度晃动、暂时的局部遮挡等对目标外观影响较小和光照等对场景影响较小的干扰因素时，网络能够有较好的进行跟踪，但是当目标一旦发生快速运动、外观相似等对当前帧目标外观影响较大和背景复杂、对比度不明显等对背景影响较大的干扰时，由于没有及时地更新目标模板和网络参数，可能导致搜索图像无法覆盖目标，从而使最后的相似度结果为错误的，随着跟踪过程中发生的错误的累加，将导致跟踪不可恢复。

为了更好地探究 Siamese 网络的特性，我将 Siamese 网络的基础网络 AlexNet 用 VggNet 进行替换，并通过改变最后一层网络的神经元感受野(RF)、步长(STR)、特征输出尺寸(OFS)等条件进行试验，实验结果如表 1 所示，其中，原始网络中输入图片为 127 × 127，RF = 87，STR = 8，OFS = 6。

Table 1. Correspondence table of Siamese network influencing factors
表 1. Siamese 网络影响因素对应表

RF	127	111	103	103	103	95	87	87	79
STR	8	8	8	16	4	8	8	8	8
OFS	1	3	4	2	7	5	6	16	7
Siamese (AlexNet)	0.561	0.572	0.582	0.553	0.585	0.587	0.589	0.553	0.579
Siamese (VggNet)	0.569	0.574	0.578	0.542	0.581	0.591	0.592	0.561	0.580

如表 1 所可以有如下结论：

1) 当网络步幅(STR)从 4 增加到 8 或者 16 时，性能明显的下降，说明 Siamese 网络在跟踪时更适应中级特征(步长 4 或者 8)，这些特征在目标定位时能使性能更精准。

2) 当感受野(RF)覆盖输入样本 60%~80%的像素大小时效果最佳，大的感受野虽然能覆盖更多图像信息，但是会导致提取的特征对目标的位置不敏感，而感受野太小将会无法捕捉目标足够的外观信息，因此，需要根据输入的大小进行选择。

3) 当输出特征尺寸(OFS)小于 3 时，由于特征太小缺乏目标的空间结构描述，因此在计算图像相似度中不稳健，因此，跟踪精度下降。

5. 结语

本文以 Siamese 网络为基础，针对复杂场景下不同影响因素进行了实验探索，研究了不同网络结构下的算法性能。实验结果表明，Siamese 网络采用第一帧作为目标模板所以小幅度的形变、暂时的遮挡、运动模糊等干扰对跟踪的影响较小，但是受于剧烈运动导致的形变、长时间遮挡和外观相似等干扰因素的影响较大；Siamese 网络的神经元感受野、步长、特征输出尺寸等条件对 Siamese 跟踪框架的影响：Siamese 网络更适应中级的神经元感受野、步长、特征输出尺寸。

参考文献

[1] Wu, Y., Lim, J. and Yang, M.H. (2013) Online Object Tracking: A Benchmark. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Portland, OR, 23-28 June 2013, 2411-2418. https://doi.org/10.1109/CVPR.2013.312

[2] 黄凯奇，陈晓棠，康运锋，谭铁牛. 智能视频监控技术综述[J]. 计算机学报, 2015, 38(6): 1093-1118.

[3] 高文，朱明，贺柏根，吴笑天. 目标跟踪技术综述[J]. 中国光学, 2014, 7(3): 365-375.

[4] 王鑫，徐立中. 图像目标跟踪技术[M]. 北京：人民邮电出版社, 2012.

[5] Bertinetto, L., Valmadre, J., Henriques, J.F., *et al.* (2016) Fully-Convolutional Siamese Networks for Object Tracking. *European Conference on Computer Vision* (*ECCV*), Springer, Cham, 850-865. https://doi.org/10.1007/978-3-319-48881-3_56

[6] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *International Conference on Neural Information Processing Systems* (*NIPS*), Lake Tahoe, 3-6 December 2012, 1097-1105.

[7] Girshick, R., Donahue, J., Darrell, T., *et al.* (2016) Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*), **38**, 142-158. https://doi.org/10.1109/TPAMI.2015.2437384

[8] Yi, W., Lim, J. and Yang, M.H. (2013) Online Object Tracking: A Benchmark. 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 23-28 June 2013, 2411-2418.