DOI: 10.11992/tis.201807029
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20181218.1639.009.html

联合外形响应的深度目标追踪器

孙海宇, 陈秀宏, 肖汉雄
(江南大学 数字媒体学院, 江苏 无锡 214122)

摘　要: 针对追踪器使用卷积网络提取出来的特征模板进行目标位置匹配时, 易产生响应噪声的问题, 本文提出一种联合外形响应和卷积响应的深度目标追踪方法。在当前帧中, 由前一帧提供的目标信息先分别提取卷积特征和外形信息, 然后获得相应的卷积位置响应和外形位置响应; 最后利用外形位置响应对卷积位置响应进行修正, 从而有效地抑制响应噪声。实验表明: 这种方法具有较高的位置精度, 能够提高目标跟踪的准确性。
关键词: 目标追踪; 神经网络; 卷积特征; 相关滤波; 位置响应; 外形信息; 噪声抑制; 修正; 深度学习
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2019)04−0725−08

中文引用格式: 孙海宇, 陈秀宏, 肖汉雄. 联合外形响应的深度目标追踪器 [J]. 智能系统学报, 2019, 14(4): 725–732.
英文引用格式: SUN Haiyu, CHEN Xiuhong, XIAO Hanxiong. A deep object tracker with outline response map[J]. CAAI transactions on intelligent systems, 2019, 14(4): 725–732.

A deep object tracker with outline response map

SUN Haiyu, CHEN Xiuhong, XIAO Hanxiong
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: When convolutional neural network is used as a template to locate target, noise may be unavoidable in the final location response. To solve this problem, we developed a deep object tracker by combining the convolutional position response with the outline position response. For example, in the current frame, after extracting convolutional features and the outline information from the predicted target in the previous frame, we obtained the corresponding convolutional position response and the outline position response, and the latter was used to rectify the former in controlling the noise generated in the convolutional position response. The favorable results of our deep tracker on the benchmark show that the method of integrating the outline position response into the convolutional position response can greatly improve the precision and accuracy of the tracker.
Keywords: object tracking; neural network; convolutional features; correlation filter; position response; outline information; noise suppression; rectify; deep learning

单目标跟踪是一项基础而又重要的计算机视觉任务。通常所讲的单目标跟踪是指: 在视频的首帧, 给定目标的初始状态 (如: 位置、大小), 然后在视频的后续帧中估计出目标的状态[1]。估计一个对象的运动轨迹[2] 可以达到目标跟踪的目的, 但是目标轨迹的估计在多种干扰因素的影响下易有较大的误差。常见的干扰因素有: 光照变化 (illumination variation, IV)、大小变化 (scale variation, SV)、遮挡 (occlusion, OCC)、变形 (deformation, DEF)、运动模糊 (motion blur, MB)、快速运动 (fast motion, FM)、平面内旋转 (in-plane rotation, IPR)、平面外旋转 (out-of-plane rotation, OPR)、部分显示 (out-of-view, OV)、背景杂乱 (background clutters, BC)、目标像素过少 (low-resolution, LR) 等[3]。实际上, 目标跟踪就是要在当前帧中确定与目标相关的两大要素: 位置以及大小。有很多方法可以实现该目的, 其中, 检测就是一种比较流行的方法。若在基于检测的目标跟踪方法中采用前背景分类器, 这种追踪方法又称作基于判别式模型[4]

收稿日期: 2018−07−26.　**网络出版日期**: 2018−12−20.
基金项目: 江苏省研究生科研与实践创新计划项目 (1232050205185680)
通信作者: 孙海宇. E-mail: 6161610009@vip.jiangnan.edu.cn.

的追踪方法。判别式追踪器充分利用了视频序列中每一帧的前背景信息，从而达到区分目标和背景的目的，如：Henriques 等[5] 所提出的追踪器以及 Danelljan 等[6] 提出的追踪器，从某种意义上讲，以上追踪器亦可称为模板匹配类的追踪器。这类追踪器主要通过已知的目标信息，习得一个与目标相关的滤波模板，然后使用该模板在搜索区域 (可能包含目标的区域) 进行滑动匹配，以匹配度的形式来反应匹配区域是否是目标位置，并将匹配度最高的位置作为最佳目标位置。文献 [7] 中，主要讨论的是样本采样问题，当样本的数目采集的越多时，这些样本会构成一种理论比较完善的循环结构，故而提出了循环采样的方法。文献 [5] 中，基于文献 [7]，使用样本的原始像素或者方向梯度直方图作为样本特征求解模板，然后使用该模板进行目标位置的匹配。从跟踪的定义上来看，以上主要解决的是目标追踪中的位置问题。文献 [6] 中，基于文献 [5]，主要讨论了目标尺度问题，在求解尺度以及位置的时候，使用的是方向梯度直方图作为特征。类似的追踪器还有文献 [8-9]，而最近几年，深度学习在许多应用领域表现出了优秀的成绩[10]，目标追踪领域也不例外，自 Wang Naiyan 将深度学习算法应用到跟踪领域[11] 后，深度学习类的追踪器也涌现出不少优秀的作品，文献 [12] 中，提出了基于孪生网络的深度追踪器，其在利用目标信息提取出卷积特征之后，仅仅使用了卷积特征在搜索区域进行目标位置的匹配。文献 [13] 在文献 [12] 的基础上直接将相关滤波转化为了神经网络的网络层，使得两者合二为一，成为一个端到端的整体系统，但是依旧仅仅使用卷积网络所提取出来的特征。在最近的研究中，孪生网络受到了极大的关注。文献 [14] 提出了一种动态孪生网络通过在线学习目标的外形变化以获得目标的时域信息；文献 [15] 结合 RPN 改进了候选框的生成来提高追踪的精度；文献 [16] 则尝试通过改变训练样本的数据分布来获得更具有判别性的特征；文献 [17] 基于 FaceNet 中的 Triplet Loss 来改进损失函数。本文认为，网络卷积出来的特征其实是碎片化的，故而在进行目标位置匹配的时候，易在远离目标的地方产生位置响应噪声，从而影响目标位置的确定。因此，本文设计了一种联合外形响应的深度目标追踪器，利用外形信息的位置响应来修正卷积的位置响应，从而得到更准确的目标定位，并通过实验与其他追踪器进行了比较，验证了思路的可行性。

1　模板匹配以及外形信息

模板匹配类的追踪器，基本思想是通过衡量目标与搜索区域内各个部分的相似度，选取相似度最大的位置作为目标位置。事实上，这种思想在基于深度学习的追踪器[12] 和基于相关滤波的追踪器[5] 中均有体现。

1.1　孪生网络追踪器

在深度追踪器中，模板的匹配大都是通过孪生网络来实现的，Bertinetto 等[12] 第一次将孪生网络应用于目标追踪。图 1 展示了一个典型的孪生网络结构图，f 表示已经训练好的、权值固定的卷积网络；$\boldsymbol{Z}$ 表示上一帧的目标信息；$\boldsymbol{X}$ 表示搜索区域；$\otimes$ 表示相关操作，实践中具体表现为卷积运算。当 $\boldsymbol{Z}$ 与 $\boldsymbol{X}$ 分别经过卷积网络提取特征之后，分别得到卷积特征，$\boldsymbol{Z}_f$ 以及 $\boldsymbol{X}_f$，通过计算两者之间的相关性可得到搜索区域中每个部分与目标之间的相关程度响应 (图中的红点表示搜索区域中红色部分与目标的相关程度，而蓝点则表示搜索区域中蓝色部分与目标的相关程度)，当得到响应图之后，通过三线性插值，最大响应值的位置便可作为当前帧的目标位置。

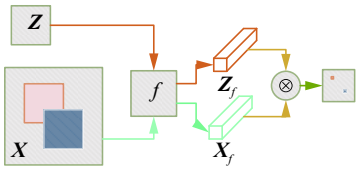


图 1　孪生网络追踪器结构图

Fig. 1　Siamese network tracker structure

1.2　相关滤波追踪器

相关滤波类的追踪器最早是由文献 [18] 提出来的，文献 [7] 在其基础上发展了循环采样以及引入了核方法，但在文献 [19] 中，将目标位置和大小分开考虑，与本文将追踪中两大要素分而治之的思想更为契合，因此为本文所选用。为了获取目标位置的响应图，相关滤波类的追踪器要找到一个最优化的滤波器 $\boldsymbol{H}$，该滤波器由 $\boldsymbol{H}^i$ 所构成。$\boldsymbol{H}$ 通过最小化如下的代价函数获得：

$$l=\left\|\sum_{i=1}^{d}\boldsymbol{H}^i\cdot\boldsymbol{F}^i-\boldsymbol{G}\right\|^2+\lambda\sum_{i=1}^{d}\left\|\boldsymbol{H}^i\right\|^2 \tag{1}$$

式中：$\boldsymbol{H}^i$ 为特征的第 i 维度上的滤波器；$\boldsymbol{F}^i$ 为特征的第 i 维度上的值；$\boldsymbol{G}$ 是对应的位置响应标签，由峰值点在目标中心处的高斯函数获得；并且 $\boldsymbol{H}^i$、$\boldsymbol{F}^i$、$\boldsymbol{G}$ 都是 $M\times N$ 大小的；$\cdot$ 代表了循环卷积运算；$\lambda\geqslant 0$，为正则项系数，易得该式的解为

$$\boldsymbol{H}^i=\frac{\mathcal{G}_*\circ\mathcal{F}^i}{\sum\limits_{k=1}^{d}\mathcal{F}_*^k\circ\mathcal{F}^k+\lambda} \tag{2}$$

式中: $\mathcal{G}$、$\mathcal{F}$ 代表了式 (1) 中 $\boldsymbol{G}$、$\boldsymbol{F}$ 的离散傅里叶变换, 下标 $*$ 代表取该对象的复共轭, $\circ$ 表示哈达玛积。将式 (2) 中 $\mathcal{G}_*\circ\mathcal{F}^i$ 记作 $\boldsymbol{A}_t^i$ (t 表示第 t 时刻), 将分母中加号前的部分记作 $\boldsymbol{B}_t$, 那么第 t 时刻的分子分母可用式 (3) 和式 (4) 近似更新, 而第一个时刻的 $\boldsymbol{H}$ 可由式 (2) 获得。

$$\boldsymbol{A}_t^i=(1-\alpha)\boldsymbol{A}_{t-1}^i+\alpha\mathcal{G}_*^t\circ\mathcal{F}_t^i \tag{3}$$

$$\boldsymbol{B}_t=(1-\alpha)\boldsymbol{B}_{t-1}+\alpha\sum_{k=1}^{d}\mathcal{F}_{t*}^k\circ\mathcal{F}_t^k \tag{4}$$

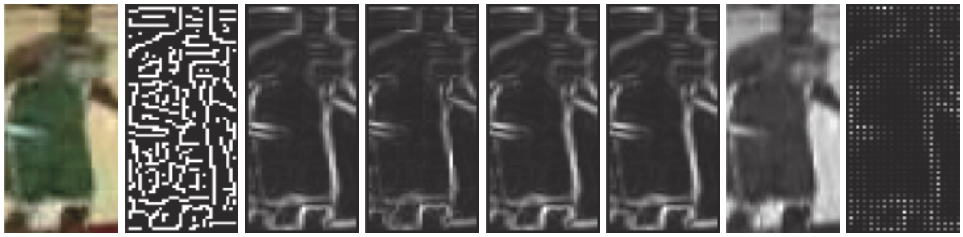
式中 α 为学习率, 则新的目标位置响应可计算得到:

$$\boldsymbol{P}=D^{-1}\left\{\frac{\sum_{i=1}^{d}\boldsymbol{A}_*^i\circ\boldsymbol{Z}^i}{\boldsymbol{B}+\lambda}\right\}=D^{-1}\{\boldsymbol{H}_*\circ\boldsymbol{Z}\} \tag{5}$$

式中: $\boldsymbol{P}$ 表示位置响应图, 响应值最大的位置为目标的位置; D^{-1} 为离散傅里叶变换的逆变换; $\boldsymbol{Z}$ 是新的一帧中的特征, 与式 (1) 中的 $\boldsymbol{F}$ 相对应。

1.3　目标外形信息

在人类视觉中, 目标的外形信息具有重要的意义, 倘若缺少了目标的外形信息, 人类就会产生‘一叶障目’的视觉障碍, 因此在过去的几十年中, 学者们对于目标的外形信息有着大量的研究[20-23]。目标的外形信息一般存在于目标与背景之间, 能够有效地突出目标物, 为双眼提供一个良好的聚焦区域。由于目标与背景在外形上存在较大的差异, 因而学者们常常使用微分的方式来检测目标的外形信息, 常见的方式有 Canny、Sobel、Roberts、Scharr、Prewitt、Hog, 此外, 还有小波变换等方式, 其检测效果如图 2 所示。



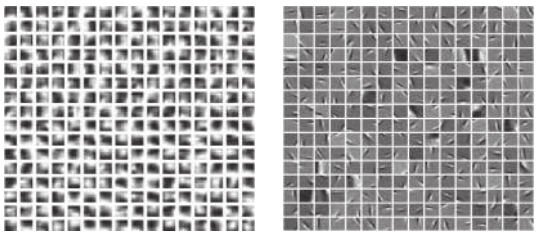
(a) 原图　(b) Canny　(c) Sobel　(d) Roberts　(e) Scharr　(f) Prewitt　(g) Gabor　(h) Hog

图 2　外形信息检测方法

Fig. 2　The samples of methods to detect outlines

2　联合外形响应的深度目标追踪器

本文在可视化卷积网络特征后, 观察到其与文献 [24] 所述的局部性、方向性的特征具有相似性后 (如图 3), 更加验证了本文联合外形信息的想法。图 3(a) 为卷积网络所提取的特征, 图 3(b) 为稀疏编码所求得的基, 可以发现图 3(a) 的特征与未完全处理的基具有很高的相似性。如前所述, 提取目标外轮廓信息有相当多的方法, 鉴于方向梯度直方图在追踪问题上的广泛应用, 本文选择使用方向梯度直方图来提取外形信息。



(a) 卷积网络特征　(b) 局部性、方向性特征

图 3　特征对比

Fig. 3　Features comparison

2.1　网络结构

本文提出的联合外形响应的深度目标追踪器的结构如图 4 所示, 它由两个部分组成, 一个是由卷积网络所构成的位置匹配部分, 这个部分主要使用卷积网络提取的目标特征进行位置匹配, 称之为卷积匹配部分, 另一个是利用外形信息使用相关滤波进行位置匹配, 称之为滤波匹配部分。

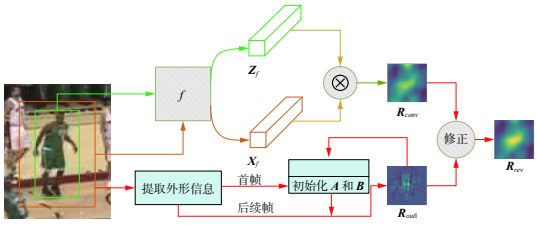


图 4　本文目标追踪器结构

Fig. 4　The architecture of the proposed deep tracker

2.1.1　卷积匹配部分

在追踪器的卷积匹配部分仅使用卷积网络提取的特征进行目标位置的匹配, 同时融合了尺度考虑, 同文献 [12] 中所表示的那样, 本文仅使用了 3 种尺度, 追踪器的输入同文献 [12] 一样是一对样本, 一个是在初始帧中标记出来的目标, 用 $\boldsymbol{Z}$ 来表示, 其维度是 $\boldsymbol{H}_z\times\boldsymbol{W}_z\times3$, 另一个是在当前帧中, 以上一帧目标中心为中心的包含背景的目标

搜索区域，用 $\boldsymbol{X}$ 表示，其维度是 $\boldsymbol{H}_x \times \boldsymbol{W}_x \times 3$，将两者通过权值固定的卷积网络 f，f 是文献 [25] 中所提出的 AlexNet 网络 (不包含全连接层)，提取出对应的卷积特征 $\boldsymbol{Z}_f$ 和 $\boldsymbol{X}_f$ 之后，通过相关操作得出目标的位置响应。

2.1.2　滤波匹配部分

在追踪器的滤波匹配部分，主要使用方向梯度直方图来提取目标的外形信息。由于目标的外形信息常常存在于目标与背景之间，为了提取目标的外形信息，需要包含一些背景信息，因此本文直接在 $\boldsymbol{X}$ 中进行目标的外形信息提取。如果当前帧是首帧的话，就使用式 (2) 初始化滤波匹配时所需要的 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，如果当前帧不是首帧的话，就利用式 (5) 求得目标的位置响应，然后再利用式 (3) 和式 (4) 更新 $\boldsymbol{A}$ 和 $\boldsymbol{B}$。

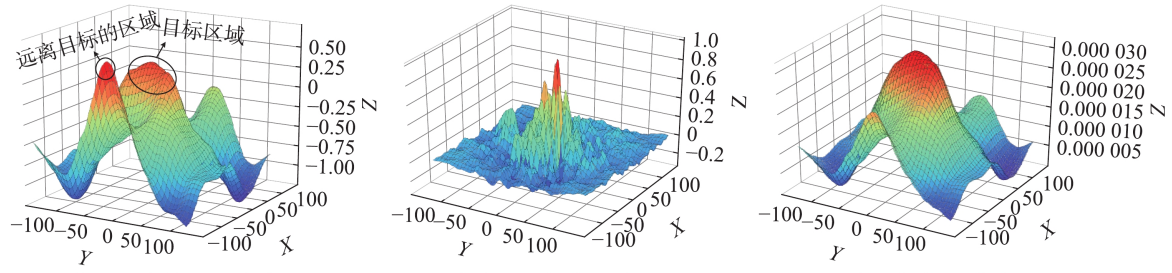
2.2　修正部分

在没有使用外形信息对位置响应进行修正的情况下，由于卷积特征中多是类似于局部的、方向性的特征，这种碎片化的特征容易导致在远离目标的区域处产生极大的位置响应点，从而形成位置响应噪声。如图 5 中卷积匹配的位置响应图所示 (X、Y 轴无十分重要的物理意义，Z 轴表示相关性程度，数值越大表示该处与目标的相关性越大，则颜色越红)。相对的，在搜索区域中，目标的外形占比一定大于局部特征的占比，所以利用了外形信息的滤波匹配的位置响应大多会集中在目标区域处，如图 5 中滤波匹配的位置响应图所示，联合该外形信息的位置响应，可以有效地突出目标的所在区域，使得位置响应集中在目标区域处，从而达到抑制噪声，避免位置发生漂移的目的。如图 5 中修正的位置响应图所示，可以看到目标的位置响应图在修正后，抑制住了左边的噪声响应。联合该外形位置响应，涉及到数据融合技术，可以使用加权平均法、贝叶斯估计法[26]、卡尔曼滤波法[27] 等，本文为验证想法直接采用了最为简单的加权平均法：

$$\boldsymbol{R}_{\text{rev}} = \eta \boldsymbol{R}_{\text{conv}} + (1-\eta)\boldsymbol{R}_{\text{outl}} \tag{6}$$

式中：η 是权重值；$\boldsymbol{R}_{\text{conv}}$ 是卷积匹配的位置响应；$\boldsymbol{R}_{\text{outl}}$ 是利用外形信息的滤波匹配的位置响应；$\boldsymbol{R}_{\text{rev}}$ 是最终经过修正之后的位置响应。三者之间的关系可参考图 4，由于修正前 $\boldsymbol{R}_{\text{conv}}$ 和 $\boldsymbol{R}_{\text{outl}}$ 的量纲不一致，故而在修正前，先分别对两个位置响应进行如下的归一化处理：

$$\boldsymbol{R}_{\text{norm}} = \frac{\boldsymbol{R}_{\text{old}}}{\max(\boldsymbol{R}_{\text{old}})} \tag{7}$$



(a) 卷积匹配的位置响应　(b) 滤波匹配的位置响应　(c) 修正的位置响应

图 5　位置响应的变化

Fig. 5　The transform of position response map

3　实验配置以及评价指标

3.1　实验细节

卷积匹配部分，网络的构成同文献 [12] 一样 (去除了网络的全连接部分)，权重的取值是经过 405 650 次随机梯度下降得到的，网络训练的数据集是从 ILSVRC-2015 视频数据集[29] 中提取出的 4 417 个视频序列，网络训练的损失函数为

$$l_{\text{train}} = N^{-1}\sum_{i,j}\boldsymbol{G}_{ij}\times\left[-\ln\left(\text{sigmoid}\left(\boldsymbol{R}_{\text{conv}}^{ij}\right)\right)\right] + (1-\boldsymbol{G}_{ij})\times\left[-\ln\left(1-\text{sigmoid}\left(\boldsymbol{R}_{\text{conv}}^{ij}\right)\right)\right] \tag{8}$$

式中：N 为 $\boldsymbol{G}$ 中的元素个数；i、j 为 $\boldsymbol{G}$ 中元素索引；sigmoid 函数为：

$$\text{sigmoid}(x) = \frac{1}{1+\exp(-x)} \tag{9}$$

训练时，网络的输入是一对样本，并且，执行一次梯度下降使用 8 对样本，一对样本中，$\boldsymbol{Z}$ 的维度是 $127\times127\times3$，$\boldsymbol{X}$ 的维度是 $255\times255\times3$，经过网络 f 后，$\boldsymbol{Z}_f$ 的维度是 $6\times6\times256$，$\boldsymbol{X}_f$ 的维度是 $22\times22\times256$，$\boldsymbol{R}_{\text{conv}}$ 经过三线性插值后，维度为 255×255，学习率采用动态学习率，初始值为0.01，然后使用如下的指数衰减法进行衰减：

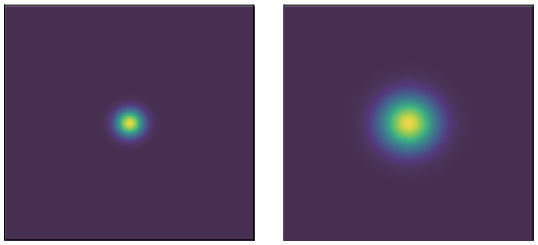
$$l = l_{\text{init}} \times r^{\frac{s_{\text{global}}}{s_{\text{decay}}}} \tag{10}$$

式中：l_{init} 是初始学习率；r 是初始学习率的基本保留率，本文中取值为 0.868 5；s_{global} 表示当前是第几次梯度下降；s_{decay} 表示经过几次梯度下降

后, 执行一次学习率的衰减, 本文取值 6 650。滤波匹配部分不需要进行训练, 其中, 构造目标位置响应标签的高斯函数:

$$g = \exp\left(\frac{x^2+y^2}{2\sigma^2}\right) \tag{11}$$

式中: g 是 $\boldsymbol{G}$ 中的元素; x、y 表示目标位置, σ 控制了响应结果中与目标相关的范围大小, 如图 6 所示, 本文的取值为 16。为了使得输入的数据依旧保持 3 个维度, 外形信息的提取方法同文献 [5] 中一样, 使用文献 [28] 中所使用的多通道方向梯度直方图, 然后按照 2.1.2 节所述方法进行目标位置的匹配, 修正部分的权重值 η 是通过二分法确定的, 本文最终的取值为 0.967 8。



(a) σ=10　　(b) σ=20

图 6　σ 效果图

Fig. 6　The effect of σ

3.2　实验设备

本文使用 Tensorflow[30] 框架来实现模型, 版本为 1.4.0, 实验是在配置为 i5-7300HQ 2.5 GHz CPU, GeForce GTX1050 GPU 的笔记本中运行的。

3.3　测试数据集以及评价指标

3.3.1　指标

目标追踪需要解决两个问题: 位置和大小, 因此, 评价一个追踪器的优劣往往通过精度图和成功图来描述[4]。精度图是指在不同的中心误差下, 目标追踪的成功率所构成的图; 成功图是指在不同的重叠率下, 目标追踪的成功率所构成的图。其中, 中心误差是指: 追踪器所输出的目标框的中心与标签目标框的中心之间的误差, 常用欧氏距离表示, 单位是像素; 重叠率 o 的定义为

$$o = \frac{\boldsymbol{A}_{\text{track}} \cap \boldsymbol{A}_{\text{groundtruth}}}{\boldsymbol{A}_{\text{track}} \cup \boldsymbol{A}_{\text{groundtruth}}} \tag{12}$$

式中: $\boldsymbol{A}_{\text{track}}$ 表示由追踪器所追踪到目标区域; $\boldsymbol{A}_{\text{groundtruth}}$ 表示标签所表示的目标区域。对于某一帧而言, 如果该帧的中心误差小于某个阈值或者重叠率大于某个阈值, 则认为追踪器在该帧上的追踪结果是可靠的、成功的, 而一个视频序列中, 成功帧数的占比, 称之为成功率。

3.3.2　测试数据集

本文使用目标追踪测试平台 (object tracking benchmark, OTB)[1,3] 中 CVPR-2013、OTB-50 和 OTB-1003 个数据集对现有的算法进行评估。这 3 个数据集分别有 51、50 和 100 个视频序列, 每个视频序列都包含了 IV、SV、OCC、DEF、MB、FM、IPR、OPR、OV、BC、LR 中的多个干扰因素。在这些干扰因素的影响下, 追踪测试平台统计追踪器的成功率, 以精度图和成功图的形式来反应追踪器的追踪性能。

3.3.3　外形信息的时间花费

本文在不同分辨率图上测试了提取外形信息以及获得其对应的位置响应所花费的时间, 如图 7 所示。分别给出了提取外形信息的时间花费和获取外形信息对应的位置响应所花费的时间。从图中可以看出, 外形信息的提取时间相对较为合理 (和图像分辨率之间接近线性关系); 同时, 获取位置响应的时间相对较长, 有待进一步改进。

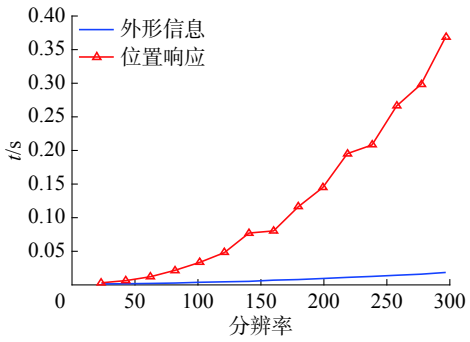


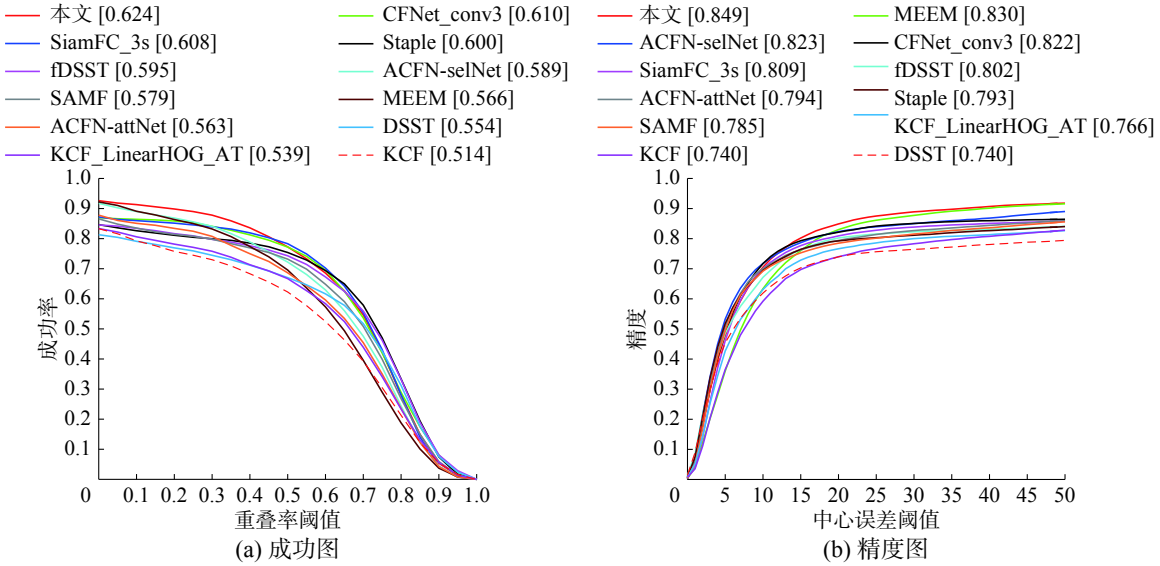
图 7　时间花费

Fig. 7　The results of elapsed time

3.4　平台测试结果

本文在目标追踪测试平台上和近几年优秀的追踪器 CFNet_conv3[13]、SiamFC_3s[12]、Staple[8]、fDSST[7]、ACFN-selNet[31]、SAMF[9]、LCT[33]、MEEM[32]、ACFN-attNet[31]、DSST[19]、KCF[5] 进行了比较, 其结果如图 8 所示, 这里仅给出了 CVPR-2013 的结果, 并利用图 9, 直观展示了部分追踪效果 (更多的结果数据请访问文献 [34]), 图 8(a) 代表数据集的成功图结果, 图 8(b) 代表数据集的精度图结果。从图 8(b) 的结果中可以看出, 在中心误差阈值很大的情况下, 本文追踪器依旧有着优秀的成功率, 说明在追踪的过程中, 本文的追踪器发生了较少的边框漂移, 反应到图 8(a) 的成功图中, 可以看到, 在重叠率阈值很小的情况下, 本文追踪器的成功率依旧优秀, 而很多追踪器的成功率却不理想, 说明他们在追踪的过程中, 发生了较多的边框漂移现象, 导致了目标丢失; 而本

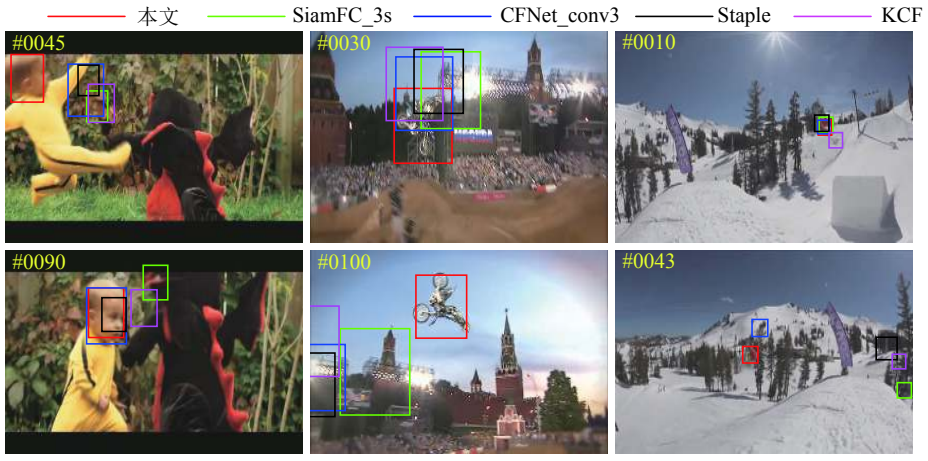
文追踪器在进行位置确定的时候，利用了外形信息来抑制原本位置响应中的噪声点，所以具有较少的边框漂移现象。以上的结果表明本文的追踪器具有优秀的追踪效果。



(a) 成功图　　(b) 精度图

图 8　测试结果

Fig. 8　The results of object tracking benchmark



(a) DragonBaby　　(b) MotorRolling　　(c) Skiing

图 9　追踪效果直观感受

Fig. 9　The direct feeling of our tracker

3.5　直观效果

在追踪测试平台的测试序列中，每个序列都包含了多个干扰因素。从这么多的测试序列中取得优秀的追踪效果是相当不容易，由于本文使用了外形信息来对目标的位置响应进行噪声抑制，所以从上面的测试平台给出的追踪结果可知，本文的追踪器具有优秀的追踪能力，但由于篇幅限制，这里给出几组具有代表性的视频序列追踪效果的直观展示，如图 9 所示，正红为本文追踪器。

4　结束语

本文尝试从理解卷积特征的基础上来理解目标追踪中卷积位置响应的结果，从而指导如何修正目标跟踪中的卷积响应。通过分析可知：卷积网络抽离出的卷积特征类似于局部性、方向性的特征，是碎片化的，在进行位置匹配的时候，可以通过突出目标区域的方式来缓和这种碎片化特征的影响。和最近几年优秀的追踪器相比，该思路具有一定的可行性，能够有效提高目标位置定位的精度。接下来的工作可以进一步探究如何缩短位置响应的时间；本文卷积网络的许多特征之间具有很高的相似性，是否可以直接通过稀疏化的方式来实现抑制位置响应中的噪声也是值得研究的。

参考文献:

[1] WU Yi, LIM J, YANG M H. Online object tracking: a benchmark[C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2411–2418.

[2] 杨戈, 刘宏. 视觉跟踪算法综述 [J]. 智能系统学报, 2010, 5(2): 95–105.
YANG Ge, LIU Hong. Survey of visual tracking algorithms[J]. CAAI transactions on intelligent systems, 2010, 5(2): 95–105.

[3] WU Yi, LIM J, YANG M H. Object tracking benchmark[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1834–1848.

[4] 管皓, 薛向阳, 安志勇. 深度学习在视频目标跟踪中的应用进展与展望 [J]. 自动化学报, 2016, 42(6): 834–847.
GUAN Hao, XUE Xiangyang, AN Zhiyong. Advances on application of deep learning for video object tracking[J]. Acta automatica sinica, 2016, 42(6): 834–847.

[5] 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, 2015, 37(3): 583–596.

[6] DANELLJAN M, HÄGER G, KHAN F S, et al. Discriminative scale space tracking[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(8): 1561–1575.

[7] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy, 2012: 702–715.

[8] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: complementary learners for real-time tracking[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1401–1409.

[9] LI Yang, ZHU Jianke. A scale adaptive kernel correlation filter tracker with feature integration[C]//Proceedings of European Conference on Computer Vision. Zurich, Switzerland, 2014: 254–265.

[10] ALOM M Z, TAHA T M, YAKOPCIC C, et al. The history began from alexNet: a comprehensive survey on deep learning approaches[J] arXiv: 1803.01164, 2018.

[11] WANG Naiyan, YEUNG D Y. Learning a deep compact image representation for visual tracking[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 809–817.

[12] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 850–865.

[13] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5000–5008.

[14] GUO Qing, FENG Wei, ZHOU Ce, et al. Learning dynamic Siamese network for visual object tracking[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1781–1789.

[15] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with Siamese region proposal network[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8971–8980.

[16] ZHU Zheng, WANG Qiang, LI Bo, et al. Distractor-aware Siamese networks for visual object tracking[C]//Proceedings of European Conference on Computer Vision. Munich, Germany, 2018: 103–119.

[17] DONG Xingping, SHEN Jianbing. Triplet loss in Siamese network for object tracking[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 472–488.

[18] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. California, USA, 2010: 2544–2550.

[19] DANELLJAN M, HÄGER G, KHAN F S, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of the 25th British Machine Vision Conference. Linköping, Sweden, 2014: 1–5

[20] CANNY J. A computational approach to edge detection[J]. IEEE transactions on pattern analysis and machine intelligence, 1986, PAMI-8(6): 679–698.

[21] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. California, USA, 2005: 886–893.

[22] DERICHE R. Using Canny's criteria to derive a recursively implemented optimal edge detector[J]. International journal of computer vision, 1987, 1(2): 167–187.

[23] ELDER J H, ZUCKER S W. Local scale control for edge detection and blur estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(7): 699–716.

[24] OLSHAUSEN B A, FIELD D J. Emergence of simple-

cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381(6583): 607–609.

[25] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097–1105.

[26] BLEIHOLDER J, NAUMANN F. Data fusion[J]. ACM computing surveys (CSUR), 2009, 41(1): 1.

[27] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of basic engineering, 1960, 82(1): 35–45.

[28] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(9): 1627–1645.

[29] RUSSAKOVSKY O, DENG Jia, SU Hao, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211–252.

[30] ABADI M, BARHAM P, CHEN Jianmin, et al. Tensorflow: a system for large-scale machine learning[C]//Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. Savannah, USA, 2016: 265–283.

[31] CHOI J, CHANG H J, YUN S, et al. Attentional correlation filter network for adaptive visual tracking[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4828–4837.

[32] ZHANG Jianming, MA Shugao, SCLAROFF S. MEEM: robust tracking via multiple experts using entropy minimization[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 188–203.

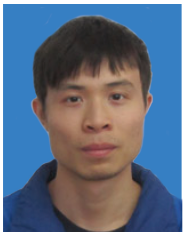
[33] MA Chao, YANG Xiaokang, ZHANG Chongyang, et al. Long-term correlation tracking[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5388–5396.

[34] SUN H. Y data[EB/OL]. https://github.com/SMZCC/A_proposed_deep_tracker.

作者简介：

孙海宇，男，1993 年生，硕士研究生，主要研究方向为图像处理、目标跟踪、深度学习相关算法。

陈秀宏，男，1964 年生，教授，博士后，主要研究方向为数字图像处理和模式识别、目标检测与跟踪、优化理论与方法。发表学术论文 100 余篇。



肖汉雄，男，1991 年生，硕士研究生，主要研究方向为模式识别和数字图像处理、人脸识别、深度学习相关算法。