结合多模板的多域卷积神经网络视觉跟踪算法

王鹏翔[1]，郭敬滨[1]，谭文斌[2]，李醒飞[1]
（1. 天津大学 精密测试技术与仪器国家重点实验室，天津 300072；2. 天津商业大学机械工程学院，天津 300134）

摘要：为了适应视觉跟踪过程中目标外观变化，提高视觉跟踪算法的鲁棒性，本文基于卷积神经网络（Convolutional Neural Network，CNN）并结合多域学习法与多模板管理，提出一种通过树形结构管理多模板的多域卷积神经网络（Multi-Domain CNNs with Multiple Models in a tree structure）视觉跟踪算法。首先使用大量已标记目标位置的视频数据预训练多域结构的 CNN，使 CNN 卷积层可从图像中提取出适用于跟踪任务的特征。然后在跟踪时中对 CNN 全连接层进行微调以适应跟踪目标，并使用树形结构管理存储不同时间段的目标模板得到模板树。使用模板树综合评价待检测帧，估计目标位置。最后按照一定规则将新模板添加进模板树，完成模板的更新。实验表明，该算法对跟踪过程中目标外观的变化有着良好的适应性，同时多模板可抑制 CNN 在跟踪时产生的模板漂移问题。

关键词：视觉跟踪；深度学习；卷积神经网络；多域学习；多模板
中图分类号：TP391 **文献标识码：**A **文章编号：**1001-8891(2018)01-0047-08

A Multidomain CNN that Integrates Multiple Models in a Tree Structure for Visual Tracking

WANG Pengxiang[1]，GUO Jingbin[1]，TAN Wenbin[2]，LI Xingfei[1]
(1. *State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin* 300072, *China*;
2. *School of Mechanical Engineering, Tianjin University of Commerce, Tianjin* 300134, *China*)

Abstract：To solve the problem of visually tracking a target that changes its appearance and improve the robustness of visual tracking, we propose a convolutional neural network (CNN)-based algorithm that combines a multidomain learning framework and multiple models stored in a tree structure. First, the multidomain CNN is pretrained with many videos containing tracking ground truths, so that its convolutional layer can extract features appropriate for visual tracking. During tracking, the fully connected layers are fine-tuned online to fit the target appearance, and the multiple target appearance models are managed in a tree structure. Then, the model tree is used to estimate the target's state in a new frame. Finally, a new model is updated along a path in the model tree. The algorithm produces outstanding performance when a target abruptly changes its appearance. Furthermore, the model tree can fix the problem of drift during online learning with the CNN.

Key words：visual tracking，deep learning，Convolutional Neural Network (CNN)，multi-domain learning，multiple models

0 引言

视觉跟踪是机器视觉领域的一个重要分支。跟踪过程中目标外观的变化是视觉跟踪最难解决的问题之一。为适应这种变化，需要跟踪算法快速提取并学习目标新的特征。传统的跟踪算法[1-4]采用人工设计的低层次特征（如 SIFT，HOG）表达目标，这种特征在跟踪过程中表现出提取效率低、适应性不强等弊端，特别是对于彩色图像，很难从中提取出合适的低层次特征。近年来，深度学习得到广泛的关注，而通过深度学习则可以从数字图像中提取出更加适用于视觉跟踪任务的高层次抽象特征。

由于 CNN 对图像信息的处理能力优越，因此该技术在机器视觉领域备受重视，并占据着重要地位。主

收稿日期：2017-04-01；**修订日期**：2017-10-26.
作者简介：王鹏翔（1992-），男，山东即墨人，硕士研究生，主要从事视觉跟踪、机器视觉、深度学习方面的研究。
基金项目：精密测试技术及仪器国家重点实验室开放基金资助项目（PIL1407）；天津市科技兴海项目（KJXH2012-11）。

要体现在图像分类[5-7]、物体检测[8]、图像语义分割[9-11]等方面。由于它的权值共享网络结构更类似于生物神经网络，减少了网络参数，降低了网络训练复杂度。特别是对于彩色图像，CNN 卷积层可从中直接提取出特征，与传统特征相比，简化了特征提取和数据重建的过程。但目前一些基于 CNN 的跟踪算法[12-13]仍是使用大量用于分类任务的数据（如 ImageNet[14]）对神经网络进行预训练。虽然使用这类数据可以让 CNN 从中学习一定的特征信息，但是这种特征信息与传统的低层特征类似，跟踪任务中效率较低、适用性差。由于视觉跟踪任务的目的是在任意物体中定位跟踪目标，而并非为图像分类，所以要使 CNN 卷积层能够从图像中提取出适用于跟踪任务的特征，便需要使用大量专门用于视觉跟踪的数据训练多域结构的 CNN[15]。

为了适应新的目标外观，需要在跟踪过程中通过在线学习用新的数据训练 CNN 的全连接层。但传统在线学习的前提条件是目标外观的变化平滑，即在跟踪过程中目标外观不会突变。并且使用新的数据训练卷积神经网络会使其快速丢失先前学习的信息[16]。神经网络的这种缺陷会使得在目标被遮挡或目标暂时丢失时，背景物体的特征被误当做目标特征，引起模板的漂移。所以传统的在线学习法不能够满足 CNN 跟踪的需要。

为此，本文提出了一种多域训练 CNN 方法[15]与树形结构管理目标模板[17]相结合的视觉跟踪算法。该算法首先使用大量已标记目标位置的视频序列预训练多域结构的 CNN。然后在跟踪的过程中采用树形结构进行目标模板的管理，在更新模板的同时保留不同时间段模板的历史信息，从而解决模板漂移问题[18]。这种结合在跟踪过程中既充分发挥了 CNN 对数字图像处理的优势，又通过树形结构弥补了 CNN 在线学习时易遗忘的缺点。通过实验证明该方法跟踪精确度高、成功率高，可快速适应目标外观的变化，并且在遮挡、光照变化，外观变化、低分辨率、杂乱背景等不良条件表现出良好的鲁棒性。

1 多域 CNN 的预训练

1.1 多域 CNN 结构

通过多域学习方法预训练的 CNN 在跟踪时有着良好的鲁棒性。本文所使用的多域 CNN 结构如图 1 所示。首先将 107×107 的 RGB 图像作为输入。之后使用 5 个隐藏层完成对图像的特征提取与分类，其中包括 3 个卷积层和 2 个全连接层。网络的最后是 K 个用于多域学习的输出层分支，这 K 个分支分别对应着 K 个域，每个域对应着一个已标记出目标位置的跟踪任务视频训练数据。卷积层的构造与 VGG-M 网络[5]的构造方法相同。卷积层后是两个全连接层，各有 512 个神经单元，采用 ReLu 激活函数并且使用 dropout 方法避免过拟合。最后的 K 个输出层分别有两个输出神经元用于分类每个域中的目标与背景。输出结果用一个归一化向量$[\phi(x),\ 1-\phi(x)]^{\mathrm{T}}$ 表示，元素分别代表输入图像是目标或背景的概率。本文将最后 K 个输出层合称为多域层，其他的层合称为共享层。

与用于识别、分类任务的网络结构（如 AlexNet[6]和 VGG-Nets[5,7]）相比，本文所采用的网络结构规模较小。在跟踪任务中采用这种轻量级结构主要有以下原因。首先，与需要区分多种类别的识别任务相比，跟踪任务只需区分目标与背景两个类别，所需信息承载量少。其次，随着 CNN 层数增多，从图像中提出的特征便越趋于抽象，图像的细节信息丢失越多，目标定位的精度也会随之降低[19]。最后由于跟踪目标图像的分辨率一般较小，所以需要处理的数据维度便比较小。最后，较为简单的结构可降低计算的复杂度，加快程序运行速度，保证跟踪的实时性。

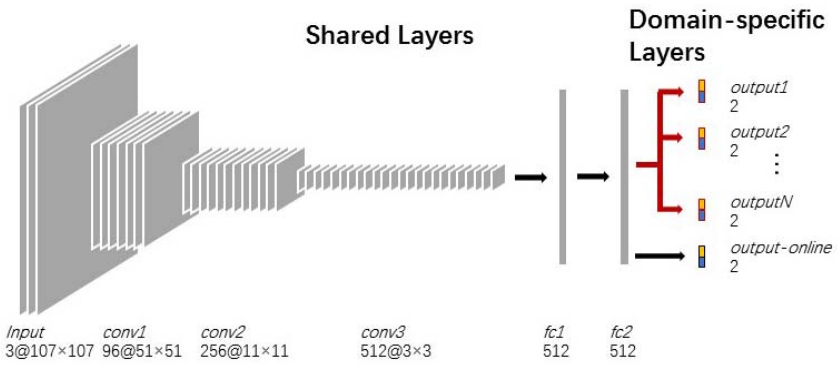

图 1 多域 CNN 结构

Fig.1 The architecture of multi-domain CNN

1.2 多域 CNN 训练方法

多域卷积神经网络的训练方法的本质是一种多域学习，可以从不同域中提取出共性的信息[20]。在不同的跟踪任务之中，虽然对目标与背景的定义是不同的，但却都有着光照变化、运动模糊、尺寸变化等等不良条件。在大量的训练数据中既存在着需要抛弃的个性的信息（如不同的跟踪任务中目标、背景是不同的），同时又存在需要获得的共性信息。通过多域结构便可以在训练的过程中将这种差异性信息集中到多域层，而卷积层只获得诸如鲁棒性的共性信息而忽略各个域中的个性信息。

在预训练时，首先从用于训练的视频集中的每一帧抽取 50 个正样本与 200 个负样本作为训练数据，其中正样本与目标的重叠率≥0.7，负样本重叠率≤0.5。采用随机梯度下降法（Stochastic Gradient Descent，SGD）训练多域结构 CNN，每一次迭代只有一个输出分支被激活参与计算，并只使用相对应的视频数据训练该分支。下一次迭代只将下一个输出层分支激活而其他分支不参与计算，并且训练数据变为该分支对应的视频数据。重复上述过程直至 CNN 收敛。

跟踪时多域层中输出分支 1～N 将弃之不用，建立一个新的输出层适应新的跟踪任务，而 3 个卷积层参数保持不变，只更新全连接层和新输出层。当跟踪任务开始时，将在第一帧标注出期望跟踪的目标的位置。于是在第一帧抽取出 500 个正样本与 5000 个负样本初始化训练该神经网络，而在随后的每一帧只选取 50 个正样本与 200 个负样本用于更新网络。正样本与目标重叠率≥0.7，负样本重叠率≤0.3，此处适当的降低负样本选取的重叠标准是由于每一帧跟踪结果不一定是十分准确的，以免将目标划入负样本。若跟踪任务只提供了目标图像而无背景图像，或在第一帧中目标过大从而占据绝大部分图像时，可以在跟踪的初期引入代价敏感学习模型[21]，赋予负样本预测高惩罚度，当出现背景后移除惩罚项。

2 模板树数学模型

2.1 模板树结构

本文采用树形结构管理储存多个目标模板以保证模板的连续性和多样性。主要分为目标的评估与模板的更新两个部分。这种树形结构的核心思想是为各个模板之间确定一种关系以确定跟踪过程中不同阶段对不同模板的依赖程度。模板树的结构以及工作原理如图 2 所示。图中箭头的宽度代表模板的权重，模板边框的宽度代表模板的可信度，模板之间连线的宽度代表它们之间的关联度，数字代表该模板加入模板树的次序。

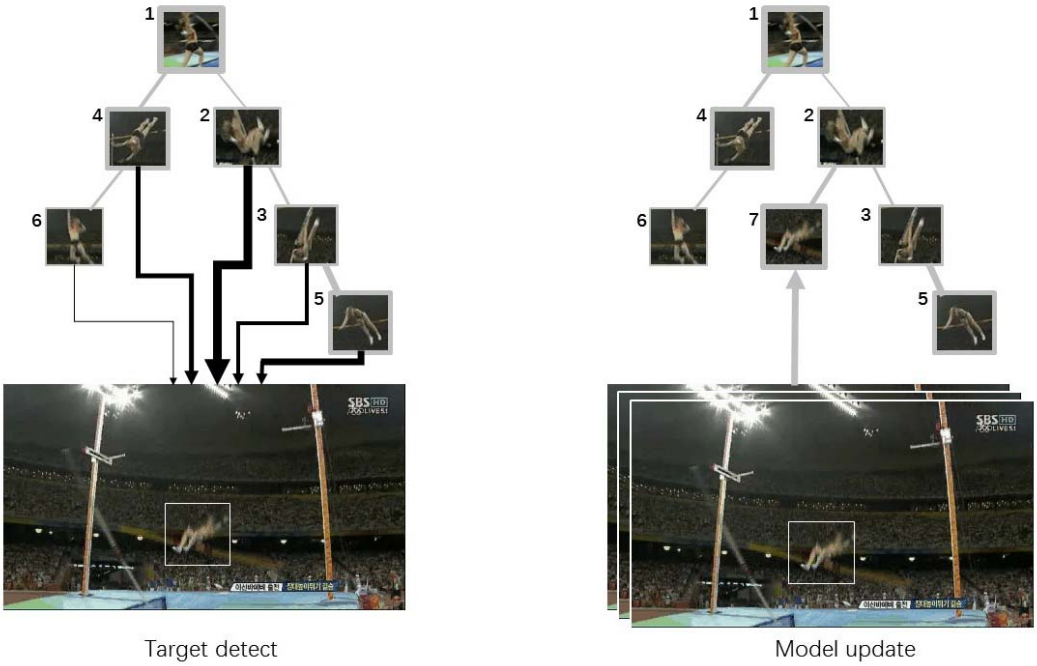

图 2 目标的检测与模板树的更新示意图。

Fig.2 Illustration of target state estimation and model update procedures

在实际应用中，为了节省储存空间，加快程序运行速度，并且由于 CNN 卷积层参数是共享的，模板树节点中并不直接储存目标的模板图像而是储存其对应的 CNN 的全连接层与输出层参数。也就是说，一个目标模板对应着一个适应该目标外观的 CNN 全连接层。并且在全连接层中包含着对负样本的记忆，在维护目标模型的同时也维护了背景模型。在跟踪一定帧数之后，这些帧会被用来训练 CNN，微调其全连接层和输出层以适应目标外观的变化。然后该 CNN 全连接层和输出层参数将被储存在模板树内。

在一个模板树结构 $T=\{V, E\}$中，一个顶点 $v\in V$ 对应着一个目标模板所对应的 CNN，有向边界$(u, v)\in E$ 示两个 CNN 之间的关系。两个 CNN 之间的关联度由下式得出：

$$s(u,v)=\frac{1}{|F_v|}\sum_{t\in F_v}\phi_u\left(x_t^*\right)$$

式中：F_v 代表用于生成 CNNv 的一段连续帧；x_t^*代表在第 t 帧中的目标；$\phi_u(\cdot)$是 CNNu 的评价函数；关联度 $s(u, v)$是目标的检测与模板树的更新时所需的关键参数。

2.2 目标的检测

在待检测帧中，首先在前一帧目标周围以正态分布取样若干（为了更加精确可加入粒子滤波等环节），然后使用模板树中的 CNN 对样本进行估计，接着使用所有CNN估计值的加权平均值作为最终估计结果。CNN 的权重是由网络在模板树中的更新路径决定的。在跟踪数帧 F_v 后，一个适应当前目标状态的 CNNv 会被生成，其拥有最大的权重。由于使用过多 CNN 进行目标估计会降低程序效率，通常只让一部分 CNN 参与计算，$V_+\subseteq V$ 代表处于激活状态的 CNN。用 x_t^1, …, x_t^N代表第 t 帧中所抽取的 N 个样本，样本 x_t^i 是目标的概率由下式得出：

$$H\left(x_t^i\right)=\sum_{v\in V_+}w_{v\to t}\phi_v\left(x_t^i\right)$$

式中：$w_{v\to t}$代表 CNNv 在第 t 帧的权重。取使得加权概率 $H(x_t^i)$最大的样本 x_t^i 作为跟踪目标 x_t^*：

$$x_t^*=\max_{x_t^i}H\left(x_t^i\right)$$

权重系数 $w_{v\to t}$的值主要由 CNNv 在当前帧 t 的相似度、模板的可靠度这两个参数决定。其中相似度 $a_{v\to t}$ 决定了 CNNv 与当前帧所抽取样本的相似程度，计算方法为：

$$a_{v\to t}=\max_{x_t^i}\phi_v\left(x_t^i\right)$$

但仅使用相似度 $a_{v\to t}$确定权重系数 $w_{v\to t}$而忽略各个 CNN 的可信度，这会使得有些神经网络在某些情况下（如目标被遮挡时遮挡物体与目标同时被当作目标，或者跟踪错误时背景被当作目标）对背景物体产生很高的相似度评分。CNN v 的可信度β_v通过递归方式得出，递归方程如下：

$$\beta_v=\min\left(s\left(p_v,v\right),\beta_{p_v}\right)$$

式中：p_v为模板树中 CNNv 父节点储存的 CNNp_v。

结合 CNN 的相似度和可靠度，可以得到它在第 t 帧的权重：

$$w_{v\to t}=\frac{\min\left(a_{v\to t},\beta_v\right)}{\sum_{v\in V_+}\min\left(a_{v\to t},\beta_v\right)}$$

随着 CNN 的层数变多，最终提取出的目标特征会变得抽象模糊，丢失边缘细节信息。这会使得在定位目标时不容易获得其精确边框，且影响之后更新 CNN 时训练样本的取样精度。本文采用 bounding box regression[22-23]来解决这种边界框松弛问题以提高目标定位精度。

首先在第一帧目标的周围取样并得到这些样本在卷积层 3 输出的特征，再用这些特征训练一个简单线性回归模型。在接下来的各帧中便可以通过这个模板对目标位置进行进一步精确。由于模型在线更新是十分耗时的，并且在线更新可能会引起模板的漂移，所以该回归模型只在第 1 帧训练，不随跟踪过程中更新。

2.3 模板的更新

使用树形结构管理模板并将相应的 CNN 储存在各个节点中，并通过一定的更新路径保证模板的可靠性，所以在得到新的训练样本后如何在模板树中选择合适更新路径是模板更新问题的关键。

在跟踪Δ（本文取 10）帧连续图像 F_z之后，在 z 节点创建一个新的 CNN。它的父 CNN 应使其可信度最大，所以连接父 CNN 的顶点 p_z可以通过公式给出：

$$p_z=\arg\max_{v\in V_+}\min\left(s\left(v,z\right),\beta_v\right)$$

使用 F_z与 F_{p_z}这两段图像序列作为训练样本，微调其父 CNN 得出新的 CNN 并储存在模板数中。模板树便增加了顶点 z 与相应的有向边界(p_z, z)。

3 实验与分析

本章节使用 Object Tracking Benchmark（OTB）[24]数据库中的视频对该算法进行仿真实验。算法通过

MATLAB 实现，在配置了 2.50 GHz 的 Inteli5CPU 与 NVIDIA GeForce GT555M 的 GPU 的计算机上运行。

使用边界框的中心误差（CE）与边界框覆盖率（O）来评价跟踪算法的精确度与成功率[25]。计算方法如下：

$$\mathrm{CE}=\sqrt{\left\|X_{\mathrm{C}}-Y_{\mathrm{C}}\right\|_F^2}$$

$$O=\left(X\cap Y\right)/\left(X\cup Y\right)$$

式中：X_{C}、Y_{C} 分别代表跟踪结果中心与真值中心的坐标。X、Y 分别是跟踪结果与真值的覆盖区域。

3.1 仿真实验

为了测试本文算法（CNN-MD-TS）跟踪效果，本章中选取了一些典型视觉跟踪算法与之对比，包括 KCF[2]、SCM[26]、Struck[27]、MUSTer[28]。在对 OTB 数据库中部分目标外观变化较大的视频的仿真实验中，本文算法对目标的形变、光照变化、遮挡等不良条件都表现出良好的适应性。图 3～图 6 展示了效果较明显的 4 个跟踪任务中的代表帧。各个图中从左至右每列依次是本文算法、KCF、SCM、Sturck、MUSTer 的跟踪结果，图中数字为每帧序号。可以看出，对于非刚性形变的目标外观变化，本文算法有着优势明显。特别对于图 4 和图 6 的彩色图像跟踪任务，由于 CNN 可直接对彩色图像进行特征提取，使得跟踪效果明显优于其他跟踪算法。

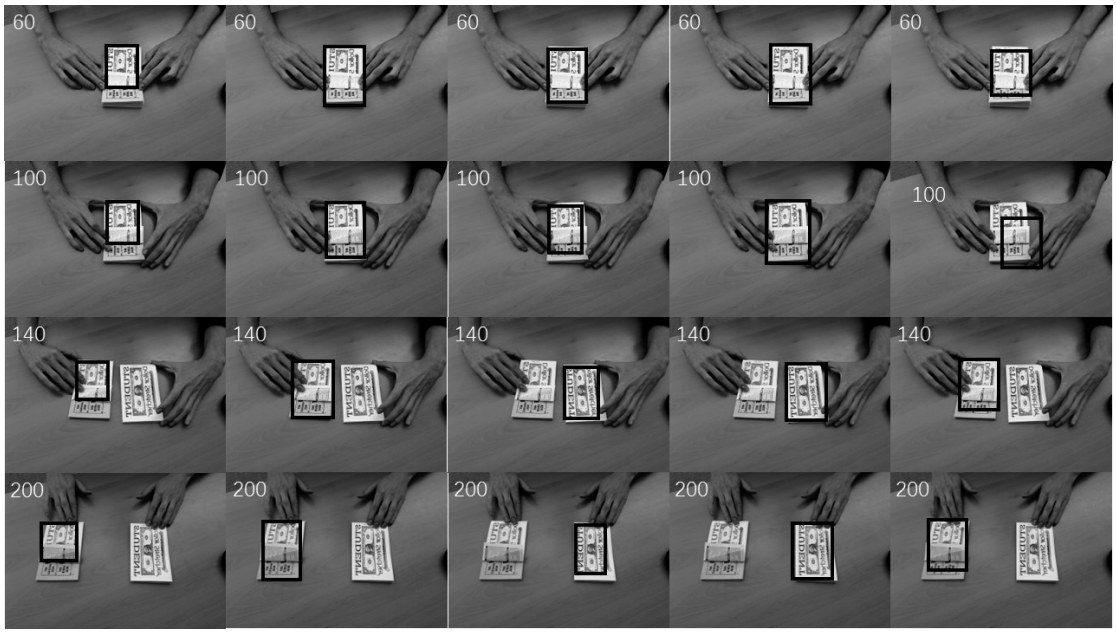

图 3 Coupon 视频跟踪效果对比 Fig.3 The results of Couponsequence

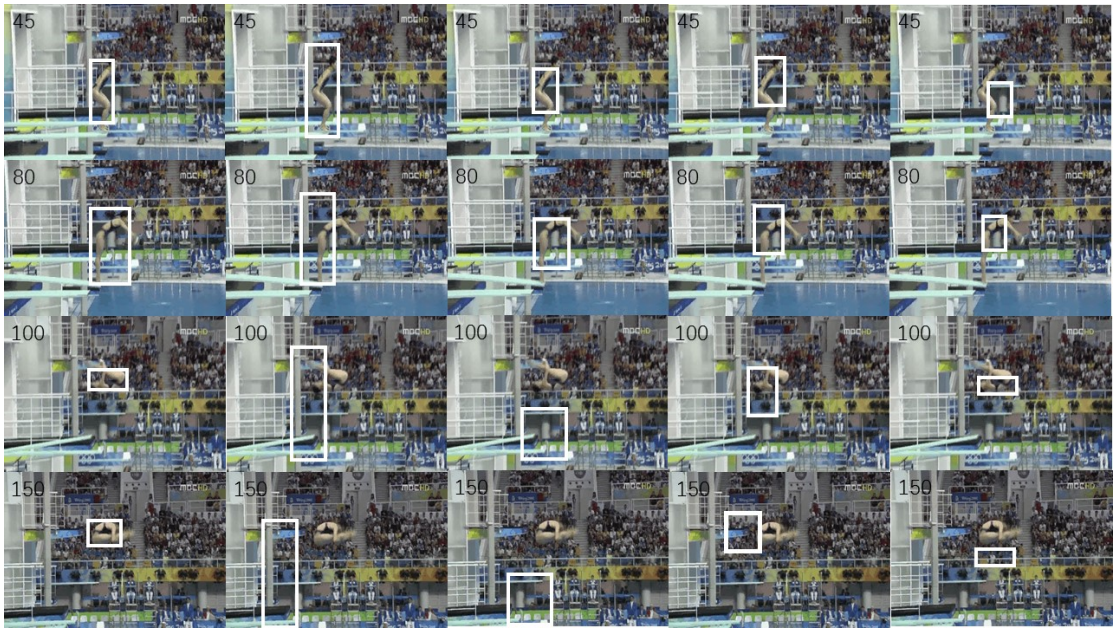

图 4 Diving 视频跟踪效果对比 Fig.4 The results of Divingsequence

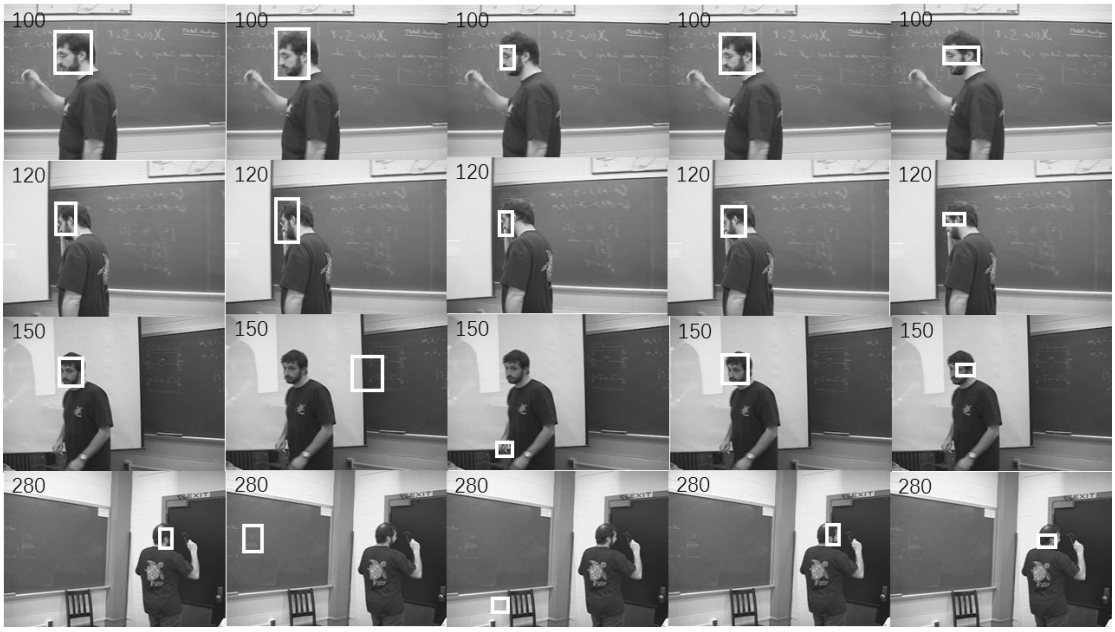

图 5 Freeman1 视频跟踪效果对比 Fig.5 The results of Freeman1sequence

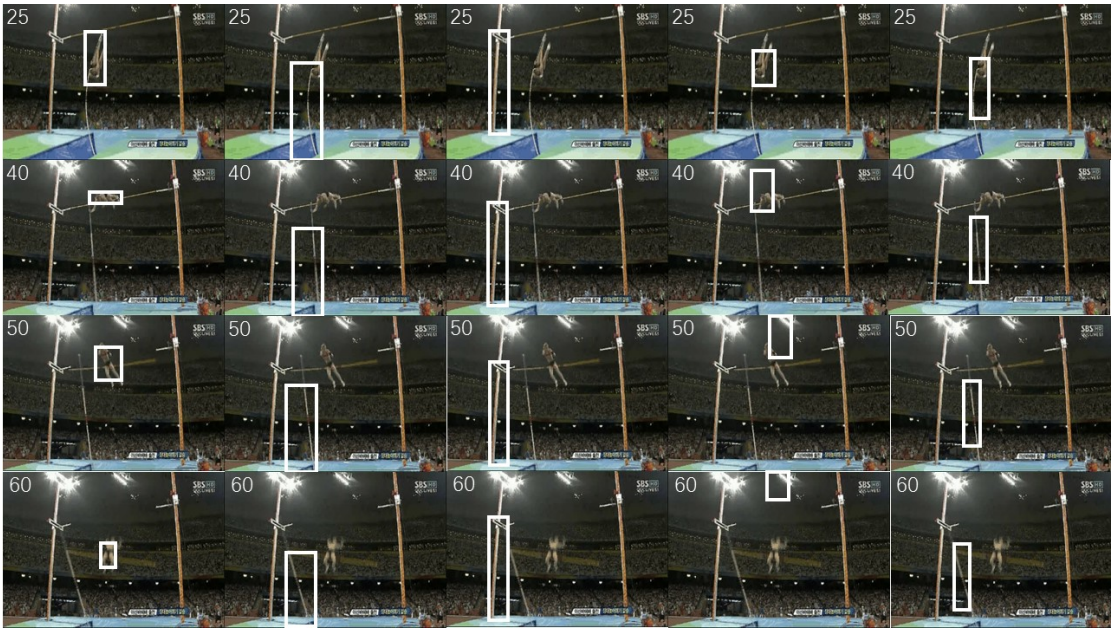

图 6 Jump 视频跟踪效果对比 Fig.6 The results of Jump sequence

3.2 模块分析

为了测试本文算法中多域训练与模板树这两个模块对跟踪性能的影响，将本文算法（CNN-MD-TS）与多域训练 CNN 算法（CNN-MD）和模板树 CNN 算法（CNN-TS）进行实验对比。多域训练 CNN 算法的预训练与本文方法相同，但在跟踪过程中只微调其全连接层与输出层以适应目标外观的变化，不进行模板的储存。前两种算法中的 CNN 的预训练使用的数据为 VOT2013[29]、VOT2014[30]和 VOT2015[31]中的 58 个视频，并不包括用于测试的 OTB 中的视频。模板树 CNN 算法的卷积层为使用 ImageNet[32]训练出的 VGG-M 网络[5]的卷积层，而目标的检测与模板的更新策略与本文方法相同。

通过实验，3 种算法对 OTB 中大部分视频序列有着很好的跟踪效果。但是对于部分目标外观变化明显且形变速度较快的跟踪任务，本文提出的结合算法有着更显著的优势，跟踪结果由表 1 给出。下面通过 Jump 视频的仿真跟踪实验对算法中两个模块进行分析。

从图 7 中可以看出，CNN-MD 由于目标的外观变化过快，将背景物体特征当作目标的特征进行学习，导致了模板的漂移，直至完全丢失目标。而采用模板树的两种算法则很好的抑制了这种漂移。图 8 绘制了 3 种算法在跟踪时覆盖率与中心误差的变化，可以看出，CNN-TS 对目标的识别精度很低，这主要是由其卷积层对跟踪物体的不敏感造成的。而本文提出的算法准确提出目标特征的同时也抑制了模板的漂移。

表 1 部分视频平均跟踪精度 Table 1 The average accuracy in several sequences

Video	O			CE		
	CNN-MD	CNN-TS	CNN-MD-TS	CNN-MD	CNN-TS	CNN-MD-TS
Diving	0.79	0.8	0.89	5.9	5.6	3.5
Coupon	0.5	0.76	0.98	64.3	6.1	2.9
Jump	0.4	0.67	0.8	70.6	9.4	5.1
Bolt2	0.91	0.84	0.95	3.3	4.5	3.4
Skating2-1	0.89	0.81	0.9	5.2	7.7	4.0

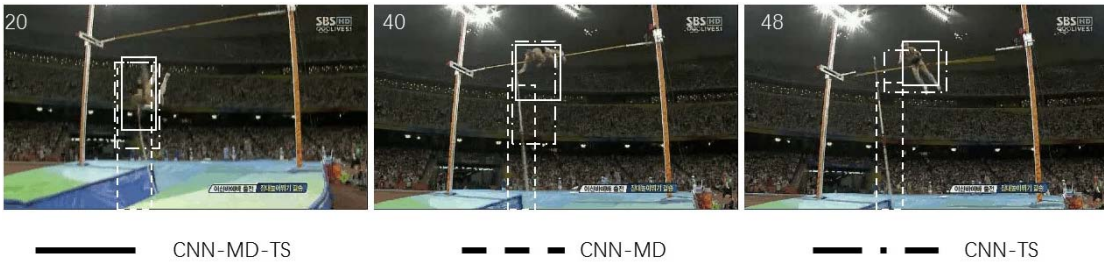

图 7 Jump 视频跟踪效果对比 Fig.7 The tracking results in Jump sequence

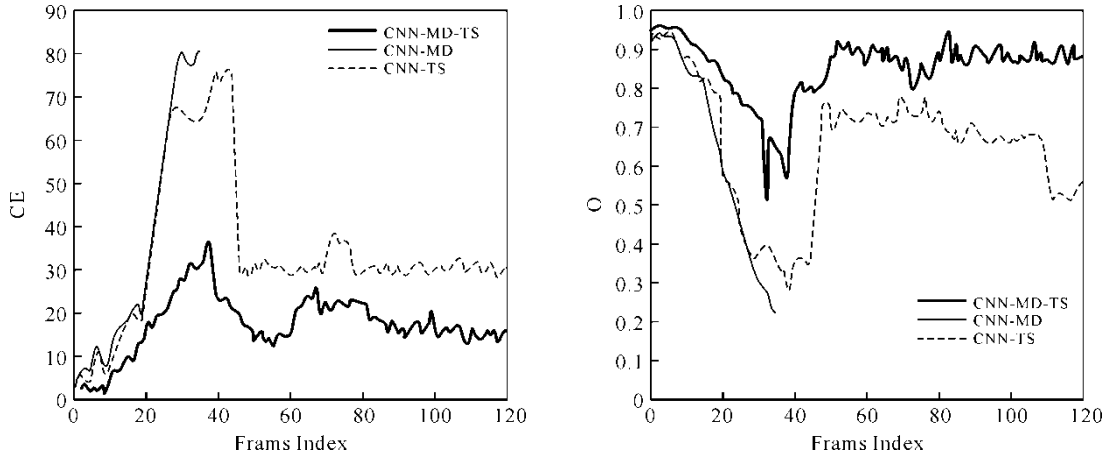

图 8 Jump 视频跟踪精度 Fig.8 The tracking accuracy in Jumpsequence

4 结论

通过多域学习法训练的 CNN 卷积层对视频跟踪任务有着良好的适应性，从图像中可直接提取出适用于跟踪任务的高层抽象特征，这种特征相比于人工设计的底层特征更容易被分类器（即神经网络中的全连接层、输出层）所区分。实验表明，该算法提高了跟踪的准确度和成功率。并且针对 CNN 训练时易遗忘的特性，模板树解决了在线学习时模板易漂移、模板单一的问题，并且使模板具有更高的可靠性。多域训练与模板树对 CNN 的跟踪性能均起着至关重要的作用。算法对跟踪时目标外观变化有着良好的适应性，并且在背景复杂、光照变化、运动模糊、遮挡等各种不良条件下，均表现出良好的鲁棒性且定位准确。

参考文献：

[1] Danelljan M, Häger G, Khan F S, et al. Accurate Scale Estimation for Robust Visual Tracking[J]. *BMVC*, 2014: 1-11.

[2] Henriques J F, Rui C, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015, **37**(3): 583-596.

[3] HONG Z, CHEN Z, WANG C, et al. MUlti-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking[J]. *Computer Vision & Pattern Recognition*, 2015: 749-758.

[4] ZHANG J, MA S, Sclaroff S. MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization[C]//*European Conference on Computer Vision*. *Springer*, *Cham*, 2014: 188-203.

[5] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the Devil in the Details: Delving Deep into Convolutional Nets[J]. *Computer Science*, 2014.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//*International Conference on*

Neural Information Processing Systems, 2012: 1097-1105.

[7] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. *Computer Science*, 2014: 1-14.

[8] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]//*Proceeding CVPR* '14 *Proceedings of the* 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 580-587.

[9] Hong S, Noh H, Han B. Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation[J]. *NIPS*, 2015: 1495-1503.

[10] LONG J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, **39**(4): 640-651.

[11] HONG S, YOU T, Kwak S, et al. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network[J]. *Computer Science*, 2015: 597-606.

[12] PAN Junting. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network[J]. *Computer Science*, 2015, 2015: 597-606.

[13] WANG N, LI S, Gupta A, et al. Transferring Rich Feature Hierarchies for Robust Visual Tracking[J]. *Computer Science*, 2015: 1-10.

[14] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. *International Journal of Computer Vision*, 2014, **115**(3): 211-252.

[15] Nam H, Han B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking[C]//*Computer Vision and Pattern Recognition, IEEE*, 2016: 4293-4302.

[16] Mccloskey M, Cohen N. Catastrophic interference in connectionist networks: The sequential learning problem[J]. *Psychology of Learning & Motivation*, 1989: 109-165.

[17] Nam H, Baek M, Han B. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking[J]. *arXiv*, 2016: 1-10..

[18] Li H, Li Y, Porikli F. DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking[J]. *BMVC*, 2014: 1-15.

[19] Hong S, You T, Kwak S, et al. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network[J]. *Computer Science*, 2015: 597-606.

[20] Daumé Iii H. Frustratingly Easy Domain Adaptation[J]. *ACL*, 2009: 1-9.

[21] ZHOU Z H. Cost-Sensitive Learning[J]. *Computer Science*, 2011, **6820**: 17-18.

[22] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object detection with discriminatively trained part-based models[J]. *Computer*, 2014, **47**(2): 6-7.

[23] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. *CVPR*, 2013: 580-587.

[24] WU Y, Lim J, YANG M H. Object Tracking Benchmark[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2015, **37**(9): 1834.

[25] Everingham M, Gool L V, Williams C K I, et al. The Pascal, Visual Object Classes (VOC) Challenge[J]. *International Journal of Computer Vision*, 2010, **88**(2): 303-338.

[26] ZHONG W. Robust object tracking via sparsity-based collaborative model[C]//*IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society*, 2012: 1838-1845.

[27] Hare S, Saffari A, Torr P H S. Struck: Structured output tracking with kernels[C]//*IEEE International Conference on Computer Vision, IEEE*, 2012: 263-270.

[28] HONG Z, CHEN Z, WANG C, et al. MUlti-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking[C]//*Computer Vision and Pattern Recognition, IEEE*, 2015: 749-758.

[29] Kristan M, Pflugfelder R, Leonardis A, et al. The Visual Object Tracking VOT2013 Challenge Results[C]//*IEEE International Conference on Computer Vision Workshops, IEEE Computer Society*, 2013: 98-111.

[30] Kristan M, Gatt A, Khajenezhad A, et al. The Visual Object Tracking VOT2013 Challenge Results[C]//*Computer Vision-ECCV* 2016 *Workshops*, 2016: 191-217.

[31] Kristan M, Pflugfelder R, Matas J, et al. The Visual Object Tracking VOT2015 Challenge Results[C]//*IEEE International Conference on Computer Vision Workshop*, 2016: 564-586.

[32] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//*Computer Vision and Pattern Recognition, IEEE Conference on*, 2009: 248-255.