深度相对度量学习的视觉跟踪

郭文^{1,2}, 游思思^{1,2}, 高君宇³, 杨小汕³, 张天柱³, 徐常胜^{3*}

1. 山东工商学院信息与电子工程学院, 烟台 264009

2. 山东省高校感知技术与控制重点实验室, 烟台 264009

3. 中国科学院自动化研究所模式识别国家重点实验室, 北京 100190

* 通信作者. E-mail: csxu@nlpr.ia.ac.cn

收稿日期: 2017-06-02; 接受日期: 2017-08-01; 网络出版日期: 2018-01-05

国家自然科学基金 (批准号: 61572296, 61603372, 61572498, 614722227, 61303086, 61307041, 61672327)、国家重点基础研究发展计划 (973 计划) (批准号: 2012CB316304)、山东省自然科学基金 (批准号: ZR2015FL020) 和模式识别国家重点实验室开放课题 (批准号: 201600024) 资助项目

摘要 传统的基于检测的跟踪方法虽然具有较好的鲁棒性, 但是这种直接将目标与背景简单分类的方式不能合理地考虑跟踪目标与背景的结构相对关系, 而这种相对结构判别性信息的缺乏使得跟踪算法极易发生跟踪漂移, 为了缓解这种问题, 本文提出了一个基于深度相对度量学习的视觉跟踪方法. 利用一个对称且权值共享的深度卷积神经网络构建深度相对度量学习模型, 通过这个模型来挖掘跟踪目标在大尺度的图像块里的结构相对关系, 然后在 Bayes 跟踪框架下利用这种相对度量最大值确定跟踪目标, 整个跟踪算法简洁有效. 通过在跟踪的基准视频序列库上的实验结果验证了本文算法在跟踪精度和跟踪成功率上的高性能.

关键词 相对属性, 度量学习, 卷积神经网络, 视觉跟踪

1 引言

视觉目标跟踪作为目前计算机视觉研究的热点问题之一, 已经在视频监控、人机交互、行为分析等领域中得到广泛的应用. 我们通常把目标跟踪看成在一个跟踪序列中寻找与被跟踪目标进行匹配的最优极值的搜索问题^[1~3]. 研究人员已经提出大量视觉跟踪的算法, 这些算法在一定范围内取得良好的效果, 但仍然存在很多实际问题有待解决. 特别是在目标的背景、形态、姿势、大小和光照条件发生强烈变化等复杂情况下, 对目标进行稳健、实时的跟踪仍然是一个具有挑战性的任务.

当前目标跟踪方法绝大多数是直接从目标前景或者背景中提取用于目标外观表达的特征并根据一定的度量来决定在跟踪序列中的某一位置是否是跟踪的区域, 这种“是”或者“不是”的二值决策方式的绝对跟踪方法, 缺乏对目标周围图像块相对结构关系的考虑, 而这种相对的结构关系对于跟踪

引用格式: 郭文, 游思思, 高君宇, 等. 深度相对度量学习的视觉跟踪. 中国科学: 信息科学, 2018, 48: 60–78, doi: 10.1360/N112017-00124
Guo W, You S S, Gao J Y, et al. Deep relative metric learning for visual tracking (in Chinese). Sci Sin Inform, 2018, 48: 60–78, doi: 10.1360/N112017-00124

图1 (网络版彩图) 相对值与绝对值对比

Figure 1 (Color online) Relative value vs. absolute value

是非常重要的^[4]。例如在图1第一行中,目前的直接跟踪的方法会利用已有的或者手工标注的第一帧目标的信息进行表观建模,然后从后续帧的搜索区域找到与之匹配的区域,在图1中我们会发现后续帧由于剧烈运动遭受严重的模糊,这时简单的决策方式会发现所搜索的区域没有与之匹配的目标区域,导致跟踪漂移的发生,进而使跟踪失败。实际上在后续帧搜索区域中已有的目标表观由于剧烈运动的影响已经发生了很大的变化,已经不适合进行“是”或者“不是”的二值决策方式,因为此时搜索区域所有的图像块对于原始的目标表观来讲都是同等的的不相似,缺乏判别性。在图1第二行中的方法与这种绝对跟踪方法不同,我们想极大程度上利用搜索区域的图像信息,并且挖掘前景与背景图像块的结构相对关系,可以在第一帧中大尺度地给出相关区域与原始目标表观的这种相对关系的模型,这样即使后续帧发生很大的模糊,跟踪方法依然可以估计出拥有最大相对关系的图像块。当然,如何学习这种大尺度的相对关系模型也是一个非常艰巨的任务,需要指出的是这里的大尺度不是指通过大尺度的离线训练图像数据库,而是指对跟踪序列初始视频帧进行大量的数据采样取得大尺度的训练样本。过去有些简单的相对关系的学习模型可以通过排序支持向量机算法^[5]或者度量学习^[6]的方法来完成,但是如果有效地处理这种大尺度的图像数据及数据里存在的复杂相对关系,这些方法也难以奏效。最近深度学习的方法引起了业内学者的极大兴趣,由于其具备强大学习视觉特征的能力^[7],被广泛地应用到计算机视觉领域中去,并且取得了很大的成功。但是大多数的跟踪领域的深度学习应用^[8~11]仍然属于绝对跟踪,究其原因主要在于深度学习需要大量的样本进行模型训练,而绝对跟踪的初始帧仅有少量的正样本提供,无法确切地利用相关信息进行合理目标表达。要克服这种问题,目前有两种解决办法,一种是采用离线训练的方法,先利用大量的样本进行预训练深度学习的模型;另外一种就是仅用少量的样本进行浅网络的训练,实际上这也不是深度学习的范畴了。

不同于前面的跟踪方法,为了能够在跟踪里合理地利用深度学习进行刻画大尺度样本的相对关系,并且能够保留样本的多样性,本文提出了一个有效的深度相对度量学习的目标跟踪算法,最大的创新点在于:首先,利用卷积神经网络学习一个有效的非线性的排序函数来描述前景与背景中大尺度图像块的相对关系,进行目标的表观表达,这种强有力的特征表达将极大地提高跟踪的效果;此外,背景与前景的图像块的相对关系模型与深度学习紧密的结合,这种相对关系模型可以使得深度学习在少量的正样本的情况下进行合理的模型训练,反过来合理的学习模型又极大地推进了对图像块相对关系的挖掘;最后,本文提出的方法能够保留样本的多样性,克服样本退化,同时能够使得特征的提取和度量学习结合在一个深度学习的过程中,提高跟踪的效果。本文在一系列有挑战性的数据库上进行了评测,这

些数据集上的实验结果验证了所提方法具有鲁棒的度量学习能力, 取得了很好的跟踪效果.

本文的结构安排如下: 第 2 节对相关工作进行了回顾; 第 3 节详尽介绍本文所提的深度相对关系学习的构造方法, 并给出了问题优化方法; 第 4 节给出所提方法在一系列跟踪数据集上的实验效果以及分析与评价; 最后在第 5 节总结全文, 并给出进一步的工作.

2 相关工作

大量的跟踪理论出现在相关文献中, 有关综述类可见文献 [1~3], 本文将仅回顾与我们工作相关的绝对目标跟踪以及采用深度学习进行目标跟踪的文献. 在目标跟踪方法中, 最重要的步骤是计算图像中候选目标区域与目标模型的相似度, 这样建立目标的表观模型就极为重要, 文献 [12] 表明能够建立强有力的特征模型是决定跟踪是否成功最重要的一环. 大多数研究都将目前的跟踪方法分成生成式跟踪与判别式跟踪两类, 而根据我们上面的讨论, 这些方法基本上都是属于绝对跟踪方法, 它们都是直接考虑跟踪样本的结果, 而没有考虑跟踪样本中蕴含的丰富的相对关系. 经典的生成式目标跟踪的方法, 如 Comaniciu 等 [13] 和 Collins [14] 提出的基于核的均值漂移方法, 采用 Bhattacharyya 系数作为度量候选目标与目标模板之间的相似度, 使用核加权赋予距目标中心近的像素更大的权重, 进一步提高目标直方图与周围背景区域直方图的区分度, 最终通过均值漂移迭代的方法求得目标的位置. Ross 等 [15] 提出了一种增量式学习低维子空间表示的跟踪方法, 构建可以自动更新的表观模型, 解决由于刚体运动或者有限的变形运动的表观变化. 文献 [16] 采用低秩稀疏学习来描述粒子滤波框架下粒子之间的关系. Porikli 等 [17] 采用协方差算子作为目标的表观表达, 来描述多种特征融合时的互相关关系, 并且采用了李群的更新策略. Li 等 [18] 为了缓解基于主成分分析 (PCA) 子空间跟踪方法中过度拟合问题, 提出了基于最大后验概率 (MAP) 的 PCA 子空间学习算法. Zhou 等 [19] 为了提高跟踪特征的判别性, 则采用了多特征融合的相似度计算方式, 并进一步结合上下文信息来加强这个相似度的判别性.

而绝对跟踪中的另一类是判别式跟踪方法, 它将计算图像中候选目标区域的似然转换为一个目标和背景的二值分类问题. 文献 [20] 利用压缩感知的概念对目标表观建模, 然后采用朴素 Bayes 分类器实现实时跟踪. Avidan [21] 提出了一种支持向量机的跟踪方法, 这种跟踪方法通过一个离线训练的支持向量机分类器嵌入到光流跟踪器, 桥接了跟踪器与分类器之间的联系, 在其后续工作中, Avidan [22] 才真正意义上从分类器的角度去考虑跟踪问题, 通过自适应提升算法来训练弱分类器集成强分类器得到前景与背景的置信度图, 最终在置信度图上进行模式搜索完成目标的定位. Collins 等 [23] 则另辟蹊径, 从特征的判别性角度考虑, 在线学习一个判别性特征选择器将前景从背景中区别出来. 文献 [24] 提出了一种在线特征提升的跟踪算法, 该算法利用第一帧的标记数据初始化分类器, 然后在线地对分析器进行更新. Guo 等 [25] 通过概率性类标来对不确定性进行先验建模, 提出了一个基于不确定性的最大置信度提升跟踪方法, 缓解了判别式跟踪的跟踪漂移问题. Saffari 等 [26] 则提出了一种基于随机决策树集成的随机森林学习算法, 并应用到跟踪中, 与文献 [24] 相比, 由于分类器精度的提高得到了更加可靠的跟踪结果. Babenko 等 [27] 将跟踪问题定义为一个多示例的学习问题, 其中的定义采用了概率的表达, 在一定程度上提高了跟踪效果. 文献 [28] 则为了减少跟踪的累积误差, 采用了一种所谓的“PN learning”的调节机制, 通过对正负样本的约束来维持跟踪器的性能. Zhang 等 [29] 提出了一个熵值最小化的多专家跟踪方法来处理跟踪的漂移问题, 这是一个有潜力的跟踪框架. 最近核相关跟踪方法 [30~33] 又有很大轰动效应, 一系列利用循环矩阵多特征在 Fourier 变换域进行目标跟踪的方法如雨后春笋蓬勃而出. 虽然这些基于分类器的判别式跟踪方式在效果上取得了很大进步, 但是在搜索区域

图2 (网络版彩图) 大尺度训练样本

Figure 2 (Color online) Large scale training samples

内由于复杂背景、剧烈运动、遮挡、光照变化等一系列的干扰而导致前景与背景不存在区分性, 正样本性能严重退化, 这时判别式跟踪也遭遇了瓶颈, 需要大量的正样本进行学习, 而完成这样的问题几乎不可能, 从而导致跟踪的失败.

那么如何在目标跟踪少量的真实样本实现大尺度样本的学习, 并且保留样本多样性给出强有力的表观表达, 深度学习^[34]给出了答案. 当前深度学习的方法几乎横扫了分类学习的各个领域, 作为当前极为火热的算法, 跟踪领域^[8~11]也不甘示弱, Fan等^[8]利用一个离线训练的卷积神经网络实现了人员的跟踪, 这是目前能找到的跟踪领域深度学习的雏形. 文献^[35]则通过一个预先训练的堆栈降噪自动编码器学习一般图像的特征实现基于深度学习的目标跟踪, 它采用了离线预训练加在线微调的思路, 在一定程度上解决了跟踪中训练样本不足的问题. Hong等^[36]利用预训练的卷积神经网络学习了一个判别性关注度模型算法, 通过给出关注度置信度来引导支持向量机的分类器跟踪. 文献^[37]不是把卷积神经网络看成一个黑箱函数, 而是分析了网络顶层与底层在跟踪特征上的不同点, 通过特征选择提高了跟踪的速度与性能. 不同于前面预先训练卷积神经网络的深度学习跟踪方法, Li等^[38]提出了一个带有截断结构损失函数的有目标导向的卷积神经网络, 构建了在线跟踪方法. Ma等^[39]简洁有效地利用深度特征做跟踪, 其主要思路是利用相关滤波器在所有的卷积层上学习目标的表现表达, 采用由粗粒度到细粒度的机制最终准确定位目标的中心点. Cui等^[40]利用多方向递归神经网络 (multi-directional recurrent neural network) 来建模并挖掘对整体跟踪影响最大的可靠目标部分, 实际上是二维平面上的递归神经网络建模, 最终解决了跟踪中预测误差累积和传播导致的跟踪漂移问题. Bertinetto等^[41]通过使用有标注了关联的视频数据集作为训练数据, 通过双胞胎网络 (siamese network) 的结构离线地训练出一个匹配函数, 把跟踪问题转化为了在每一帧通过相邻两帧的连续性生成很多建议分布, 然后判断这些分布和第一帧标注的物体的相似性. Tao等^[42]则采用了基于双胞胎的示例搜索的跟踪方法, 实验显示了该方法对于丢失目标能够重定位的能力. 虽然深度学习在跟踪领域有了一些进展, 但是还是属于刚刚起步, 特别是在视觉跟踪里, 由于缺乏大量的训练样本, 深度学习的方法很难施展拳脚, 深度模型的魔力之一来自于对大量标注训练数据的有效学习, 而目标跟踪仅仅提供第一帧的图像作为训练数据. 为了缓解这个问题, 本文提出了一个有效的深度相对度量学习的目标跟踪算法, 即可以利用有限的标注样本生成大尺度的训练样本, 如图2所示. 需要指出的是这里的“大尺度”不是指通过大尺度的离线训练图像数据库, 而是指对跟踪序列初始视频帧进行大量的数据采样取得大尺度的训练样本. 本文通过构建深度相对度量学习模型来挖掘样本内部的结构信息, 使得基于深度学习的目标跟踪能够运行的更加合理.

图 3 (网络版彩图) 深度相对度量学习模型网络结构

Figure 3 (Color online) The network structure of deep relative metric learning model

3 深度相对度量学习的目标跟踪

本节首先提出深度相对度量学习模型 (deep relative metric learning, DRML), 并且给出其问题的数学描述. 然后将介绍深度相对度量学习网络结构, 并且构建了一个合理的损失函数来约束网络输出的可靠性, 最后将给出前向以及后向网络传播策略来优化求解该模型.

3.1 问题的数学描述

传统的度量学习是通过带有某种约束的线性变换得到距离度量, 而相对度量学习的目的就在于找到一个鲁棒的排序函数使得前景与背景能够在相对的空间领域中有足够的差异. 这里的相对度量数值定义为采样图像框 r_S 与基准结果框 r_G 的交集面积与它们并集面积的比值 Ω ,

$$\Omega = \frac{\text{area}(r_S \cap r_G)}{\text{area}(r_S \cup r_G)}. \quad (1)$$

采用 $f(x)$ 表示在某种特定度量上排序函数. 令 $\mathcal{P}$ 表示一系列有标记的正样本对集, $\mathcal{N}$ 则表示一系列无标记样本对集, 如果一个图像对 $(x_i, y_i) \in \mathcal{P}$, 那么表示图像 x_i 比 y_i 有更高的相对度量值, 而如果一个图像对 $(x_i, y_i) \in \mathcal{N}$, 那么表示图像 x_i 与 y_i 有近似的相对度量, 这样相对度量的学习可以定义为满足如下约束的排序函数 $f(x)$ 的学习^[43]:

$$\begin{cases} \forall (x_i, y_i) \in \mathcal{P}, & f(x_i) > f(y_i), \\ \forall (x_i, y_i) \in \mathcal{N}, & f(x_i) = f(y_i). \end{cases} \quad (2)$$

为了深入刻画大尺度图像对的相对度量的关系, 本文通过卷积神经网络来构建一个深度相对度量学习模型. 与常规的属性学习方法中排序函数是线性的不同, 本文学习一个非线性的排序函数; 与传统深度学习的结构不同, 本文采用的是对称双边神经结构层, 如图 3 所示. 每边的网络结构包括 5 层卷积层 ($\text{Conv}_1 \sim \text{Conv}_5$) 和 5 层全连接层 ($F_{c6} \sim F_{c10}$), 在卷积层采用与 AlexNet^[44] 相同的抽样和归一化准则. 为了简化网络, 图 3 中卷积层和全连接层结束后不再显示抽样层、归一化层和遗弃层, 遗弃操作仅在全连接层 $F_{c6} \sim F_{c7}$ 之后进行.

在训练阶段, 一个图像对和它的标注值输入卷积神经网络, 图像将通过网络的特征学习和提取进行前向传播, 最后输出的是对应图像的相关得分, 这个过程实际上可以表述为把图像映射到一个隐式的非线性特征空间. 从网络收敛速度和稳定性的角度考虑, 本文对全部卷积层和前 2 个全连接层按照 Caffe^[45] 模型进行初始化, 全连接层 $F_{c8} \sim F_{c10}$ 的维数分别设定为 1000, 500, 1.

图 4 (网络版彩图) 邻域相对值分布

Figure 4 (Color online) Distribution of neighborhood relative value

全连接层结束后, 本文设计了一个相对损失层来约束一个图像对的相对度量的输出值, 我们知道如果一个图像对里的两个图像有相似的相对度量输出, 那么他们属于同一类; 相反, 不同的图像对则有不同的输出, 而且应该所有属于正样本的样本标注的相对度量值应该处于以一个定值为中心的领域空间中, 其他则远离这个领域随机分布. 因为同一图像对的两个图像需要同等对待, 在同一层数上的双边卷积网络可以共享权重, 虽然是相同的对称双边卷积神经网络, 但是本文将两个图像对独立地通过卷积层和全连接层进行传播, 下面将介绍模型结构.

3.1.1 卷积层与全连接层结构

对于第 m^{th} 层卷积层, 其输出表示为 $C_m(x) = s(W_m * C_{m-1}(x) + b_m)$, $m \in \{1, \dots, 5\}$, “*” 是卷积算子, W_m 和 b_m 分别是卷积层的卷积核和偏置因子, 函数 $s(x) = \max(0, x)$ 是修正线性单元 (rectified linear units) 里采用的不饱和非线性激活函数. 对于第 n^{th} 层全连接层, 其输出表示为 $F_n(x) = s(W_n F_{n-1}(x) + b_n)$, $n \in \{1, \dots, 5\}$, W_n 和 b_n 分别是全连接层的权重矩阵和偏置因子, 全连接层的激活函数类似于卷积层的修正线性单元.

3.1.2 相对度量学习的损失函数

损失函数的结构界定了卷积神经网络输出结果的优劣, 为了使网络在训练阶段得到更加鲁棒的结果并且能够提高网络收敛的速度, 相对度量学习的损失函数包含了所有正负样本集的损失、各自样本与其中心的聚类相对度量损失以及正则化项. 对于一个序列 $S = \mathcal{P} \cup \mathcal{N}$ 是针对相对度量的包含所有正负样本的集合, 根据 3.1 小节的叙述, 如果一个图像对 $(x_i, y_i) \in \mathcal{P}$, 那么表示图像 x_i 比 y_i 有更高的相对度量值, 而如果一个图像对 $(x_i, y_i) \in \mathcal{N}$, 那么表示图像 x_i 与 y_i 有近似的相对度量, 而且应该所有属于正样本样本标注的相对度量值应该处于一定值为中心的领域空间中, 其他则远离这个领域分布, 如图 4 所示. 为了使模型具有自我纠错的能力, 从目标跟踪的实际出发, 以初始帧的相对度量值作为这个邻域中心的相对度量值. 这也可以看成目标的模板模型, 这样模型可以随时根据模板的模型进行自我修正.

本文将一个图像对的对比约束 $\mathcal{L}_p$ 、相似度约束 $\mathcal{L}_n$ 和中心聚类约束损失 $\mathcal{L}_o$ 分别定义如下:

$$\begin{aligned}\mathcal{L}_p &= s(\alpha - (\mathbf{f}(x_i) - \mathbf{f}(y_i))), \\ \mathcal{L}_n &= \frac{1}{2}((\mathbf{f}(x_i) - \mathbf{f}(y_i))^2), \\ \mathcal{L}_o &= \frac{1}{2}s(\beta - ((\mathbf{f}(y_i) - \mathbf{f}(o))^2 - (\mathbf{f}(x_i) - \mathbf{f}(o))^2)),\end{aligned}\quad (3)$$

其中, 函数 $s(x) = \max(0, x)$ 是修正线性单元里采用的不饱和非线性激活函数. $\mathbf{f}(\cdot)$ 表示最后一个全连接层的输出. 对比约束 $\mathcal{L}_p$ 保证约束输入图像正样本的相对度量是最大的, 相似度约束 $\mathcal{L}_n$ 保证所有负样本的相对度量值学习是最小的, 中心聚类约束损失 $\mathcal{L}_o$ 则控制着跟踪模型自我纠错的能力.

根据式 (3), 把整体网络的相对度量学习的损失函数定义为

$$\begin{aligned}\mathcal{L} &= \sum_{(x_i, y_i) \in \mathcal{P}} \frac{1}{N^p} \mathcal{L}_p(x_i, y_i) + \lambda_1 \sum_{(x_i, y_i) \in \mathcal{N}} \frac{1}{N^n} \mathcal{L}_n(x_i, y_i) \\ &\quad + \lambda_2 \sum_{(x_i, y_i) \in \mathcal{P}} \frac{1}{N^p} \mathcal{L}_o(x_i, y_i, o) + \frac{1}{2} \sum_{m=1}^M \|\mathbf{W}_m\|_F^2,\end{aligned}\quad (4)$$

其中, N^p 与 N^n 分别表示正负样本集的个数; 系数 α 是对比约束阈值, 用于约束图像 x_i 比 y_i 具有更大的相对度量学习值, 并且两者之间的距离大于 α ; 系数 β 是聚类损失约束阈值, 用于约束图像正样本 x_i 比 y_i 对邻域中心 o 最近的具有更大的相对度量学习值, 并且两者之间的距离大于 β ; λ_1 、 λ_2 是调节因子, 用于对各项进行平衡.

式 (4) 的第 2 项保证所有负样本相对度量值学习是最小的. 第 4 项是一个包含网络模型所有参数的平衡项, 用于提高网络的精度. 第 3 项能够确保图像 x_i 比 y_i 和邻域中心相对度量具有更大的相对度量学习值. 例如, 第 3 项可以看成两种情况: 第 1 种, 如果 $(\mathbf{f}(x_i) - \mathbf{f}(o)) - |\mathbf{f}(x_i) - \mathbf{f}(o)| < \beta$, 那么第 3 项将成为一个正值, 最小化损失函数也将使得此项变为零; 另外一种, 如果 $(\mathbf{f}(x_i) - \mathbf{f}(o)) - |\mathbf{f}(y_i) - \mathbf{f}(o)| > \beta$, 那么这个损失值将成为我们所期望的零. 损失函数的第 1 项也是类似的情况, 不再赘述. 在深度相对度量学习模型里, 视觉特征和度量的排序是同一个网络里联合完成的, 并且能够相互促进.

3.2 优化求解

本文的 DRML 模型的优化方法与传统的神经网络类似, 采用了随机梯度下降的方法.

(1) 向前传播. 在向前传播中, 对于每个图像对 (x_i, y_i) 而言, 图像 x_i 首先作为输入 I_1 在深层相对度量学习模型的第 1 个分支中传播, $I_1 \rightarrow C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow F_6 \rightarrow F_7 \rightarrow F_8 \rightarrow F_9 \rightarrow F_{10}$. 与此同时, 图像 y_i 也会作为输入 I_2 在第 2 支的神经网络中传播, $I_2 \rightarrow C'_1 \rightarrow C'_2 \rightarrow C'_3 \rightarrow C'_4 \rightarrow C'_5 \rightarrow F'_6 \rightarrow F'_7 \rightarrow F'_8 \rightarrow F'_9 \rightarrow F'_{10}$. 全连接网络的最后一层输出图像相对度量值的强弱程度. 然后, 基于全连接网络的输出 F_{10} 和 F'_{10} , 整个网络的损失值通过相对损失函数的约束计算得到.

(2) 向后传播. 在向后传播算法中, 首先计算式 (4) 中的损失函数对于相对损失层的输入向量 $\mathbf{f}(x)$ 的偏导数. 然后, 网络的损失层的误差会向前传播到剩下的全连接和卷积层网络中. 全连接层和卷积网络中的参数 $\{\mathbf{W}_m\}_{m=1}^M$ 和偏置向量 $\{\mathbf{b}_m\}_{m=1}^M$ 会被更新. 尽管损失函数中的误差是在两支网络中被分别传播. 对于所有的图像对, 每一层网络中的权重矩阵和偏置向量的更新值最终会被累加起来, 两支网络中的参数会被同步. 对于图像对 (x_i, y_i) , 损失函数对于图像 x_i 和 y_i 的梯度的计算是根据链式法则在两支网络上分别进行传播的. 当图像 x_i 和 y_i 贡献的梯度被计算出来, 每层网络的权重和偏

置向量的变化会被累加起来. 然后所有层中的参数会被更新和同步. 相对损失函数层和全连接层中的偏导数计算将在下面详细介绍, 卷积网络中的偏导数是按照传统卷积网络中的类似方式来计算的.

在相对损失层的偏导数计算中, 使用 $\mathbf{f}(x)$ 表示对于最后一个全连接层 F_{10} 的输出. 为了使公式表达简单, 采用如下标记, 这里令 $\mathbf{d}_i^{(1)} = \mathbf{f}(x_i) - \mathbf{f}(y_i)$, $\mathbf{d}_i^{(2)} = (\mathbf{f}(y_i) - \mathbf{f}(o))^2 - (\mathbf{f}(x_i) - \mathbf{f}(o))^2$, 用 $\text{sign}(x)$ 表示二值函数 (如果 $x > 0$ 输出 1, 否则输出 0), 则式 (4) 中的损失函数关于 $\mathbf{f}(x)$ 的偏导数可以按照如下方式计算:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{f}(x_i)} = \begin{cases} \frac{\lambda_1}{N^n} \mathbf{d}_i^{(1)}, & (x_i, y_i) \in \mathcal{N}, \\ \frac{1}{N^p} \left(\lambda_2 (\mathbf{f}(x_i) - \mathbf{f}(o)) \text{sign}(\beta - \mathbf{d}_i^{(2)}) - \text{sign}(\alpha - \mathbf{d}_i^{(1)}) \right), & (x_i, y_i) \in \mathcal{P}, \end{cases} \quad (5)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{f}(y_i)} = \begin{cases} -\frac{\lambda_1}{N^n} \mathbf{d}_i^{(1)}, & (x_i, y_i) \in \mathcal{N}, \\ \frac{1}{N^p} \left(\text{sign}(\alpha - \mathbf{d}_i^{(1)}) - \lambda_2 (\mathbf{f}(y_i) - \mathbf{f}(o)) \text{sign}(\beta - \mathbf{d}_i^{(2)}) \right), & (x_i, y_i) \in \mathcal{P}. \end{cases} \quad (6)$$

在全连接层中的偏导数计算中, 简单起见, 用 $\mathbf{f}(x) = \mathbf{h}_M(x) = s(\mathbf{W}_M \mathbf{h}_{M-1}(x) + \mathbf{b}_M)$ 表示最后一层全连接网络的输出. 式 (4) 中的损失函数关于第 M 个全连接层 F_{10} 中的参数 $\mathbf{W}_M$ 的偏导数可以按照如下方式计算:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_M} = \sum_{(x_i, y_i) \in \mathcal{P}} (1 + \lambda_2) \left(\frac{\partial \mathcal{L}(x_i)}{\partial \mathbf{W}_M} + \frac{\partial \mathcal{L}(y_i)}{\partial \mathbf{W}_M} \right) + \sum_{(x_i, y_i) \in \mathcal{N}} \lambda_1 \left(\frac{\partial \mathcal{L}(x_i)}{\partial \mathbf{W}_M} + \frac{\partial \mathcal{L}(y_i)}{\partial \mathbf{W}_M} \right) + \mathbf{W}_M, \quad (7)$$

其中, $\frac{\partial \mathcal{L}(x_i)}{\partial \mathbf{W}_M}$ 表示图像 x_i 对于整个损失函数关于权重 $\mathbf{W}_M$ 的偏导数的贡献, $\frac{\partial \mathcal{L}(y_i)}{\partial \mathbf{W}_M}$ 表示图像 y_i 的贡献, 根据偏导的链式法则, 可以得到

$$\frac{\partial \mathcal{L}(x_i)}{\partial \mathbf{W}_M} = \frac{\partial \mathcal{L}}{\partial \mathbf{f}(x_i)} \frac{\partial \mathbf{h}_M(x_i)}{\partial \mathbf{W}_M}, \quad (8)$$

$$\frac{\partial \mathcal{L}(y_i)}{\partial \mathbf{W}_M} = \frac{\partial \mathcal{L}}{\partial \mathbf{f}(y_i)} \frac{\partial \mathbf{h}_M(y_i)}{\partial \mathbf{W}_M}, \quad (9)$$

其中, $\mathbf{h}_M(x)$ 的偏导数按如下方式计算:

$$\frac{\partial \mathbf{h}_M(x)}{\partial \mathbf{W}_M} = s'(\mathbf{W}_M \mathbf{h}_{M-1}(x) + \mathbf{b}_M) (\mathbf{h}_{M-1}(x))^T, \quad (10)$$

其中, s' 表示对函数 s 求导, $|\cdot|^T$ 表示矩阵的转置.

3.3 跟踪方法

在深度相对度量学习的基础上, 将深度相对学习的模型嵌入 Bayes 跟踪框架下, 整个跟踪流程包括 3 个部分: 训练阶段、测试阶段和更新阶段, 跟踪流程图如图 5 所示.

在训练阶段, 首先对任意需要跟踪的图像序列的第一帧进行初始化, 采用方框 (boundingbox) 手动标注需要跟踪的目标, 为了深度相对学习网络训练的需要, 利用初始化和相对位置结构来获得大尺度的图像块, 简单的, 可以随机的在图像序列的前几帧图像区域内采样大量样本. 当然, 随着跟踪的进行, 获得的样本会越来越多, 那么相对度量学习网络的训练也会越来越好.

测试阶段要解决的是一个新的图像帧到来后, 如何在新的图像帧里定位出初始化的目标区域, 本文采用粒子滤波随机采样样本的候选区域, 对每一个采样的候选图像块通过深度相对学习模型得到一

图 5 (网络版彩图) 跟踪方法流程图

Figure 5 (Color online) Flowchart of tracking method

个相对度量值, 那么在这些候选目标图像块中选择与初始化图像块相对度量值最优的作为目标的实际状态. 从 Bayes 的角度出发, 跟踪可以看成是一个状态的估计过程, 通过 Bayes 推理可以估计出目标的状态. 这里令 $\mathcal{X}_t$ 表示目标在第 t 帧的状态, 在获得一系列观测状态 $\mathcal{Y}_t = \{y_1, y_2, \dots, y_t\}$, 状态的后验概率可以通过 Bayes 得到估计:

$$p(\mathcal{X}_t|\mathcal{Y}_t) \propto p(\mathcal{Y}_t|\mathcal{X}_t) \int p(\mathcal{X}_t|\mathcal{X}_{t-1})p(\mathcal{X}_{t-1}|\mathcal{Y}_{t-1})d\mathcal{X}_{t-1}, \quad (11)$$

其中, $p(\mathcal{Y}_t|\mathcal{X}_t)$ 表示似然度函数; $p(\mathcal{X}_t|\mathcal{X}_{t-1})$ 则代表动态模型, 表示基于前一状态的当前状态的概率密度. $p(\mathcal{Y}_t|\mathcal{X}_t)$ 和 $p(\mathcal{X}_t|\mathcal{X}_{t-1})$ 决定了跟踪的整个过程, 粒子滤波通过一系列的权重采样估计出目标状态的概率分布. 本文采用图像块的深度相对度量学习值作为选择候选目标区域的指标, 则似然度函数定义为

$$p(\mathcal{Y}_t|\mathcal{X}_t) = \exp(-\gamma S_t), \quad (12)$$

其中, γ 是一个控制 Gauss 核函数的形状的常数, S_t 表示在第 t 帧某候选目标基于得到的相对度量学习值的指数衰减函数.

由于跟踪过程的目标表观受到光照、噪声等影响会发生很大的变化, 因此跟踪过程需要自适应的在每一帧中更新网络, 为了简单, 在更新阶段我们仅利用当前帧深度相对度量值来更新卷积神经网络.

4 实验分析与评价

为了对本文方法进行性能评测, 我们在两个相关的基准库上进行了测试. 一个是视觉跟踪器 2013 基准库 (benchmark) ^{[2]1)}, 它将最近几年在目标跟踪中常用的实验视频序列和跟踪方法的源代码进行了归纳整理. 2013 基准库中有 50 个有挑战性的视频用于算法跟踪性能的评测. 另外一个 VOT2014 视频库²⁾, 即视频跟踪竞赛基准库, 全部 25 个视频用于视频性能测评. 这些视频的基准结果 (ground truth) 都已经被人工标出. 本文选取的视频序列包含了复杂情况下跟踪的所有跟踪难度属性, 这也是跟踪问题难点所在, 其中难点属性分别包括表示复杂背景 (background clutter)、表示运动模糊 (motion blur)、表示尺度变化 (scale variation)、表示光照变化 (illumination variation)、表示遮挡 (occlusion)、表示快速运动 (fast motion)、表示平面内旋转 (in plane)、表示平面外旋转 (out off plane)、超出视角 (out of view) 等, 由于参与评测的视频太多, 就不一一列举单个视频的难点属性, 在后面的定性分析中

1) http://cvlab.hanyang.ac.kr/tracker_benchmark_v10.html.

2) <http://www.votchallenge.net/vot2014/index.html>.

会介绍部分视频的难点属性. 下面将展示评价方法效果的实验结果和一些当前最先进的方法进行比较的结果.

4.1 实验设定和评价指标

4.1.1 参数设定

本文实验的硬件环境为 multi-Core CPU, NVIDIA K20 GPU, 软件环境为 Ubuntu, Matlab 2013a, 速度性能为 1.3 帧/秒 (FPS). 本文基于深度学习库 Caffe 来实现深度相对度量学习模型. 在实验中, 数据输入层的图像在没有图片切分的情况下统一调整为 227×227 , 同时前 5 个卷积网络层是原始 Caffe 中提供的 AlexNet 的设置, 其余是我们新建的, 其中全连接层 $F_8 \sim F_{10}$ 的维数分别设定为 1000, 500, 1, 最后一个全连接层 F_{10} 的输入就是相对强弱值. 在相对损失函数层, 参数 α 、 β 、 λ_1 、 λ_2 分别设置为 1, 1, 1, 5×10^{-5} . 由于本文用于训练跟踪的图片属性相对不多, 为了避免过拟合, 把前 4 个卷积层的学习率设置为 0, 第 5 个卷积层、 F_6 、 F_7 的学习率设置为 0.001, $F_8 \sim F_{10}$ 的网络学习率定为 0.01. 学习率的设置与网络特征的层级有一定关系, 据文献 [46] 所知, 网络的底层卷积网络输出的是角点和边缘特征, 中层部分反映的是更为复杂的纹理特征, 顶层部分是输出类别相关的特征. Bayes 框架的粒子数设为 400, 似然度函数中 γ 为 1. 确定后的参数因子在整个实验中保持不变.

4.1.2 评价指标

定量度量一帧图像中跟踪框的准确度, 有两种常见的评价指标:

(1) 中心位置误差: 跟踪轨迹的中心位置与人工标记的基准结果中心位置的欧式距离.

(2) 重合率^[47]: 跟踪的跟踪框 r_T 与基准结果框 r_G 的交集面积与它们并集面积的比值, $s_{olp} = \frac{\text{area}(r_T \cap r_G)}{\text{area}(r_T \cup r_G)}$.

本文使用平均中心位置误差和平均重合率度量算法在一个视频上的跟踪效果, 根据成功率曲线^[2]评价算法在所有视频上的整体效果. 成功率曲线定义为将重合率大于给定阈值 θ_{olp} 的视频帧做为成功跟踪的帧, 当阈值 θ_{olp} 从 0 变化到 1 时, 记录成功跟踪的帧数占视频总帧数的比率, 得到成功率曲线, 使用成功率曲线下的面积 (AUC) 作为目标跟踪算法在所有视频上整体跟踪效果的度量. 根据基准库的设定, 本文使用一次通过评估 (one-pass evaluation, OPE) 策略与其他的当前先进的跟踪器进行比较.

4.2 实验结果与分析

4.2.1 比较的方法

我们将本文的目标跟踪算法与 Benchmark2013 里所有 29 个跟踪器进行了比较, 根据基准库里定量比较的标准, 排名靠前的 10 个跟踪器将用于排序显示, 其中包括³⁾: 使用结构跟踪器 (Struck)、循环结构核跟踪器 (CSK)、在线提升跟踪器 (OAB)、稀疏协同模型跟踪 (SCM)、跟踪器分解 (VTD)、采样跟踪器 (VTS)、上下文跟踪 (CXT)、正负约束学习跟踪器 (TLD)、多任务稀疏学习跟踪器 (MTT) 等. 在 VOT2014Benchmark 里, 我们与所有的 37 个跟踪器进行了比较, 根据基准库里定量比较的标准, 将所有的跟踪器结果进行了排序显示.

3) 各种跟踪器的缩写均来自于 Benchmark^[2], 为节省篇幅不再单独引用.

图 6 (网络版彩图) 总体排名前十的跟踪器整体成功率和精确度曲线

Figure 6 (Color online) (a) Overall success and (b) precision rate curves of rank top10

4.2.2 在 Bechmark2013 上定量的实验结果及分析

本文的跟踪算法在测试视频上的成功率曲线如图 6 所示, 图 6(a) 为测试视频上的整体成功率曲线, 图 6(b) 为测试视频的整体精确度曲线. 根据基准库的评价方法中一般取重叠比的阈值为 0.5 的位置来看跟踪成功率: 若某帧中心距离误差小于给定的阈值, 则认为该帧的跟踪结果是正确的, 文献 [2] 将该阈值设定为 20 个像素. 在图 6 测试视频中, 成功率达到了 0.741, 精确度达到了 0.852, 而且本文的跟踪器的成功率和精确度曲线处于排名前十的比较方法的上方, 具有最大曲线下的面积, 表明本文的跟踪器在测试视频集上具有较好的整体性能.

为了全面评估跟踪算法在不同难点属性上的性能, 图 7 和 8 给出了各个跟踪器在各个难点属性上成功率和精确度的分析曲线. 通过分析曲线可以看出, 在 9 个难点属性里, 在跟踪精度上本文提出的算法在以下 8 个属性: 复杂背景、变形、快速运动、光照变化、低分辨率、运动模糊、遮挡、平面外旋转取得了第一, 分别达到了 0.82, 0.908, 0.744, 0.768, 0.819, 0.72, 0.788, 0.842; 而在跟踪的成功率上本文提出的跟踪算法在 7 个属性: 复杂背景、变形、光照变化、低分辨率、遮挡、平面外旋转、超出视角取得了第一, 分别达到了 0.781, 0.797, 0.704, 0.800, 0.725, 0.729, 0.708; 本文所提算法的跟踪精度虽然在超出视角的属性上排在第二, 但是在跟踪的成功率上仍然取得了第一. 通过整体的实验结果可以看出, 本文算法在整个对比实验中有最好的整体跟踪性能, 在单个难度属性上也大部分取得了很好的效果, 虽然在部分属性上与相关算法相比有一点差距, 但本文算法无论从精度还是成功率上都有不错的平衡, 整体性能超过了参与比较的主流算法.

4.2.3 在 Bechmark2013 上定性的实验结果及分析

为了可视化的显示跟踪的效果, 本文结合部分测试视频的难点属性和部分视频的部分跟踪结果进行了针对性的分析, 如图 9 所示, 其余的部分跟踪结果如图 10 和 11 所示.

Tiger2. 该测试视频展示的是一个在实验室环境下的玩具在灯光变化以及绿色植物遮挡情况下的往复运动. 该视频的主要难点属性在于尺度变化、目标旋转、光照变化和目标的遮挡. 参与评测的跟踪器在视频的第 57 帧出现分化, 此时物体有微小的旋转, SCM 跟踪器已经偏移跟踪的目标; 此后目标在 100 到 300 帧之间由于上述跟踪难点的出现, 除了我们的跟踪器以及 CSK 和 Struck 外均发生了较大的偏移; 而后在 300 到 350 帧左右视频发生了较大的光照变化和遮挡, 即使一直跟踪不错的

图 7 (网络版彩图) 在不同跟踪视频属性上的精确度曲线

Figure 7 (Color online) Tracking precision rate curves on different attributes. (a) Background clutter; (b) deformation; (c) fast motion; (d) illumination variation; (e) low resolution; (f) motion blur; (g) occlusion; (h) out of plane rotation; (i) out of view

Struck 也偏离的跟踪目标, 而本文的跟踪算法则一直能够很好地处理出现的复杂情况, 有效地利用深度学习的强大判别能力将跟踪目标与背景判别出来. 从测试结果中可以看出本文算法在有尺度变化、目标旋转、光照变化和目标的遮挡等情况下仍然能够保证跟踪精度, 体现了本文跟踪算法的鲁棒性.

Skiing. 该测试视频展示的是一个在自然环境下的人的滑雪运动. 该视频的主要难点属性在于目标旋转、光照变化、复杂背景和运动模糊. 该视频长度不长, 但是难度不小. 参与评测的跟踪器在视频的第 10 帧出现分化, 此时由于复杂背景以及光照的变化, 绝大多数跟踪器已经偏移跟踪的目标; 此后帧中, 目标在 30 帧到达运动最高点, 除了我们的跟踪器外所有的跟踪器由于复杂背景及光照变化都跟踪到了稀疏的树上. 由于深度相对度量学习能够尽最大可能的挖掘目标与背景之间的相对性, 因而在后续的视频帧中可以看到我们的跟踪器能够很好地跟踪目标, 这表明本文的跟踪算法能够很好地处理复杂背景以及运动模糊的情形.

MotorRolling. 该测试视频展示的是山地车表演比赛. 该视频的主要难点属性在于目标旋转、尺度的巨大变化和运动模糊. 该视频在第 10 帧左右出现了跟踪器分化, 跟踪器 Struck、TLD 和本文的算法在当前都基本能够跟上, 而从此开始由于目标要发生较大的形变, 且摄像机的视角方向的不同,

图 8 (网络版彩图) 在不同跟踪视频属性上的成功率曲线

Figure 8 (Color online) Tracking success rate curves on different attributes. (a) Background clutter; (b) deformation; (c) fast motion; (d) illumination variation; (e) low resolution; (f) motion blur; (g) occlusion; (h) out of plane rotation; (i) out of view

使得目标在下降过程中尺度发生较大改变; 在 50 到 60 帧之间, 除了我们的跟踪器外所有的跟踪算法都丢失了; 从 60 帧可以看到虽然本文算法跟踪上了, 但是尺度上处理不是很理想, 导致跟踪有一些偏移, 但是跟踪算法能够保留样本的多样性, 使得跟踪方法在运动模糊发生时仍然能够保持足够的判别性. 从该视频全局来看, 我们的跟踪算法的跟踪效果良好, 虽然后来 TLD 由于其自身的重检测机制, 能够恢复到目标上, 但是效果并不理想.

Jogging1. 该视频展示了自然场景下的行人慢跑. 该视频的主要难点在于目标的遮挡和视频抖动. 虽然该视频是一个长帧视频, 但从跟踪效果来看难度不大. 从 60 到 85 帧之间由于跟踪目标在慢跑的过程经过灯柱发生了遮挡, 此时 MTT, VTD, VTS 和 Struck 偏离了跟踪目标, 它们或定位到灯柱上, 或跟踪发生错误, 其余的大部分跟踪算法则基本很好地处理了这个遮挡问题.

4.2.4 在 VOT2014Benchmark 上定量的实验结果及分析

本小节将展示 VOT2014Benchmark 上的实验结果与分析, 相关的实验评价细节详见文献 [48], VOT 作为一个目标跟踪的竞赛也是久负盛名, 它的评价指标主要有原始 (raw) 的排序、鲁棒性 (ro-

图 9 (网络版彩图) 部分视频跟踪结果 (I)

Figure 9 (Color online) Part of sequences tracking results (I)

图 10 (网络版彩图) 部分视频跟踪结果 (II)

Figure 10 (Color online) Part of sequences tracking results (II)

bustness) 排序和平均期望重叠率 (average expected overlap) 排序.

我们与所有跟踪器⁴⁾的比较实验结果展示在图 12(a) 中, 通过该图可以看出本文的跟踪算法的精确度排在了最上边的最好位置, 在精确度与鲁棒性之间展现了非常好的平衡性, 通过很少的跟踪失败获得了较高的精度. 在反映平均期望重叠率中 (图 12(b)), 本文算法的结果排在第三位, 仅次于竞赛官方设置的未公开发表的两个算法 PLT₁₃ 与 PLT₁₄. VOT2014Benchmark 为了展现跟踪算法在视觉难点属性的鲁棒性, 也设置了如下视觉属性: 摄像机运动 (camera motion)、光照变化 (illumination change)、遮挡 (occlusion)、尺度变化 (size change) 和运动变化 (motion change). 图 13 给出了跟踪

4) VOT 实验图的跟踪器的缩写均来自于 <http://www.votchallenge.net/vot2014/>.

图 11 (网络版彩图) 部分视频跟踪结果 (III)

Figure 11 (Color online) Part of sequences tracking results (III)

图 12 (网络版彩图) 整体的精确度和平均期望重叠率排序曲线

Figure 12 (Color online) (a) Overall precision and (b) average expected overlap rate rank curves

难点属性的鲁棒性排序结果. 通过图 13 可以看出本文的跟踪方法在所有属性上也显示了极好的平衡性能.

图 13 (网络版彩图) 在不同跟踪视频属性上的鲁棒性排序曲线
Figure 13 (Color online) The robustness rank curves on different attributes

5 结论

本文从解决合理的利用深度学习进行刻画大尺度样本相对结构关系的角度出发, 提出了一个有效的深度相对度量学习的目标跟踪算法, 不同于当前直接将前景与背景分类的方法, 本文尽最大可能挖掘前景与背景的结构相对关系, 并且能够保留样本的多样性, 这种跟踪目标与背景之间的结构差异恰是跟踪成功的关键. 本文基于卷积神经网络提出了深度相对度量学习模型, 通过该模型来建立有效的跟踪目标特性, 并将其应用到目标跟踪, 同时在基准跟踪库上进行了大量的评测实验. 通过实验结果可以看出本文提出的跟踪算法无论从跟踪精度和跟踪成功率上都超越了评测算法. 需要指出的是, 本文的重点在于深度相对度量学习跟踪算法, 仅简单地利用了 Bayes 跟踪框架, 下一步的工作首先将构建更加合理的卷积神经网络结构以提高算法的精度, 其次考虑重构 Bayes 跟踪框架, 改善跟踪更新模块的机制.

参考文献

- 1 Zhang S, Yao H, Sun X, et al. Sparse coding based visual tracking: review and experimental comparison. *Pattern Recogn*, 2013, 46: 1772–1788
- 2 Wu Y, Lim J, Yang M H. Online object tracking: a benchmark. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 2013. 2411–2418
- 3 Smeulders A W, Chu D M, Cucchiara R, et al. Visual tracking: an experimental survey. *IEEE Trans Pattern Anal Mach Intel*, 2014, 36: 1442–1468

- 4 Yang M, Pei M T, Wu Y W, et al. Learning online structural appearance model for robust object tracking. *Sci China Inf Sci*, 2015, 58: 032106
- 5 Bai Y C, Tang M. Robust tracking via weakly supervised ranking SVM. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 2012. 1854–1861
- 6 Jiang N, Liu W, Wu Y. Learning adaptive metric for robust visual tracking. *IEEE Trans Image Process*, 2011, 20: 2288–2300
- 7 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intel*, 2013, 35: 1798–1828
- 8 Fan J L, Xu W, Wu Y, et al. Human tracking using convolutional neural networks. *IEEE Trans Neural Netw*, 2012, 21: 1610–1623
- 9 Hu D, Zhou X S, Wu J J. Visual tracking based on convolutional deep belief network. In: *Proceedings of International Workshop on Advanced Parallel Processing Technologies*. Berlin: Springer, 2015. 103–115
- 10 Li H X, Li Y, Porikli F. DeepTrack: learning discriminative feature representations by convolutional neural networks for visual tracking. *IEEE Trans Image Process*, 2016, 25: 1834–1848
- 11 Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking. In: *Proceedings of International Conference on Advances in Neural Information Processing System*, Lake Tahoe, 2013. 809–817
- 12 Wang N Y, Shi J P, Yeung D Y, et al. Understanding and diagnosing visual tracking systems. In: *Proceedings of IEEE International Conference on Computer Vision*, Santiago, 2015. 3101–3109
- 13 Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking. *IEEE Trans Pattern Anal Mach Intel*, 2003, 25: 265–277
- 14 Collins R T. Mean-shift blob tracking through scale space. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Madison, 2003
- 15 Ross D, Lim J W, Lin R S, et al. Incremental learning for robust visual tracking. *Int J Comput Vision*, 2008, 77: 125–141
- 16 Zhang T Z, Ghanem B, Liu S. Robust visual tracking via structured multi-task sparse learning. *Int J Comput Vision*, 2013, 102: 367–383
- 17 Porikli F, Tuzel O, Meer P. Covariance tracking using model update based on Lie Algebra. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2006. 728–735
- 18 Li K, He F Z, Chen X. Visual tracking algorithm based on MAP multi-subspace incremental learning. *Sci Sin Inform*, 2016, 46: 476–495 [李康, 何发智, 陈晓. 基于 MAP 多子空间增量学习的目标跟踪算法. *中国科学: 信息科学*, 2016, 46: 476–495]
- 19 Zhou Y, Bai X, Liu W, et al. Similarity fusion for visual tracking. *Int J Comput Vision*, 2016, 118: 337–363
- 20 Zhang K H, Zhang L, Yang M H. Real-time compressive tracking. In: *Proceedings of European Conference on Computer Vision*, Florence, 2012. 866–879
- 21 Avidan S. Support vector tracking. *IEEE Trans Pattern Anal Mach Intel*, 2004, 26: 1064–1072
- 22 Avidan S. Ensemble tracking. *IEEE Trans Pattern Anal Mach Intel*, 2007, 29: 261–271
- 23 Collins R, Liu Y, Leordeanu M. Online selection of discriminative tracking features. *IEEE Trans Pattern Anal Mach Intel*, 2005, 27: 1631–1643
- 24 Grabner H, Bischof H. Online boosting and vision. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, New York, 2006. 260–267
- 25 Guo W, Cao L, Han T X, et al. Max-confidence boosting with uncertainty for visual tracking. *IEEE Trans Image Process*, 2015, 24: 1650–1659
- 26 Saffari A, Leistner C, Santner J, et al. Online random forests. In: *Proceedings of IEEE International Conference on Computer Vision Workshops*, Kyoto, 2009. 1393–1400
- 27 Babenko B, Yang M, Belongie S. Visual tracking with online multiple instance learning. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Miami, 2009. 983–990
- 28 Kalal Z, Matas J, Mikolajczyk K. “P-N” learning: bootstrapping binary classifiers by structural constraints. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, San Francisco, 2010. 49–56
- 29 Zhang J M, Ma S G, Sclaroff S. MEEM: robust tracking via multiple experts using entropy minimization. In: Pro-

- ceedings of European Conference on Computer Vision. Berlin: Springer, 2014. 188–203
- 30 Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 2544–2550
 - 31 Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of the 12th European Conference on Computer Vision, Florence, 2012. 702–715
 - 32 Henriques J F, Caseiro R, Martins P, et al. High-Speed tracking with kernelized correlation filters. *IEEE Trans Pattern Anal Mach Intel*, 2015, 37: 583–596
 - 33 Liu T, Wang G, Yang Q. Real-time part-based visual tracking via adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4902–4912
 - 34 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521: 436–444
 - 35 Wang N Y, Yeung D Y. Learning a deep compact image representation for visual tracking. In: Proceedings of Advances in Neural Information Processing Systems, Lake Tahoe, 2013. 809–817
 - 36 Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network. In: Proceedings of the 32nd International Conference on Machine Learning, Lille, 2015. 597–606
 - 37 Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015. 3119–3127
 - 38 Li H X, Li Y, Porikli F. Deeptrack: learning discriminative feature representations by convolutional neural networks for visual tracking. *IEEE Trans Image Process*, 2016, 25: 1834–1848
 - 39 Ma C, Huang J B, Yang X K, et al. Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, 2015. 3074–3082
 - 40 Cui Z, Xiao S T, Feng J S, et al. Recurrently target-attending tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1449–1458
 - 41 Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016. 850–865
 - 42 Tao R, Efstratiou G, Smeulders A W M. Siamese instance search for tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 1420–1429
 - 43 Parikh D, Grauman K. Relative attributes. In: Proceedings of International Conference on Computer Vision, Barcelona, 2011. 503–510
 - 44 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, 2012. 1097–1105
 - 45 Jia Y, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of ACM Conference on Multimedia, Orlando, 2014. 675–678
 - 46 Matthew D Z, Rob F. Visualizing and understanding convolutional networks. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2014. 818–833
 - 47 Everingham M, Van G L, Williams C K, et al. The pascal visual object classes challenge. *Int J Comput Vision*, 2010, 88: 303–338
 - 48 Cehovin L, Leonardis A, Kristan M. Visual object tracking performance measures revisited. *IEEE Trans Image Process*, 2016, 25: 1261–1274

Deep relative metric learning for visual tracking

Wen GUO^{1,2}, Sisi YOU^{1,2}, Junyu GAO³, Xiaoshan YANG³, Tianzhu ZHANG³ & Changsheng XU^{3*}

1. School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264009, China;
2. Key Laboratory of Sensing Technology and Control in Universities of Shandong, Yantai 264009, China;
3. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

* Corresponding author. E-mail: csxu@nlpr.ia.ac.cn

Abstract While traditional tracking-by-detection methods have some robustness in object tracking, their simple classification of the target and the background cannot model the relative structural relationship between the target and the background. It is the lack of relative structural discriminative information that always causes a tracker drifting away. In order to alleviate the tracking drifting problem, we propose a new visual travel approach based on deep relative metric learning. In this study, we design a deep relative metric learning model with a symmetric and shared-weight deep convolutional neural network. Through such a network, we can explore the relative structural relationship between the target and the background in large-scale image patches. Then, the highest score of relative metric is used to locate the tracking object in the Bayesian tracking framework. The whole tracking algorithm is simple and effective. Experimental results on the tracking benchmark show that the proposed algorithm achieves a better tracking precision rate and success rate than other state-of-the-art tracking methods.

Keywords relative attribute, metric learning, convolutional neural network (CNN), visual tracking

Wen GUO was born in 1978. He received his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2012. Currently, he is an associate professor at Shandong Technology and Business University. His research interests include computer vision and multimedia.

Sisi YOU was born in 1994. She received her bachelor's degree in electrical science from Shandong Technology and Business University, Yantai, in 2016. Currently, she is a master candidate at Shandong Technology and Business University. Her research interest lies in computer vision.

Tianzhu ZHANG was born in 1982. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2011. Currently, he is an associate professor at the Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision and multimedia.

Changsheng XU was born in 1969. He received his Ph.D. degree from Tsinghua University, Beijing, in 1996. Currently, he is a professor at the Institute of Automation, Chinese Academy of Sciences. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision.