融合显著性与运动信息的相关滤波跟踪算法

张伟俊 [1, 2] 钟 胜 [1, 2] 徐文辉 [1, 2] WU Ying [3]

摘 要 主流的目标跟踪算法以矩形模板的形式建立被跟踪物体的视觉表征, 无法有效区分目标与背景像素, 在背景复杂、目标非刚体形变、复杂运动等挑战性因素影响下容易出现模型偏移的问题, 导致跟踪失败. 与此同时, 像素级的显著性信息与运动先验信息作为人类视觉系统有效区分目标与背景、识别运动物体的重要信号, 并没有在主流目标跟踪算法中得到有效的集成利用. 针对上述问题, 提出目标的像素级概率性表征模型, 并且建立与之对应的像素级目标概率推断方法, 能够有效利用像素级的显著性与运动观测信息, 实现与主流的相关滤波跟踪算法的融合; 提出基于显著性的观测模型, 通过背景先验与提出的背景距离模型, 能够在背景复杂的情况下得到高辨识度的像素级图像观测; 利用目标与相机运动的连续性来计算目标和背景的运动模式, 并以此为基础建立基于运动估计的图像观测模型. 实验结果表明, 提出的目标表征模型与融合方法能够有效集成上述像素级图像观测信息, 提出的跟踪方法总体跟踪精度优于多种当下最先进的跟踪器, 对跟踪场景中的背景复杂、目标形变、平面内旋转等挑战性因素具有更好的鲁棒性.

关键词 视觉目标跟踪, 运动分析, 显著性检测, 像素级概率模型, 相关滤波

引用格式 张伟俊, 钟胜, 徐文辉, WU Ying. 融合显著性与运动信息的相关滤波跟踪算法. 自动化学报, 2021, **47**(7): 1572−1588

DOI 10.16383/j.aas.c190122

Correlation Filter Based Visual Tracking Integrating Saliency and Motion Cues

ZHANG Wei-Jun[1, 2] ZHONG Sheng[1, 2] XU Wen-Hui[1, 2] WU Ying[3]

Abstract Rectangle template is a popular target representation adopted by mainstream visual tracking methods. However, by including some background clutter as part of the target representation, the model is likely to drift away from the target gradually and result in tracking failure, especially in challenging situations such as background clutter, target deformation and complex motions. Meanwhile, motion and saliency cues, which play important roles in distinguishing targets from the background and identifying moving objects in the human vision system, have not been modeled into existing tracking methods. To solve these problems, we propose a foreground probabilistic inference formulation that collects pixel-level observations from different sources, and a unified framework integrating the pixel-level model with a widely used correlation filter based method. A saliency-based observation model is proposed by introducing background prior and a distance-based model, which provides reliable evidence to resolve confusion caused by appearance similarity between targets and the background. By taking advantage of continuity and inertia of both target and camera motion, we discover motion patterns in the spatial domain to distinguish targets from the background, and introduce a pixel-level motion-based observation model. Experiments demonstrate that the proposed method outperforms some of the state-of-the-art methods, and shows better robustness in challenging situations such as background clutter, target deformation and in-plane rotation.

Key words Visual tracking, motion analysis, saliency detection, pixel-level probabilistic model, correlation filter

Citation Zhang Wei-Jun, Zhong Sheng, Xu Wen-Hui, Wu Ying. Correlation filter based visual tracking integrating saliency and motion cues. *Acta Automatica Sinica*, 2021, **47**(7): 1572−1588

随着越来越多的智能机器的普及应用, 计算机视觉作为机器的 “眼睛”, 担负着感知和理解外部世界的功能, 成为一项迫切的需求. 视觉目标跟踪[1−2]的主要任务是在视频图像序列中建立目标的运动轨迹, 在智能视频监控[3]、自动驾驶[4]、人机交互[5]、机器人导航[6]、医学诊断[7] 等领域均有广泛的应用. 这些上层算法应用的性能很大程度上受限于目标跟踪算法的性能, 因此提高目标跟踪算法的鲁棒性、准确

收稿日期 2019-03-03 录用日期 2019-07-30
Manuscript received March 3, 2019; accepted July 30, 2019
国家重点研发计划 (2016YFF0101502) 资助
Supported by the National Key Research and Development Program of China (2016YFF0101502)
本文责任编委 赖建煌
Recommended by Associate Editor LAI Jian-Huang

1. 华中科技大学人工智能与自动化学院 武汉 430074 中国 2. 华中科技大学多谱信息处理技术国家级重点实验室 武汉 430074 中国 3. 美国西北大学电子工程与计算机系 埃文斯顿 伊利诺伊州 60208 美国
1. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China 2. National Key Laboratory of Science and Technology on MultiSpectral Information Processing, Huazhong University of Science and Technology, Wuhan 430074, China 3. Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208, USA

率与实时性, 能够为各领域的发展提供必要的技术支撑与理论促进, 具有重大的意义.

在视觉目标跟踪技术的众多分支中, 针对通用物体的在线目标跟踪技术由于不需要使用预训练的物体模型, 对跟踪任务执行的场景、被跟踪物体的类别、形状、运动模式均无特殊的限定与要求, 存在极其广泛的应用需求, 因此成为众多计算机视觉系统与应用的底层关键技术之一, 近十年来一直是计算机视觉领域中一个非常活跃的研究课题. 与此同时, 由于存在目标及场景先验知识缺乏, 物体及环境变化不可预测等诸多因素, 与已知物体类别的跟踪[8−9] 相比, 对建模方法的适应性有着更高的要求. 要长时间准确定位目标, 算法必须适应目标及场景的各种变化, 典型的变化包括目标尺度变化、非刚体形变、背景干扰、快速运动与复杂运动等, 这些都给通用物体的在线跟踪任务带来了极大的挑战. 尽管近年来在理论和应用上均取得了显著的进展[10−11], 在线目标跟踪的研究仍有很多关键问题亟待解决, 其中之一是被跟踪物体的表征与建模, 即目标表征问题.

无论是经典的生成式模型[12−13] (Generative model), 还是近年来较为主流的判别式模型[14−15] (Discriminative model) 以及基于深度学习的方法[16−18] 都使用了外接目标区域的矩形模板来表征被跟踪目标. 虽然这些算法在刚性物体跟踪上取得了很好的效果, 但是大部分缺乏能够十分有效区分目标与背景像素的机制. 由于目标模型里包含了一部分背景区域, 随着噪声和误差的累积, 模型容易慢慢偏移到背景上面去, 同时也比较难对目标的形状变化实现自适应调整, 在背景干扰、目标形变明显或者复杂运动的场景下容易逐渐丢失目标.

与此不同的是, 人类视觉系统能够明确地区分目标与背景的区域, 并不以矩形模板的形式表征和建模物体. 研究表明, 人类的视觉机制具有异常突出的数据筛选能力, 能够快速有效地识别复杂场景中的显著性区域, 准确定位感兴趣的目标[19−20]. 人类能够轻松实现对目标的稳定跟踪, 视觉注意机制扮演了重要的角色. 因此, 在目标跟踪算法中建模显著性机制, 对其提供的像素级观测信息进行集成利用, 以提高跟踪算法的鲁棒性与准确率, 具有重要的意义.

人类处理运动物体的另一个特点是具备关于运动的先验知识, 知道属于同一个物体的像素有同样的运动趋势. 认知与心理学的研究[21−22] 表明, 几个月大的婴儿就已经有关于自由物体连续和平滑运动的知识, 能够根据这些知识辅助预测和判断物体的走向. 这些关于物体显著性和运动的知识, 目前都没有在目标跟踪方法中被很好地建模与集成利用.

上述像素级先验信息没有被有效利用, 一个重要的原因是当前主流的目标跟踪模型使用了基于矩形模板的目标表征模型, 无法有效地融合这些像素级的图像观测. 因此, 本文提出使用像素级概率性目标表征模型, 将目标跟踪任务建模为一个像素级目标概率的贝叶斯推断 (Bayesian inference) 问题, 在每一帧使用前后帧的像素关联来向前传递目标概率, 再进一步融合当前帧显著性模型和运动观测模型提供的像素级图像证据, 递推地产生目标概率图. 该模型提供了与当前主流矩形模板目标表征模型互补的信息, 可以用来预测目标位置, 与使用矩形模板目标表征的算法进行融合决策, 提升目标跟踪算法在背景干扰、目标形变、复杂运动等场景下的鲁棒性. 同时, 像素级的目标概率图也可产生目标分割结果, 为视频目标分割、增强现实以及行为分析等应用和研究提供帮助.

1 相关工作

目前较为主流的视觉跟踪算法使用判别式模型, 在已跟踪图像序列上采集目标与非目标样本训练分类器, 通过对新图像上采样的候选目标矩形框进行分类判决来完成跟踪任务[14−15], 也被一些研究者称为检测−跟踪 (Tracking-by-detection) 框架. 其中, 基于岭回归 (Ridge regression) 分类器的算法由于可以利用循环矩阵的特性, 将空间域的训练样本转换到频域进行加速计算, 得到基于相关滤波 (Correlation filter, CF) 的算法实现[14], 具备算法速度与准确率俱佳的特点, 吸引了大量的研究和改进工作[23−25].

为了避免基于矩形模板的目标表征模型受到目标形变、背景噪声以及误差累积的影响, 导致算法目标模型偏移的问题, 一些跟踪算法[26−27] 采用基于子块的模型 (Part-based model) 来进行目标表征, 以减少背景区域对模型的干扰, 对目标形变和遮挡等常见挑战性因素具有一定的自适应能力. 但是, 相对于单个矩形模板的表征方法, 基于子块的目标表征模型存在参数较多, 模型较为复杂, 需要灵活处理如何选择和更新子块等问题, 在长时间跟踪 (Long-term tracking) 过程中仍然无法保持足够的算法鲁棒性, 限制了其进一步的应用.

另一类跟踪算法将目标分割引入到跟踪过程中, 得到像素级的目标模型, 目标表征更为精确. Fan 等使用了抠图技术 (Image matting) 对目标前背景进行分割, 并把分割结果反馈到跟踪过程中[28].

Godec 等使用 Grabcut 算法进行目标分割, 并在每帧使用分割结果指导下一帧的检测[29]. Bibby 等使用水平集 (Level set) 进行目标的分割, 以处理目标形状变化[30]. 这一类算法存在的问题是, 模型极大依赖于图像分割算法的鲁棒性, 在背景干扰严重的情况下, 单帧的分割误差对后续操作影响较大, 容易循环积累, 导致后面的模型出现偏差.

上述基于分割的目标表征方法对每个像素是否在目标物体上做出了确定性的判决, 像素级目标概率模型则在此基础上进一步改进, 对像素点是否在目标上进行概率性的估计. 这样的模型相当于对目标进行了软分割, 在建立了像素级目标模型的同时, 对于分割误差有更高的容忍度. Oron 等对目标进行了像素级的建模, 并把像素概率推断融入到 Lucas-Kanade 目标跟踪框架之中[31]. Possegger 等针对目标和背景分别建立了颜色直方图来作为分类器, 对每个像素给出目标概率推断[32]. Son 等使用了梯度提升决策树算法 (Gradient boosting decision Tree, GBDT) 来作为分类器给出目标与背景的分类[33]. Duffner 等则综合使用霍夫投票 (Hough voting) 与颜色直方图进行像素分类器的建模[34].

在文献 [34−36] 中目标跟踪与分割问题被联合建模, 不同来源的像素级图像特征由一个贝叶斯推断框架进行融合. 虽然这些工作和本文的贝叶斯推断方法有相似之处, 但在传递概率和像素级似然概率的建模方式等方面, 都和本文提出的方法有很大的区别. 此外, 贝叶斯推断也被广泛应用到多物体估计[37]、识别[38] 与跟踪[39] 问题当中. 在这些任务中运动信息通常被用来关联不同帧之间检测到的目标.

近年来, 视觉显著性检测[19−20] 作为一项新兴的课题, 吸引了大量的研究, 它通过模拟人类视觉注意机制对图像信息进行筛选处理, 选取优先处理区域, 提供给其他较上层的计算机视觉算法进行使用. 显著性检测领域的代表算法包括由 Itti 等提出的基于空间域的计算模型[40]、Hou 等提出的基于频谱域的方法[41] 等. 此外, 基于测地距离 (Geodesic distance) 和基于最小障碍距离 (Minimum barrier distance, MBD) 的显著性算法[42−43] 使用背景先验和距离度量来衡量像素点的显著性, 在数据集上取得了很好的效果, Zhang 等在此基础上提出的加速算法[44] 由于较快的计算速度和出色的检测效果而受到关注. 一些研究者尝试将视觉显著性计算模型引入跟踪系统中, 通过模拟人类特有的视觉选择性注意机制为采样提供先验知识, 从而提高跟踪效率[45].

目标跟踪领域的另一个近期发展趋势是深度学习与卷积神经网络 (Convolutional neural network, CNN) 技术的应用. 一部分研究者在判别式跟踪模型框架内通过使用对目标表征能力更强的 CNN 特征, 来获得更好的跟踪效果[46−48], 其中 Choi 等通过对 CNN 特征进行压缩来保证算法的实时性[48]. 另外一部分研究者则通过构造和训练端对端 (End-to-end) 的卷积神经网络来完成跟踪任务, 其中 Bertinetto 等提出的全卷积孪生神经网络 (Fully-convolutional siamese networks, SiamFC) 是十分具有代表性的工作[16], Valmadre 等提出的 CFNet 算法在此基础上将相关滤波器建模为深度神经网络的一个层[17], 使得算法集成了深度学习与相关滤波技术的优点. 此外, Hong 等使用卷积神经网络特征通过后向传播 (Back-projecting) 技术构造目标的显著图[46], Choi 等在跟踪算法中通过训练深度回归网络 (Deep regression network) 建立注意力机制[18], Gladh 等在基于深度学习方法的跟踪框架内引入了深度运动特征[49], 这些工作虽然采用了与我们截然不同的建模方式, 但是与本文具有相似的出发点, 认为注意力机制以及基于运动的图像观测能够提供与现有模型呈现互补性的信息, 从而有效提升目标跟踪算法的精度与鲁棒性.

2 多目标表征融合跟踪框架

本文使用了基于检测器的目标跟踪框架, 其核心思想是根据已跟踪的目标采集正负样本训练检测器, 在待跟踪图像上通过一定的预测搜索策略产生大量的候选样本, 使用之前训练的检测器对这些样本是目标物体的概率进行估计, 选取最佳选项作为跟踪输出结果.

具体的, 在第 t 帧的时候, 在图像 I^t 中采集大量候选目标样本形成集合 Q^t, 从中选择一个作为目标矩形框 p^t, 以使得目标函数最大化:

$$p^t = \arg\max_{p \in Q^t} f(T(I^t, p); \theta^t) \tag{1}$$

其中, $T(I, p)$ 是一个图像变换, 对图像 I 中的矩形窗口 p 提取一定的特征描述符, 构成目标的视觉表征, 评估函数 $f(T(I, p); \theta)$ 再对视觉表征数据根据模型参数 θ 赋值一个分数. 第 t 帧的模型参数 θ^t 根据之前帧的图像观测与目标位置的集合 $\{(I^i, p^i)\}_{i=1}^{t-1}$ 来进行选择.

在每一帧, 目标跟踪问题的核心转化为评估函数 $f(I, p)$ 的构造与求解. 为了融合互补的跟踪模型, 充分利用不同类型图像特征和目标表征方式的优势, 把评估函数 $f(I, p)$ 设置为两个分数的线性组合, 两个分数 $f_{\text{pxl}}(I, p)$ 和 $f_{\text{tmpl}}(I, p)$ 分别基于像素级目标

表征模型和矩形框表征模型来进行计算, 加权系数分别为 γ 和 $1-\gamma$:

$$f(I,p)=\gamma f_{\text{pxl}}(I,p)+(1-\gamma)f_{\text{tmpl}}(I,p) \tag{2}$$

图 1 给出了多目标表征模型融合跟踪框架的示意图. 基于矩形框目标表征的相关滤波器模型、基于像素级概率性目标表征的运动模型和显著性模型均通过上一帧 (训练帧 I^{t-1}) 提供的目标邻域图像数据进行模型训练, 在当前帧 (测试帧 I^t) 对搜索区域中的候选目标框位置进行评估分数的求解. 其中, 相关滤波器模型与当前帧数据直接进行求解可得到评估分数 $f_{\text{tmpl}}(I^t,p)$; 运动模型和显著性模型结合当前帧图像数据求解得到像素级目标似然概率图, 再进一步通过本文提出的转化方法得到 $f_{\text{pxl}}(I^t,p)$. 两种目标表征模型的评估分数线性融合之后, 应用式 (1) 定位最优的目标位置 p^t.

2.1　基于像素级目标表征的评估分数

基于像素级的目标表征模型计算候选样本的评估分数 $f_{\text{pxl}}(I,p)$ 时, 可使用该样本矩形框内每个像素的目标概率来进行估计. 具体的, 对于已知目标位置, 矩形目标框 p 内的像素位置集合记为 $H\subset\mathbf{Z}^2$, 目标概率函数 $\psi_I(\boldsymbol{x}):H\to\mathbf{R}$ 是从每个像素位置 $\boldsymbol{x}$ 到目标概率分数的映射, 定义评估函数为:

$$f_{\text{pxl}}(I,p)=\frac{1}{|H|}\sum_{\boldsymbol{x}\in H}\psi_I(\boldsymbol{x}) \tag{3}$$

即取目标框内像素目标概率分数的平均值来作为目标框的评估分数. 像素点目标概率函数 $\psi_I(\boldsymbol{x})$ 的建模与求解是本方法的重点, 具体在第 4 节进行详细的阐述.

在实际算法实现中, 我们取搜索区域 Ω, 计算搜索区域上每个像素位置到目标概率分数的函数映射 $\psi_I(\boldsymbol{x})$. 再通过积分图像[50] (Integral image) 计算得到密集采样的每个目标框位置的目标函数 $f_{\text{pxl}}(I,p)$. 具体的, 在第 t 帧计算目标函数 $f_{\text{pxl}}(I,p)$ 的时候, 根据第 $t-1$ 帧已经估算的目标中心位置 $\boldsymbol{x}=[x,y]^{\text{T}}$ 以及尺度大小 $[d_1,d_2]^{\text{T}}$, 取中心位置为 $[x,y]^{\text{T}}$, 尺度大小为 $\left[d_1+\alpha\sqrt{d_1d_2},d_2+\alpha\sqrt{d_1d_2}\right]^{\text{T}}$ 的矩形区域作为搜索区域 Ω 来进行积分图像的计算 (如图 1 左下图中的黑色虚线框所示), 其中系数 α 控制了搜索区域相对于目标尺度的扩大比例, 经验性地设置为 1. 每一帧根据目标函数 $f(I,p)$ 和式 (1) 确定新的目标位置之后, 使用判别式尺度空间跟踪器[23] (Discriminative scale space tracking, DSST) 进行新的目标尺度的估算.

2.2　基于矩形框目标表征的评估分数

在矩形框目标模型中, 对于矩形目标框 p 内的采样网格点集合 $T\subset\mathbf{Z}^2$, 定义 D 通道的特征图像 $\phi_I:T\to\mathbf{R}^D$, 基于矩形框模型的分数定义为特征图像 ϕ_I 的一个线性组合:

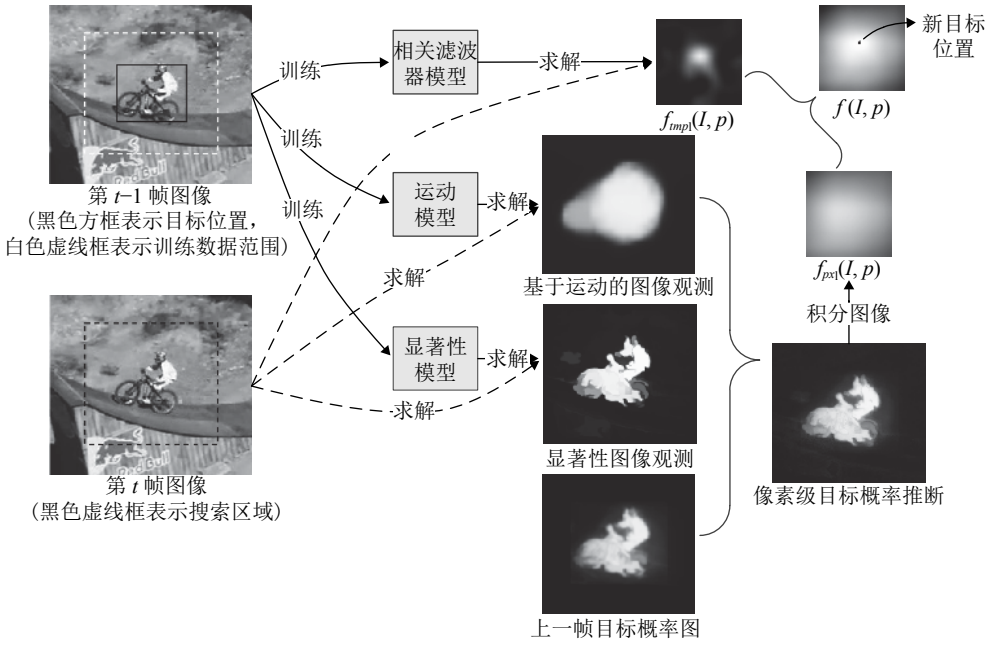


图 1　总体跟踪流程图

Fig. 1　Overall tracking procedure

$$f_{\text{tmpl}}(I,p;h)=\sum_{\boldsymbol{x}\in T}(h[\boldsymbol{x}])^{\mathrm{T}}\phi_I(\boldsymbol{x}) \tag{4}$$

在这里使用了常见的基于矩形框目标表征的判别式相关滤波器[23]进行统计建模, 参数 $h\in\mathbf{R}^{|T|\times D}$ 对应于模型中的滤波器参数, $h[\boldsymbol{x}]$ 表示其在像素位置 $\boldsymbol{x}$ 处对应的长度为 D 的向量. 本文使用步长为 4 的特征网格, 提取方向梯度直方图 (Histogram of oriented gradient, HOG) 特征来与基于矩形框的目标表征模型配合使用.

该算法利用目标图像平移产生的循环样本近似表示密集采样的训练样本, 使用循环的样本数据集来训练岭回归分类器. 岭回归的本质是一种加入正则化的最小二乘法, 对病态数据有很好的拟合能力. 假设训练样本特征图像记为 f , 其第 l 个通道特征表示为 $f^l, l\in\{1,\cdots,D\}$. 记相关滤波器为 h , 由 D 个单通道滤波器 h^l 组成. 多通道相关滤波器算法的目标是最小化相关滤波响应结果与期望的输入响应结果 g 之间的 L_2 残差, 即

$$\varepsilon=\left\|\sum_{l=1}^{D}h^l\star f^l-g\right\|^2+\lambda\sum_{l=1}^{D}\|h^l\|^2 \tag{5}$$

其中, $\star$ 表示循环相关操作, g 表示相关滤波训练输入, 由一个峰值位于 f 中心的高斯函数生成, 表示通过循环移位得到的训练样本与目标重叠度高的取为正样本, 偏离目标较远的取为负样本. 公式后半部分是一个权重系数为 λ 的正则化项, 用来防止过拟合.

上式是一个线性最小二乘问题, 通过求偏导并化简即可计算得到分类器参数 h^l , 其闭式解为:

$$h^l=\left(\sum_{k=1}^{D}{f^k}^{\mathrm{T}}f^k+\lambda U\right)^{-1}{f^l}^{\mathrm{T}}g \tag{6}$$

其中, ${f^l}^{\mathrm{T}}$ 表示矩阵 f^l 的转置, U 为单位矩阵.

将时域的卷积转化为频域的点乘, 能极大地降低计算量, 保障算法的实时性. 文献 [51] 通过应用帕塞瓦尔公式 (Parseval′s formula) 以及离散傅里叶变换的特性, 推导得到相关滤波器 h^l 在频域内的闭式解为

$$H^l=\frac{\overline{G}F^l}{\sum\limits_{k=1}^{D}\overline{F^k}F^k+\lambda} \tag{7}$$

式中, 大写字母表示相应变量的离散傅里叶变换, 乘法除法均为矩阵对应元素相乘或相除, $\overline{F^k}$ 和 $\overline{G}$ 分别表示 F^k 和 G 的复数共轭形式.

在仅考虑单个目标样本的情况下, 式 (7) 进行离散反傅里叶变换给出了最优滤波器 h . 在目标跟踪过程中, 目标的外观会发生变化, 为了能持续跟踪目标, 需要考虑不同时刻 t 的目标样本 $\{f_i\}_{i=1}^{t}$, 对滤波器进行在线更新. 在第 t 帧图像上进行目标跟踪时, 相关滤波器 h 在频域内的更新公式为

$$A_t^l=(1-\eta)A_{t-1}^l+\eta\overline{G}{F_t}^l \tag{8}$$

$$B_t^l=(1-\eta)B_{t-1}^l+\eta\sum_{k=1}^{D}\overline{{F_t}^k}{F_t}^k \tag{9}$$

其中, A_t^l 和 B_t^l 分别是式 (7) 中滤波器 H_t 的分子和分母, η 是模型的更新系数.

在第 t 帧的时候, 使用密集采样的搜索策略, 在搜索区域上采集候选样本, 远离目标中心的样本通过目标中心样本循环移位产生, 则通过相关滤波器输出可一次性计算每个候选样本的评估分数. 对搜索区域提取相应的特征图像, 转化成频域特征图像 Φ_t , 可以方便地在频域内计算相关滤波器输出, 再转换为时域滤波结果, 即

$$\hat{G}=\frac{\sum\limits_{l=1}^{D}\overline{A_{t-1}^l}\Phi_t^l}{B_{t-1}+\lambda} \tag{10}$$

得到的频域滤波器输出 $\hat{G}$ 进行离散傅里叶反变换即可得到相应位置的评估分数.

3 像素级概率性目标表征模型

第 2.1 节提出基于像素级概率性目标表征模型 $\psi_I(\boldsymbol{x})$ 来构造评估函数. 本节对 $\psi_I(\boldsymbol{x})$ 的建模与求解进行详细的介绍.

3.1 像素级贝叶斯推断框架

以 $\boldsymbol{x}=[x,y]^{\mathrm{T}}$ 记搜索区域 Ω 中的像素位置, 使用 $\boldsymbol{z}_{\boldsymbol{x}}^t$ 表示像素位置 $\boldsymbol{x}$ 在第 t 帧的图像观测向量, $c_{\boldsymbol{x}}^t\in\{0,1\}$ 表示像素位置 $\boldsymbol{x}$ 在第 t 帧的类别 (以 0 表示背景, 1 表示目标), $\boldsymbol{z}_{\boldsymbol{x}}^{1:t}$ 表示从 $\boldsymbol{z}_{\boldsymbol{x}}^1$ 到 $\boldsymbol{z}_{\boldsymbol{x}}^t$ 的图像观测的集合. 对于每个像素位置 $\boldsymbol{x}$, 属于类别 $C\in\{0,1\}$ 的概率通过当前帧以及之前帧的所有图像观测进行推断. 有别于传统的对候选目标框是目标的概率进行估计的做法, 像素级概率性目标表征模型对搜索区域的像素点在目标上的概率进行估计, 将目标跟踪任务建模为一组并行的像素级目标概率估计问题, 如图 2 所示. 概率的推断通过递推的贝叶斯模型完成, 分为预测过程和更新过程, 分别集成动态模型与观测模型.

在预测过程中, 上一时刻的后验概率分布 $p\left(c_{\boldsymbol{x}'}^{t-1}|\boldsymbol{z}^{1:t-1}\right)$ 通过动态模型 $p\left(c_{\boldsymbol{x}}^t|c_{\boldsymbol{x}'}^{t-1}\right)$ 转移得到当前时刻的预测概率分布:

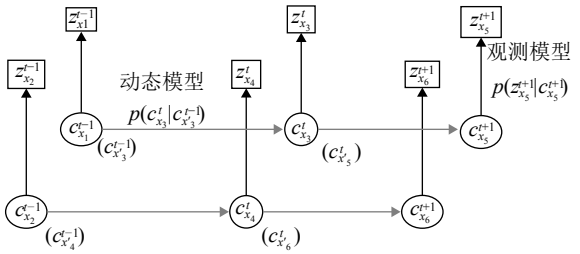


图 2　像素级目标概率推断模型的贝叶斯网络示意图

Fig. 2　Bayesian network representation of pixel-level target probabilistic inference model

$$p\left(c_{\boldsymbol{x}}^{t}=C|\boldsymbol{z}^{1:t-1}\right)= \int p\left(c_{\boldsymbol{x}}^{t}=C|c_{\boldsymbol{x}'}^{t-1}\right)p\left(c_{\boldsymbol{x}'}^{t-1}|\boldsymbol{z}^{1:t-1}\right)\,\mathrm{d}c_{\boldsymbol{x}'}^{t-1} \tag{11}$$

其中, $\boldsymbol{x}'$ 表示第 t 帧的像素位置 $\boldsymbol{x}$ 在第 $t-1$ 帧对应的像素位置.

更新过程是得到当前时刻的观测信息 $\boldsymbol{z}^t$ 之后, 将式 (11) 计算得到的预测概率分布根据目标的观测方程进行更新, 得到:

$$p\left(c_{\boldsymbol{x}}^{t}=C|\boldsymbol{z}^{1:t}\right)\propto p\left(\boldsymbol{z}_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}}^{t}=C\right)p\left(c_{\boldsymbol{x}}^{t}=C|\boldsymbol{z}^{1:t-1}\right) \tag{12}$$

在本方法中像素级的目标观测概率 $p\left(\boldsymbol{z}_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}}^{t}\right)$ 假定由独立的两部分组成, 一部分来自显著性检测模型, 另一部分来自基于运动估计的模型, 即

$$p\left(\boldsymbol{z}_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}}^{t}\right)\approx p_{\mathrm{S}}\left(D^{t}(\boldsymbol{x})|c_{\boldsymbol{x}}^{t}\right)p_{\mathrm{M}}\left(M^{t}(\boldsymbol{x})|c_{\boldsymbol{x}}^{t}\right) \tag{13}$$

其中, $D^t(\boldsymbol{x})$ 表示像素位置 $\boldsymbol{x}$ 在时间 t 的背景距离测度, $M^t(\boldsymbol{x})$ 表示像素位置 $\boldsymbol{x}$ 在时间 t 的后向光流矢量, 分别为两个子观测模型在第 t 帧的观测量.

本文提出的像素级模型对图像中的运动信息有两个方面的利用. 一方面, 时间上的运动关联通过动态模型 $p\left(c_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}'}^{t-1}\right)$ 来进行表示, 在第 3.2 节具体阐述; 另一方面, 空间上的运动关联通过基于运动估计的观测模型 $p_M\left(M^{t}(\boldsymbol{x})|c_{\boldsymbol{x}}^{t}\right)$ 来进行建模, 在第 3.4 节具体阐述. 另外, 基于显著性信息的像素级模型 $p_{\mathrm{S}}\left(D^{t}(\boldsymbol{x})|c_{\boldsymbol{x}}^{t}\right)$ 在第 3.3 节具体展开. 这三方面的信息可分别理解为基于时间域运动连续性的显著性估计、基于空间域运动连续性的显著性估计、以及基于本帧空间域信息的显著性估计, 三者通过一个统一的贝叶斯推断框架进行融合, 构成完整的时空显著性模型. 以下对每一部分进行详细介绍.

3.2　动态模型

动态模型 $p\left(c_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}'}^{t-1}\right)$ 建模了当前帧与上一帧像素之间的运动关联, 通过最新的光流估计算法[52] 来获得, 在式 (11) 的预测过程中使用. 算法给出了亚像素级别的光流匹配结果, 记像素位置 $\boldsymbol{x}$ 在第 t 帧的后向光流场为 $M^t(\boldsymbol{x})$, 则像素位置 $\boldsymbol{x}$ 的在第 $t-1$ 帧在对应的位置为:

$$\boldsymbol{x}'=\boldsymbol{x}+M^{t}(\boldsymbol{x}) \tag{14}$$

由于 $\boldsymbol{x}'$ 可以是亚像素级别的位置, 而且对于搜索区域中的像素 $\boldsymbol{x}\in\Omega$ 已经有上一帧的像素级目标概率估计 $p\left(c_{\boldsymbol{x}}^{t-1}|\boldsymbol{z}^{1:t-1}\right)$, 因此式 (11) 中像素位置 $\boldsymbol{x}'$ 的上一帧目标概率 $p\left(c_{\boldsymbol{x}'}^{t-1}|\boldsymbol{z}^{1:t-1}\right)$ 可以通过插值来计算获得.

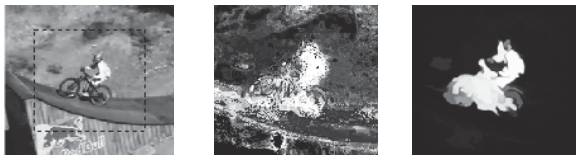
状态传递概率 $p\left(c_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}'}^{t-1}\right)$ 代表了对前一帧估计结果的置信程度, 理论上应当大于 0.5, 值越大代表上一帧的结果置信程度越高. 实验发现区间 $[0.8,1]$ 内的值对于实验结果并没有大的影响, 在本文中采用了 $p\left(c_{\boldsymbol{x}}^{t}|c_{\boldsymbol{x}'}^{t-1}\right)=1$ 即最高的置信程度.

Duffer 等提出的方法[34] 也使用了像素级贝叶斯模型, 不过他们的方法并没有显式地建模帧与帧之间的像素关联, 而是假定像素位置之间是独立的, 以简化计算复杂度. 这样的假设仅仅适用于帧率较高, 帧间目标相对位移很小的场景, 导致该方法无法很好地处理目标形变、快速运动等更复杂的情况.

3.3　基于显著性的观测模型

在目标跟踪任务中, 常见的像素级观测模型是基于颜色特征, 对于目标和背景在线地建立统计直方图[32] 或者训练分类器[29]. 这样的模型很容易受到背景噪声的影响 (如图 3(b) 所示), 在目标与背景表观特征具有相似性的情况下很难建立有效的判别模型. 本方法受到显著性检测文献的启发, 综合地考虑空间距离和颜色距离两方面的信息. 一方面, 和背景区域在空间距离上更远的像素, 属于目标的概率更高. 另一方面, 和背景区域在颜色上差异更大的像素, 属于目标的概率也更高. 因此, 本文提出一种新型的观测模型, 综合衡量目标邻域像素与已知背景区域在颜色和空间两个维度的距离, 来估计像素的目标概率, 在目标和背景颜色特征十分相似的情况下仍然能够给出非常鲁棒的估计结果. 图 3(a) 给出了跟踪图像的样例, 黑色虚线框内是目标概率待估计的搜索区域, 图 3(c) 显示了本章模型给出的像素级目标似然概率估计, 相比于颜色直方图模型[32] 得到的结果, 显著性模型给出的估计结果明显对背景噪声的干扰更加鲁棒.

背景先验理论[42] 假设大多数图像边界区域是背景, 以此为基础的图像边界先验方法已被应用于物体显著性检测任务中, 在实验中展示了可靠的结果. 本方法在目标跟踪场景中对背景先验进行应用, 根据第 $t-1$ 帧目标所在位置和尺度取扩大区域作为搜索区域, 假设该区域之外的图像为背景区域, 当前帧的观测量定义为每个像素点到背景区域的最



(a) 图像及搜索区域 (a) Example image and search region (b) 基于颜色的估计 (b) Color-based estimation (c) 基于显著性的估计 (c) Saliency-based estimation

图 3 基于颜色与基于显著性的目标似然概率估计结果对比

Fig. 3 Results of color-based and saliency-based target likelihood estimation

短距离.

理论上, 需要定义待求解区域 Ω 中每个像素点到背景的距离. 假定基于二维单通道图像区域来进行计算, 把区域中一条从某像素点 $\boldsymbol{x}$ 到背景种子点的路径记为 $\pi = \langle \pi(0), \cdots, \pi(k) \rangle$, 其中 $\pi(i)$ 和 $\pi(i+1)$ 是区域 Ω 中相邻的两个像素, 设 $\mathcal{F}(\pi)$ 为路径消耗, S 为背景种子像素集合, 则要计算的路径图 $D(\cdot)$ 定义为:

$$D(\boldsymbol{x}) = \min_{\pi \in \pi_{S,\boldsymbol{x}}} \mathcal{F}(\pi) \tag{15}$$

其中, $\Pi_{S,\boldsymbol{x}}$ 表示连接 S 和 $\boldsymbol{x}$ 之间的所有路径. 两个像素点 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 之间的路径长度有如下性质 $f(\boldsymbol{x}_1 \to \boldsymbol{x}_2) = f(\boldsymbol{x}_2 \to \boldsymbol{x}_1) \geq 0$. 路径消耗的定义取决于不同的应用. 一种经典的定义是使用测地距离[42], 可以对颜色和空间位置两方面的距离进行综合的衡量. 这种方法累加了路径上所有相邻像素点的灰度差作为路径消耗, 即

$$\mathcal{F}_{\mathrm{G}}(\pi) = \sum_{i=1}^{k} |I(\pi(i-1)) - I(\pi(i))| \tag{16}$$

其中, $I(\cdot)$ 表示像素点灰度值. 文献 [43] 提出使用最小障碍距离来进行显著性检测, 其路径消耗定义为:

$$\mathcal{F}_{\mathrm{MBD}}(\pi) = \max_{i=0}^{k} I(\pi(i)) - \min_{i=0}^{k} I(\pi(i)) \tag{17}$$

最小障碍距离检测相比于经典的测地距离, 可以得出对噪声和分辨率更加鲁棒的显著性检测结果. Zhang 等提出了快速最小障碍距离算法[44], 利用光栅扫描算法, 计算每个像素与邻近像素的距离, 累加其中的最大值来进行近似计算, 使得算法的实时性得到保证.

将文献 [44] 计算得到的最小障碍距离图归一化使得最大值为 1, 并进一步应用基于对数运算的 Sigmoid 函数进行对比度拉升操作:

$$p_{\mathrm{S}}\left(D^t(\boldsymbol{x}) | c_{\boldsymbol{x}}^t = 1\right) = \frac{1}{1 + \mathrm{e}^{-b(D^t(\boldsymbol{x}) - \beta_t)}} \tag{18}$$

$$p_{\mathrm{S}}\left(D^t(\boldsymbol{x}) | c_{\boldsymbol{x}}^t = 0\right) = 1 - \frac{1}{1 + \mathrm{e}^{-b(D^t(\boldsymbol{x}) - \beta_t)}} \tag{19}$$

其中, 模型参数 β_t 是根据已跟踪图像进行在线统计得到的阈值参数, 用于区分目标与背景, b 为固定的控制系数, 用来控制对比度拉升的程度.

3.4 基于运动估计的观测模型

自由物体的运动具有连续和平滑的属性, 属于同个物体的像素通常具有一致的运动趋势, 这一先验知识有助于视觉系统有效区分目标和背景的像素区域. 本节对目标和背景的运动参数进行解算, 基于这一关于运动的先验知识建立观测模型. 在图 4 中, 背景区域中存在与目标在颜色特征上十分相似的像素点, 传统的基于颜色的分类器很难有效地区分这些像素点, 然而由于它们具有和目标明显不同的运动趋势, 通过对运动的建模可以很容易被区分开来. 对该信息进行建模表达, 可以在物体进行复杂运动 (比如旋转+平移), 其他模型难以适应目标变化的情况下提供有效信息, 得到准确的目标定位结果.

对于每个类别 $C \in \{0, 1\}$ (其中 0 代表背景, 1 代表目标) 在每一帧估算其旋转角度 θ_C 和位移矢量 $[u_C, v_C]^{\mathrm{T}}$, 所有的运动状态参数记为 $\boldsymbol{s} = [\theta_0, u_0, v_0, \theta_1, u_1, v_1]$. 记 A_1 和 $\boldsymbol{b}_1$ 分别为目标的旋转矩阵和位移矢量, A_0 和 $\boldsymbol{b}_0$ 分别为背景的旋转矩阵和位移矢量, 已知运动状态参数 $\boldsymbol{s}$ 的情况下有:

$$A_C(\boldsymbol{s}) = \begin{bmatrix} \cos(\theta_C) & -\sin(\theta_C) \\ \sin(\theta_C) & \cos(\theta_C) \end{bmatrix}, \ \boldsymbol{b}_C(\boldsymbol{s}) = \begin{bmatrix} u_C \\ v_C \end{bmatrix} \tag{20}$$

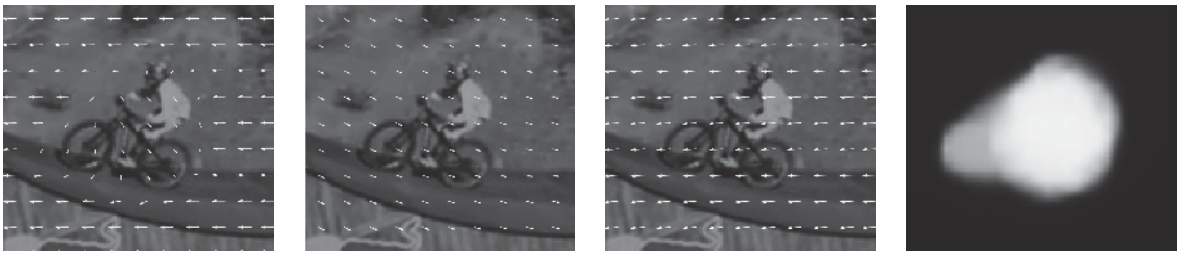
运动参数 $\boldsymbol{s}$ 计算的时候以本帧为参考帧, 即估计的运动参数定义了目标和背景从本帧到上一帧的运动. 在理想运动参数 $\boldsymbol{s}$ 已估计的情况下, $M_0(\boldsymbol{x}, \boldsymbol{s})$ 表示位置 $\boldsymbol{x}$ 属于背景区域假设下理想的后向光流矢量, $M_0(\boldsymbol{x}, \boldsymbol{s})$ 表示位置 $\boldsymbol{x}$ 属于目标区域假设下理想的后向光流矢量, 则有

$$M_C(\boldsymbol{x}, \boldsymbol{s}) = A_C(\boldsymbol{s})\boldsymbol{x} + \boldsymbol{b}_C(\boldsymbol{s}) - \boldsymbol{x} \tag{21}$$

在本模型中观测量为每个像素点的后向光流矢量 $M(\boldsymbol{x})$, 在像素点属于目标/背景的条件下, 该观测量分别是目标/背景模型理想运动矢量的一个带噪声的观测. 假定在这两种条件下, 光流矢量两个方向运动分量的观测误差均服从高斯分布且相互独立, 即

$$p(M(\boldsymbol{x}) | c_{\boldsymbol{x}} = C, \boldsymbol{s} = \hat{\boldsymbol{s}}) = \mathcal{N}(M(\boldsymbol{x}) | M_C(\boldsymbol{x}, \hat{\boldsymbol{s}}), \Sigma) \tag{22}$$

其中, $\Sigma = \mathrm{diag}\{\sigma_{u_C}, \sigma_{v_C}\}$ 是一个对角协方差矩阵, 其对角线上的元素定义了两个方向运动分量在高斯



(a) 观测到的运动场 (a) Motionfield observation estimated from image　(b) 理想的目标运动场 (b) Ideal motionfield of the target　(c) 理想的背景运动场 (c) Ideal motionfield of the background　(d) 基于运动模型的似然概率估计 (d) Motion-based likelihood estimation

图 4　基于目标与背景运动模型的似然概率估计示意图

Fig. 4　Demonstration of likelihood estimation based on motion models of target and background

观测模型中的方差. 图 4 (a) ~ (c) 分别给出了实际运动场 $M(\boldsymbol{x})$、估计的目标运动场 $M_1(\boldsymbol{x}, \hat{\boldsymbol{s}})$ 以及背景运动场 $M_0(\boldsymbol{x}, \hat{\boldsymbol{s}})$ 的可视化样例.

假设对于当前帧得到了 $\boldsymbol{s}$ 的最佳估计 $\hat{\boldsymbol{s}}$, 基于运动估计的像素级似然概率通过下式进行估计:

$$p_{\mathrm{M}}(M(\boldsymbol{x})|c_{\boldsymbol{x}} = C) \approx p(M(\boldsymbol{x})|c_{\boldsymbol{x}} = C, \boldsymbol{s} = \hat{\boldsymbol{s}}) \quad (23)$$

在式 (20) ~ (23) 中为了公式的简洁省略掉了上标 t.

为了对最佳的运动参数 $\hat{\boldsymbol{s}}$ 进行解算, 需要根据预测过程得到的概率分布 $p\left(c_{\boldsymbol{x}}^t = C|\boldsymbol{z}^{1:t-1}\right)$ 分别选取目标/背景的候选点, 使用相应的光流矢量来估计理想的目标/背景运动模型. 光流矢量描述了前后帧之间的像素匹配关系, 由于光流的估计不可避免地存在误差, 在进行运动模型解算的时候选用鲁棒估计算法来剔除误差较大的匹配点.

计算机视觉应用中常用的鲁棒估计算法包括随机抽样一致性 (Random sample consensus, RANSAC) 算法、M 估计抽样一致性 (M-estimator sample and consensus, MSAC) 算法和随机抽样最大似然估计 (Maximum likelihood estimation by sample and consensus, MLESAC) 算法等[53]. 这些算法根据随机表决的原理来计算模型参数, 基本思想是选择一个小的数据点集, 对其进行拟合, 查看有多少其他点可以匹配到拟合的物体上, 继续 l 次迭代直至达到某个终止条件时找到有较大概率的模型参数.

本文使用 MLESAC 算法, 从 l 次抽样中选择使得似然度最大的那次抽样的模型, 得到运动模型参数 $\hat{\boldsymbol{s}}$, 并进一步通过式 (23) 得到了基于运动估计的似然概率, 提供给像素级贝叶斯推断模型与多目标表征融合跟踪框架进行融合计算, 最终得出目标位置. 图 1 详细展示了基于运动估计的像素级似然概率, 基于显著性的像素级似然概率, 与基于矩形框目标表征的相关滤波器模型响应如何从图像观测中计算产生, 并且由一个统一的框架融合, 产生最终跟踪结果.

4　实验结果与分析

本节首先选取了几个具有代表性的跟踪序列, 和与本文方法相关性较大的最新算法进行了定性的比较, 以验证本文的方法动机与所取得的效果. 其次, 在广泛使用的目标跟踪标准数据集 (Object tracking benchmark) OTB-100[10] 上定量比较所提出的算法和目前主流目标跟踪算法的性能, 对总体性能以及不同挑战性因素影响下的性能分别进行分析, 以明确方法的优劣势与适用场景. 最后, 为进一步客观评估本文算法, 使用最近一期视觉物体跟踪挑战赛 (Visual object tracking challenge) 的数据集 VOT2018[11], 与所有参与测试的最新算法进行了比较, 对算法的短期跟踪性能进行评估.

本文使用 Matlab 实现了提出的目标跟踪算法, 部分模块使用 C++ 实现, 实验在配备有 3.1GHz i7 CPU 和 8GB RAM 的计算机上进行, 在实验中的平均帧率为 21 帧每秒 (Frames per second, FPS).

4.1　典型序列跟踪结果分析

本节定性比较提出的算法和与本文算法相关度较高的几个主流方法, 包括经典的相关滤波器算法 DSST[23] 和 SRDCF[24], 深度学习的代表性方法 CFNet[17], 以及使用深度学习技术建模注意力机制的 ACFN[18] 方法. 本节从 OTB-100 数据集中选择 8 个具有代表性的序列进行着重分析, 截取样例帧在图 5 中进行展示分析. 这些序列包含了目前最新跟踪算法致力于解决的各种视觉挑战, 包括尺度变化、背景复杂、目标形变、平面内旋转等.

图 5 中的 8 个序列从上到下总体上按照跟踪难度递增排列, 前 3 个为形状比较规则方正, 在跟踪过程中形变不大的序列, 前两个尺度变化较平缓,

第 3 个尺度变化较为明显. 第 4 到第 6 序列包含了明显的目标非刚体形变, 跟踪难度明显加大, 最后两个序列除了形变还包含持续的平面内旋转运动, 是目前主流跟踪方法普遍难以实现鲁棒跟踪的场景. 此外, 多个跟踪序列包含不同程度的复杂背景干扰. 以下针对每个挑战性因素进行分别讨论.

尺度变化通常由相机和目标之间的距离变化引起, 是目标跟踪中比较常见的一个视觉挑战, 图 5 中几乎所有的序列都包含了不同程度的尺度变化. 从前面 3 个较为简单的序列可以看出, 对于此类形状较为方正, 跟踪过程中形变较小的序列, 参与比较的几个主流算法几乎都能够较为准确地定位目标位置. 而在尺度估计方面, 本文算法和 DSST 算法具备一定的优势, 在几乎所有尺度变化的地方都能够及时更新目标框大小. 尤其在第 2 、第 3 个序列的后面阶段, 本文方法与 DSST 算法的尺度估计明显更为精确合理. 由于本文提出了多目标表征融合框架, 对基于矩形框和像素级表征的方法进行融合, 基于矩形框部分选用 DSST 作为基线方法, 因此继承了该算法在处理尺度变化上的优势.

背景复杂的情况在目标跟踪场景中也十分常见. 一种情况是场景里存在和被跟踪目标外观十分相似的其他疑似目标, 典型例子是图 5 第 5 行的 Soccer 序列; 另一种是背景本身颜色等特征和目标相似, 难以通过分类器有效辨识, 图 5 第 2、第 7 和第 8 行的 Singer2、Diving 和 MotorRolling 序列均在跟踪过程中的部分片段存在此类情况. 从图中可以看出, 在这些场景里面, 本文的算法受益于基于显著性先验和运动先验的观测模型, 对背景干扰具有很强的鲁棒性, 在其他多个主流算法跟踪失败的情况下仍然获得了十分准确的跟踪结果.

目标形变在非刚性物体的跟踪场景里普遍存在, 典型例子是图 5 第 4 ~ 7 行的 Bolt、Soccer、Panda 和 Diving 序列. Bolt 序列中的运动员在跟踪过程中存在快速运动, 同时伴随有目标形变, 在起跑之后 SRDCF 和 CFNet 很快丢失目标, ACFN 虽然准确定位了目标位置但是尺度估计偏差较大. Soccer 序列中除了目标形变, 还存在较为明显的背景干扰和运动模糊现象, DSST、ACFN 和 CFNet 跟踪器均在不同阶段出现较大的偏移. Panda 序列中 SRDCF、DSST 和 CFNet 先后丢失目标, 而本文算法在这几个序列中均保持鲁棒而准确的跟踪效果, 在此类跟踪场景中具有明显的优势.

目标平面内旋转是目前主流目标跟踪算法面临的难题之一, 基于矩形模板的目标表征模型很难对旋转运动导致的目标形状与外观变化进行自适应调整, 导致跟踪失败. 图 5 第 7 和第 8 行显示了两个典型的例子, 序列中的跳水运动员和摩托车均存在持续的平面内旋转和移动, 且部分帧存在复杂背景干扰问题, 参与比较的几个主流算法 (包括本文的基线方法 DSST 算法) 均在目标开始运动不久之后丢失目标, 而本文算法对两个序列均做到了对目标的全程跟踪, 且具有较高的跟踪精度, 说明所提出的像素级模型、集成的显著性与运动信息与现有模型呈现互补性, 能够十分有效提高此类场景下的算法跟踪精度和鲁棒性.

4.2 OTB-100 数据集实验

OTB-100 数据集[10] 总共包含有 100 个测试序列, 数据集上的所有图像序列都已经被人工标注, 标注的真值在图像上表现为包含有目标的矩形框. 数据集本身提供了包括 Struck、SCM、TLD 等在内的 29 个经典跟踪算法的跟踪结果, 后续提出的主流目标跟踪算法大多数在该数据集上进行了评测并提供了实验结果数据, 所以使用该数据集能够很方便地评估跟踪算法的性能.

为了使得评估的效果更加公平有效, 选用后续提出的性能更优越的主流算法, 以及与本文方法相关度较高的算法进行性能比较, 包括相关滤波模型的代表性算法 DSST[23]、Staple[25]、SRDCF[24], 判别式模型代表性方法 DLSSVM[15], 使用卷积神经网络特征构造显著图的 CNN-SVM[46] 算法, 结合深度学习与相关滤波模型的 CF2[47]、CFNet[17]、ACFN[18] 和 TRACA[48] 算法. 其中 CFNet 方法作者提供了使用深度卷积神经网络不同层特征的多个版本, 本文中用于比较的是使用 conv3 特征的版本, 记为 CFNet-conv3.

4.2.1 算法评价指标

OTB-100 数据集建议采用精度图和成功率图的方式对算法性能进行衡量和比较. 精度图和成功率图分别基于中心位置误差指标和重叠率指标进行统计获得, 这两个指标也是目前比较主流的衡量跟踪器性能的标准.

中心位置误差 (Center location error, CLE) 是被广泛使用的一个评价标准, 具体指跟踪所得的目标中心位置与基准中心位置之间的欧氏距离, 单位为像素, 即

$$S_{\mathrm{CLE}} = \sqrt{\left(x^t - x_{\mathrm{g}}^t\right)^2 + \left(y^t - y_{\mathrm{g}}^t\right)^2} \tag{24}$$

其中, (x^t, y^t) 表示第 t 帧时跟踪算法计算得到的目标中心位置坐标, $(x_{\mathrm{g}}^t, y_{\mathrm{g}}^t)$ 表示该时刻视频中目标的基准中心位置坐标. 可以看出, 中心位置误差仅仅衡量了像素位置的差异, 无法反映目标尺度大小上

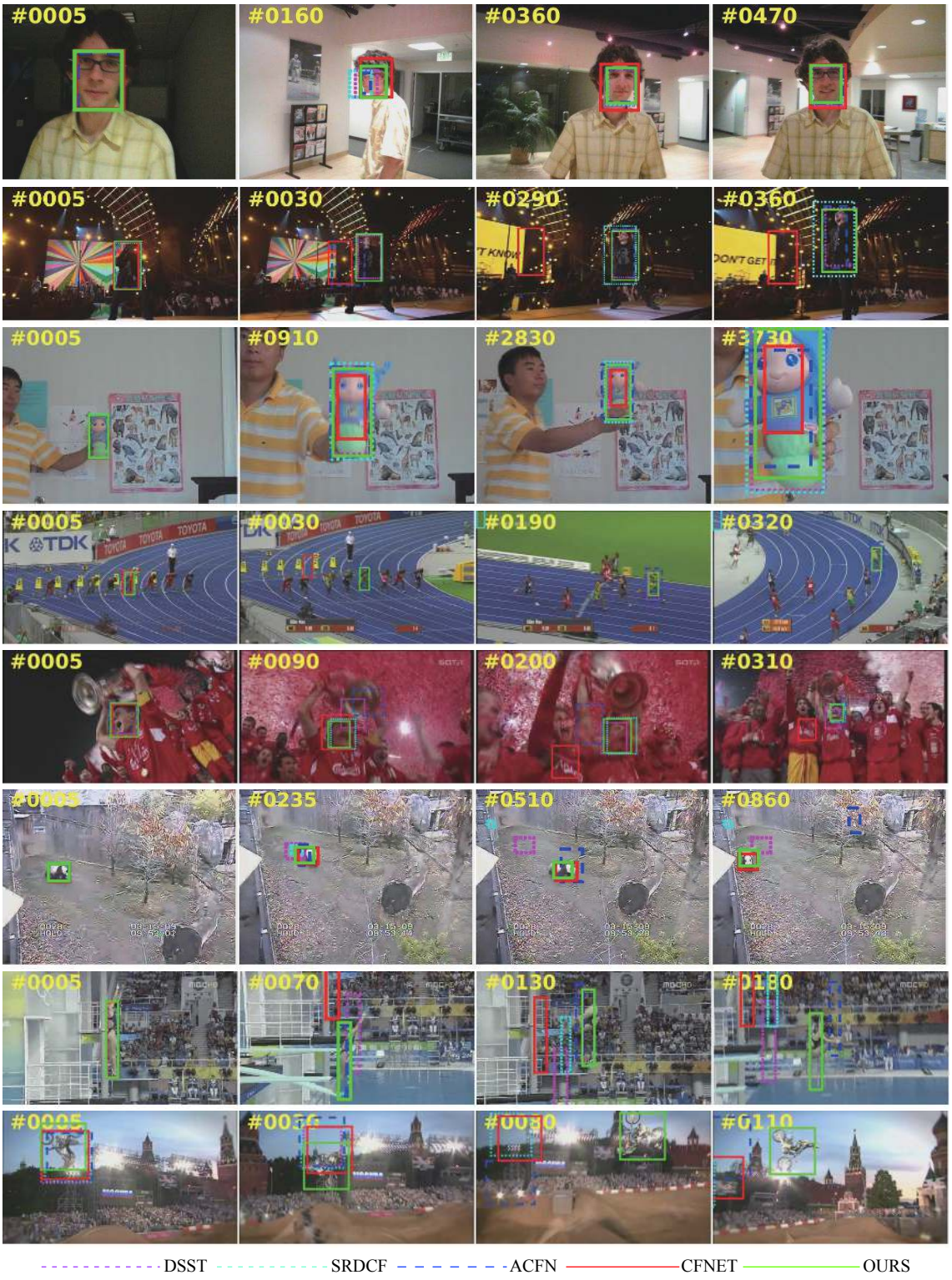


图 5　本文提出的跟踪算法和 DSST[23]、SRDCF[24]、ACFN[18]、CFNet[17] 在 8 个典型 OTB 序列上的跟踪结果 (从上往下分别是 David、Singer2、Doll、Bolt、Soccer、Panda、Diving 和 MotorRolling 序列)

Fig. 5　Tracking results using our proposed method compared with DSST, SRDCF, ACFN and CFNet on 8 OTB image sequences (From top to down: David, Singer2, Doll, Bolt, Soccer, Panda, Diving and MotorRolling

的误差. 通常定义 $S_{\mathrm{CLE}} \leq 20$ 为跟踪成功.　　精度图 (Precision plots) 对中心位置误差指标

进行统计, 横坐标为中心位置误差, 纵坐标为精度 (Precision), 表示中心位置误差小于某个阈值的视频帧数占总视频帧数的百分比. 通常使用 $S_{\mathrm{CLE}}=20$ 的精度指标衡量算法在数据集上的综合性能, 并对其进行排序.

重叠率 (Overlap) 是另一个常见的算法评价标准, 具体定义为

$$S_{\mathrm{OVL}}=\frac{A\left(R_{\mathrm{g}}^{t}\cap R^{t}\right)}{A\left(R_{\mathrm{g}}^{t}\cup R^{t}\right)} \tag{25}$$

其中, R^t 表示第 t 帧时算法输出的目标框, R_{g}^{t} 是数据集标注的标准目标框, $A(R_{\mathrm{g}}^{t}\cap R^{t})$ 表示两个区域重叠面积, $A(R_{\mathrm{g}}^{t}\cup R^{t})$ 表示两个区域并集的面积. 重叠率指标在衡量了跟踪算法输出结果与标准值在像素位置上差异的同时, 也反映了目标尺度大小估计上的准确程度, 在目标尺度估计不准确的情况下算法很难得到高的重叠率指标.

成功率图 (Success plots) 对算法在数据集各个序列上的重叠率指标进行统计, 横坐标为重叠率, 纵坐标为成功率 (Success rate), 具体表示重叠率大于某个阈值的帧数占视频总帧数的百分比. 通常使用成功率图的曲线下方面积 (Area under curve, AUC) 指标来对算法进行排序比较.

本实验采用一次通过估计 (One pass evaluation, OPE) 的方式给出跟踪算法的精度图与成功率图. OPE 评估方法从视频开头使用标注的基线矩形框进行初始化, 对算法在整个视频跟踪过程中的性能指标进行统计, 通过考察算法的长期跟踪能力评估其实用价值.

4.2.2 总体性能评估分析

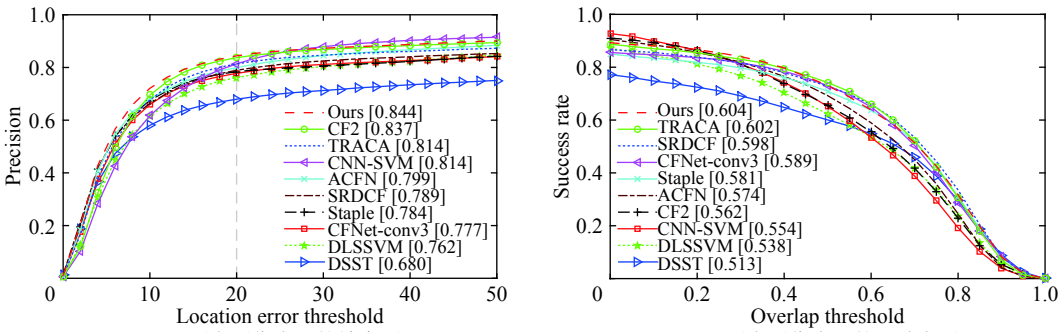
在 OTB-100 数据集上, 本文提出的目标跟踪算法与主流目标跟踪算法的性能比较结果如图 6 所示. 从图 6 中可以看出, 本文提出的算法在融合了运动与显著性信息之后, 相比于本文的基线方法, 经典的相关滤波器算法 DSST, 中心位置误差等于 20 距离精度 (即跟踪成功的帧数比例) 从 68.0 % 大幅提升到 84.4 %, 成功率图的 AUC 指标从 51.3 % 提升到 60.4 %, 充分证明了本文算法融合框架、以及所建模的运动信息与显著性信息的有效性.

图 6(a) 的精度图曲线充分说明, 本文提出的算法不仅相比于传统简单的算法模型有明显的优势, 在跟踪精度性能上也优于模型复杂和运行缓慢的 SRDCF, 以及多类整合利用了深度学习特征, 或者使用端对端深度卷积神经网络完成跟踪任务的最新主流方法. 图 6(b) 的成功率图曲线也说明了同样的结论.

4.2.3 不同挑战性因素影响下的性能分析

OTB-100 中的 100 个测试视频涵盖了现实生活中常见的包含各种复杂困难的跟踪场景, 具体使用 11 个挑战性因素属性进行了标注, 分别是光照发生变化 (Illumination variation, IV)、尺度变化 (Scale variation, SV)、目标发生遮挡 (Occlusion, OCC)、目标发生形变 (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)、跳出视野 (Out-of-view, OV)、背景复杂 (Background clutter, BC) 以及低分辨率 (Low resolution, LR).

对这些视频序列按照属性进行统计得到成功率图, 可以用于分析算法对于包含不同挑战性因素的跟踪场景的适用性与优缺点. 图 7(a) ~ (d) 显示了本文的算法最占优势的 4 个属性, 图 7(e) ~ (h) 则显示了最不占优势的 4 个属性. 从图 7 中可以看出, 本文算法对于背景复杂、尺度变化、目标发生形变这 3 个场景下的目标跟踪, 成功率图的 AUC 分



(a) 所有图像序列的精度图
(a) Precision plots of all sequences

(b) 所有图像序列的成功率图
(b) Success plots of all sequences

图 6 在 OTB-100 数据集上的一次通过估计曲线

Fig. 6 One-pass-evaluation (OPE) curves on OTB-100 dataset

别高于次优算法 1.3 %, 1.1 %, 0.6 %, 平面内旋转情况下略低于 TRACA 算法, 高于其他算法. 实验结果表明, 通过本文的算法模型整合的显著性信息能够有效减少复杂背景对于外观模型的干扰, 增强算法的鲁棒性, 所建立的运动模型也能够有效地捕捉目标与背景的运动, 带来有效信息.

另一方面, 在低分辨率、目标发生遮挡、快速运动、运动模糊 4 个困难场景下, 本文算法成功率图的 AUC 指标均低于 SRDCF, 且分别低于最优算法 9.1 %, 2.1 %, 1.4 % 和 1.2 %. 由于在上述 4 种困难场景下, 目前的光流估计算法比较难得到非常准确的光流估计, 这一实验结果表明, 本文算法模型可能会一定程度上受到光流估计算法结果的影响. 尽管如此, 本文算法在这 4 个场景下, 相对于没有引入显著性与运动信息的基线方法 DSST 算法, 在 AUC 指标上仍然分别取得了 8.5 %, 9.6 %, 12.4 % 和 11.0 % 的提升.

4.3 VOT2018 数据集实验

视觉物体跟踪挑战赛从 2013 年开始, 每年在计算机视觉顶级会议上举行. 为了将本文算法与目前最先进的方法进行客观的性能比较与分析, 我们使用最近一期的 VOT 挑战赛数据集 VOT2018[11], 与所有参与测试的最新算法进行比较. 该数据集包含 60 个测试视频, 选取的序列与 OTB 数据集相比具有更高的跟踪难度, 所有视频帧上的目标标准位置均经过人工标注.

VOT2018 竞赛提交的跟踪器总共有 72 个, 其中包括 36 个需要高性能 GPU 设备的算法, 以及 36 个仅需要 CPU 就可以运行的算法. 前者使用表征能力更强的深度神经网络特征, 在计算过程中依赖于使用高性能 GPU 来实现或接近实时效果, 总体跟踪精度和鲁棒性较高, 而后者由于不依赖于 GPU 计算, 可在各类不配备 GPU 或 GPU 算力不足的平台上使用, 其应用范围更加广泛. 本文致力于通过建模先验知识构造不依赖于 GPU 计算能力的实时跟踪器, 与后者属于同一类方法, 因此我们与后者进行了重点比较.

4.3.1 算法评价指标

VOT 数据集基于重叠率 (Overlap) 指标来评价算法跟踪性能, 计算方法与 OTB 相同, 详见式 (25), 具体测试条件和统计方式与 OTB 略有不同. OTB 侧重于评价算法的长期跟踪能力, 且在一次测试中只进行一次初始化, 不对跟踪失败的情况进行判断和重新初始化, VOT 则选用了一些跟踪难度较大的序列, 为充分利用数据集, 每次在算法跟踪失败 (重叠率小于 0) 5 帧之后对其进行重新初始化, 并且在重新初始化 10 帧之后继续统计重叠率. 具体使用精确度 (Accuracy) 和鲁棒性 (Robustness) 两个指标, 精确度计算了有效帧的平均重叠率, 鲁棒性则通过统计跟踪失败的次数来计算, 使用算法在这两个指标上的排序综合衡量了其跟踪性能.

为了在衡量算法的短期跟踪性能的同时, 减少重新初始化带来的统计偏差, VOT 挑战赛从 2015 年引入的另一个算法评价指标是平均重叠率期望 (Expected average overlap, EAO), 具体统计方法是在测试视频中截取长度较短的片段, 采用不重新初始化的方式进行一次性跟踪, 统计算法在片段上的平均重叠率, 通过对多个不同的长度下的平均重叠率求期望得到. 通过跟踪器 EAO 指标在目前最新算法中的排序, 可以比较客观地衡量其短期跟踪性能.

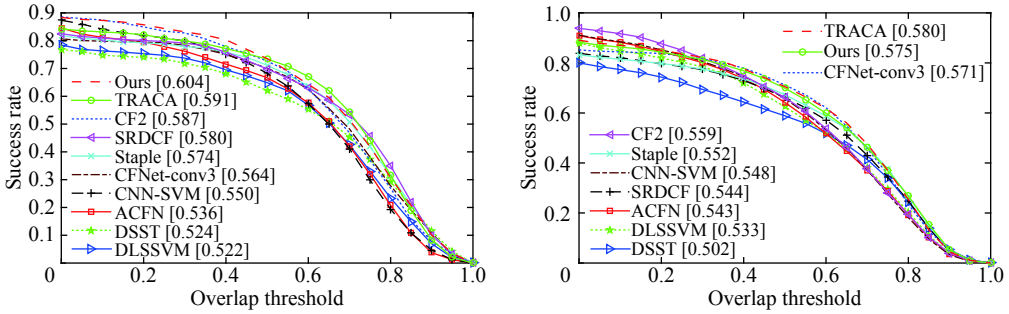
4.3.2 实验结果分析

在 VOT2018 数据集上, 本文提出的目标跟踪算法与主流目标跟踪算法的性能比较结果如图 8 所示. 图 8(a) 列出了所有参与测试的跟踪器, 图 8(b) 的精确度–鲁棒性图显示了本文算法同类方法 (不依赖于 GPU 设备的方法) 在精确度和鲁棒性两个维度上的排名, 越靠近图的右上方表明算法总体跟踪性能越好. 与目前最先进的方法 (包括已正式发表文献的最新方法, 以及部分尚未发表文献的方法) 相比, 本文的方法在精确度上排名第 4, 在鲁棒性上排名第 7, 取得较高的综合性能.

为了更进一步客观评估本文算法的性能, 呈现与目前最先进跟踪器的比较结果, 我们在图 8(c) 的平均重叠率期望排序图中列出了所有参与测试的跟踪算法, 包括 36 个依赖于 GPU 计算能力的跟踪算法和 36 个不需要 GPU 设备的算法. 不需要 GPU 设备的跟踪器在图的下方使用灰色圆圈进行了标记, 从图示中可以看出排名前 30 的跟踪器中仅有 6 个即 1/5 不需要 GPU 设备, 说明这一方面的研究有待进一步的加强, 以满足此类情况下视觉目标跟踪技术的应用需求.

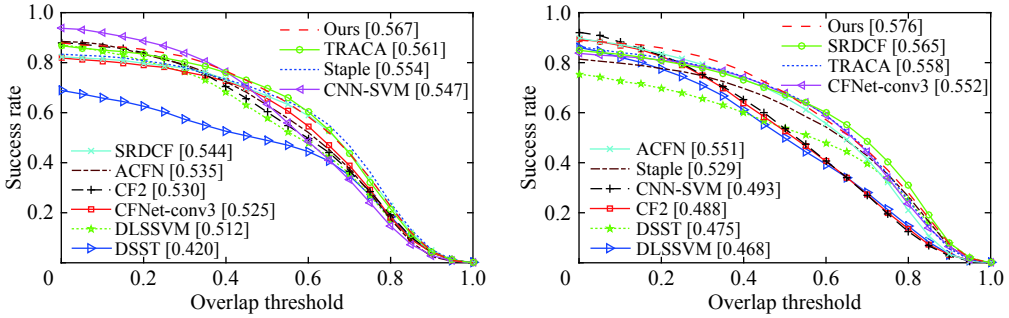
图 8(c) 的平均重叠率期望排序图显示, 我们的方法在所有参与测试的算法中排名第 31. 在所有不依赖于 GPU 设备的同类跟踪算法中, 本文方法跟踪性能优于大部分 (36 个中的 30 个) 同类方法. 值得一提的是, 跟踪性能优于我们的 6 个同类方法中包括已正式发表文献的方法, 也包括尚未公开发表文献的方法, 其中排名第 4、第 12、第 20 和第 27 的 UPDT[54]、SRCT[55]、MCCT[56]、CSRDCF[57] 为已公开发表的方法, 在 VOT2018 竞赛工具包中报告

的帧率分别为 0.43, 1.12, 1.29 和 8.75 fps, 排名第 1 和第 28 的 LADCF 和 DCFCF 方法未标注公开发表的文献, 报告的帧率分别为 0.52 和 0.18 fps, 均明显低于本文方法的 21.33 fps. 相比之下, 本文方法在精确度、鲁棒性和速度方面具有较大的综合优势, 具有较高的实用价值.



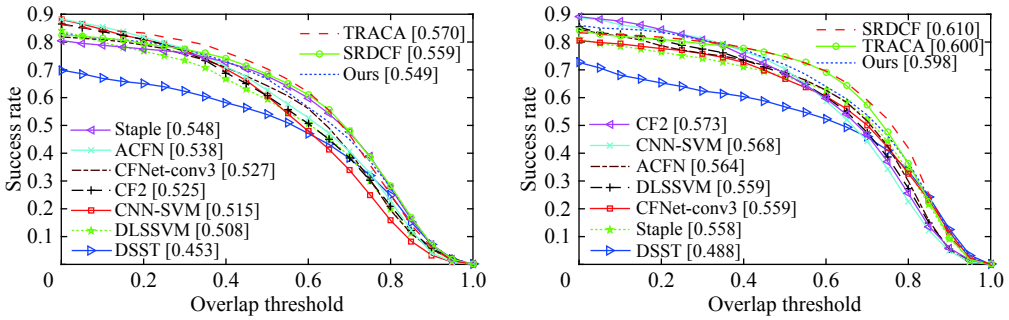
(a) 背景复杂场景下的成功率图
(a) Success plots - background clutter

(b) 平面内旋转场景下的的成功率图
(b) Success plots - in - plane rotation



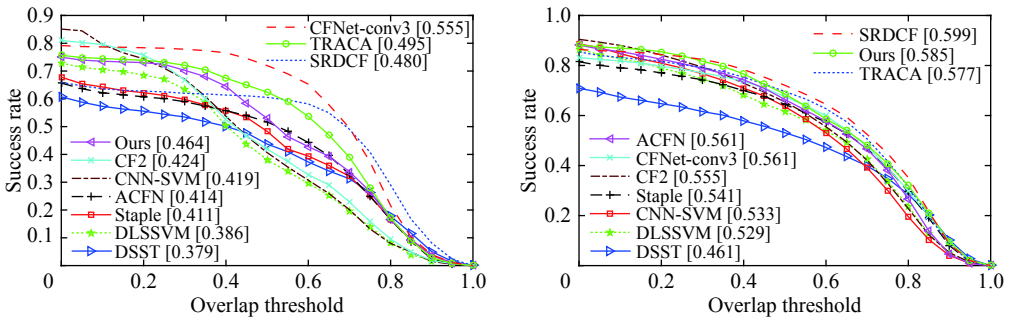
(c) 目标发生形变场景下的的成功率图
(c) Success plots - deformation

(d) 尺度变化场景下的的成功率图
(d) Success plots - scale variation



(e) 目标发生遮挡场景下的的成功率图
(e) Success plots - occlusion

(f) 运动模糊场景下的的成功率图
(f) Success plots - motion blur

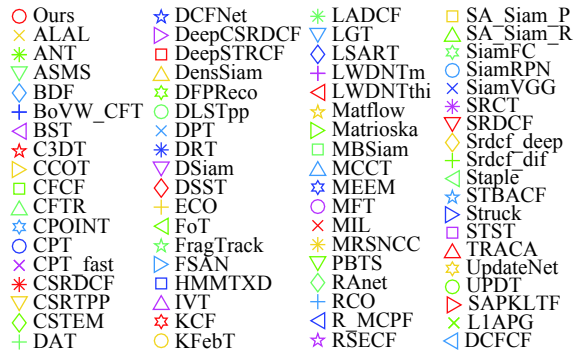


(g) 低分辨率场景下的的成功率图
(g) Success plots - low resolution

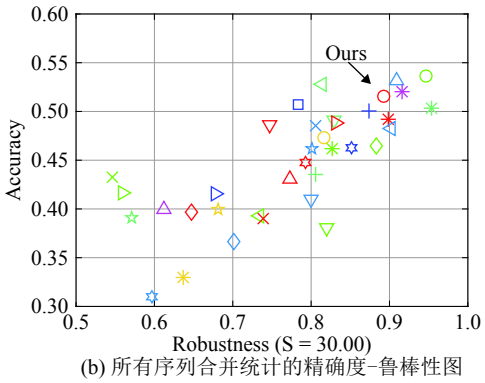
(h) 快速运动场景下的的成功率图
(h) Success plots - fast motion

图 7 在 OTB-100 数据集不同挑战性因素影响下的成功率图

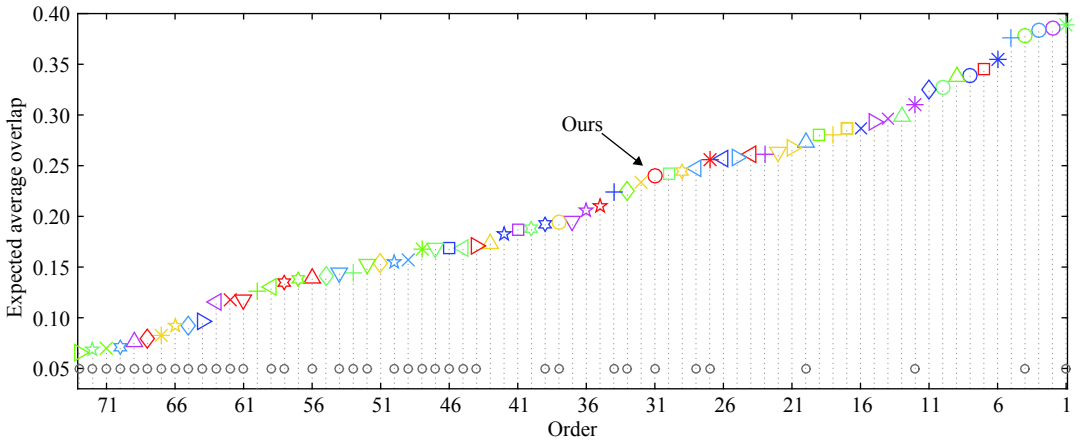
Fig. 7 Success plots on sequences with different challenging attributes on OTB-100 dataset



(a) 参与测试的跟踪器标识
(a) Labels of trackers tested



(b) 所有序列合并统计的精确度-鲁棒性图
(b) The AR plots generated by sequence pooling



(c) 平均重叠率期望图示(所有跟踪器根据性能排名从右到左排列,仅依赖于 CPU 的跟踪器在图下方使用灰色圆圈标记)
(c) Expected average overlap graph with trackers ranked from right to left (CPU trackers are denoted by gray circles)

图 8　在 VOT2018 数据集上的实验结果

Fig. 8　Experimental results on VOT2018 dataset

5　结论

本文在视觉目标跟踪的应用中, 提出了一种像素级概率性目标表征模型, 用于集成与主流的矩形框模板表征模型互补的观测信息, 并且对多目标表征模型提供的信息进行融合决策. 具体建立了感兴趣区域像素目标概率的贝叶斯推断模型, 每一帧通过上一帧的估计结果和状态传递概率预测本帧像素点的目标概率, 再融合本帧的像素级图像观测进行修正. 像素级图像观测部分建模和集成了被主流目标跟踪算法所忽略、而在人类视觉系统中十分重要的显著性信息与运动信息. 其中, 基于显著性的观测模型具体使用背景先验和最小障碍距离算法进行建模, 能够在背景干扰的情况下提供具备高辨识度的图像证据; 基于运动信息的观测模型则利用了相机与目标运动的连续性, 通过计算目标和背景的运动模式, 建立像素级的图像证据, 能够为目标复杂运动的场景提供有效决策信息. 实验结果表明, 提出的模型能够有效地融合像素级的显著性与运动信息, 增强跟踪算法在背景干扰、目标形变严重、复杂运动等挑战性跟踪场景下的鲁棒性, 与同类跟踪算法相比, 在跟踪精度、鲁棒性和运行速度方面具有较大的综合优势, 具有较高的实用价值.

References

1　Meng Lu, Yang Xu. A survey of object tracking algorithms. *Acta Automatica Sinica*, 2019, **45**(7): 1244−1260
(孟琭, 杨旭. 目标跟踪算法综述. 自动化学报, 2019, **45**(7): 1244−1260)

2　Li P X, Wang D, Wang L J, Lu H C. Deep visual tracking: Review and experimental comparison. *Pattern Recognition*, 2018, **76**: 323−338

3　Mabrouk A B, Zagrouba E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 2018, **91**: 480−491

4　Li P L, Qin T, Shen S J. Stereo vision-based semantic 3D object and ego-motion tracking for autonomous driving. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 664−679

5　Singha J, Roy A, Laskar R H. Dynamic hand gesture recognition using vision-based approach for human — Computer inter-

action. *Neural Computing and Applications*, 2018, **29**(4): 1129−1141

6 Singh P, Agrawal P, Karki H, Shukla A, Verma N K, Behera L. Vision-based guidance and switching-based sliding mode controller for a mobile robot in the cyber physical framework. *IEEE Transactions on Industrial Informatics*, 2019, **15**(4): 1985−1997

7 Gupta V, Lantz J, Henriksson L, Engvall J, Karlsson M, Persson A, et al. Automated three-dimensional tracking of the left ventricular myocardium in time-resolved and dose-modulated cardiac CT images using deformable image registration. *Journal of Cardiovascular Computed Tomography*, 2018, **12**(2): 139−148

8 Zhang Zhi-Jun, Zhong Sheng, Wu Ying, Wang Jian-Hui. Collaborative reranking: A novel approach for hand pose estimation. *Journal of Computer-Aided Design & Computer Graphics*, 2018, **30**(11): 2182−2192
(张芷君, 钟胜, 吴郢, 王建辉. 基于协同重排序的手势识别方法. 计算机辅助设计与图形学学报, 2018, **30**(11): 2182−2192)

9 Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 472−487

10 Wu Y, Lim J, Yang M H. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9): 1834−1848

11 Kristan M, Leonardis A, Matas J, Felsberg M, Pflugfelder R, Zajc L C, et al. The sixth visual object tracking vot2018 challenge results. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 3−53

12 Ross D A, Lim J, Lin R S, Yang M H. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 2008, **77**(1): 125−141

13 Zhang T Z, Ghanem B, Liu S, Ahuja N. Low-rank sparse learning for robust visual tracking. In: Proceedings of the 2012 European Conference on Computer Vision. Florence, Italy: Springer, 2012. 470−484

14 Henriques J F, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(3): 583−596

15 Ning J F, Yang J M, Jiang S J, Zhang L, Yang M H. Object tracking via dual linear structured SVM and explicit feature map. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4266−4274

16 Bertinetto L, Valmadre J, Henriques J F, Vedaldi A, Torr P H S. Fully-convolutional siamese networks for object tracking. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 850−865

17 Valmadre J, Bertinetto L, Henriques J, Vedaldi A, Torr P H S. End-to-end representation learning for correlation filter based tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5000−5008

18 Choi J, Chang H J, Yun S, Fischer T, Demiris Y, Choi J Y. Attentional correlation filter network for adaptive visual tracking. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 4828−4837

19 Itti L, Koch C. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2001, **2**(3): 194−203

20 Yu Ming, Li Bo-Zhao, Yu Yang, Liu Yi. Image saliency detection with multi-graph model and manifold ranking. *Acta Automatica Sinica*, 2019, **45**(3): 577−592
(于明, 李博昭, 于洋, 刘依. 基于多图流形排序的图像显著性检测. 自动化学报, 2019, **45**(3): 577−592)

21 Spelke E S, Katz G, Purcell S E, Ehrlich S M, Breinlinger K. Early knowledge of object motion: Continuity and inertia. *Cognition*, 1994, **51**(2): 131−176

22 Johnson S P, Aslin R N. Perception of object unity in young infants: The roles of motion, depth, and orientation. *Cognitive Development*, 1996, **11**(2): 161−180

23 Danelljan M, Hager G, Khan F S, Felsberg M. Accurate scale estimation for robust visual tracking. In: Proceedings of the 2014 British Machine Vision Conference. Nottingham, UK: BMVA Press, 2014. 1−11

24 Danelljan M, Hager G, Khan F S, Felsberg M. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4310−4318

25 Bertinetto L, Valmadre J, Golodetz S, Miksik O, Torr P H S. Staple: Complementary learners for real-time tracking. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1401−1409

26 Kwon J, Lee K M. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive Basin Hopping Monte Carlo sampling. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 1208−1215

27 Liu Da-Qian, Liu Wan-Jun, Fei Bo-Wen, Qu Hai-Cheng. A new method of anti-interference matching under foreground constraint for target tracking. *Acta Automatica Sinica*, 2018, **44**(6): 1138−1152
(刘大千, 刘万军, 费博雯, 曲海成. 前景约束下的抗干扰匹配目标跟踪方法. 自动化学报, 2018, **44**(6): 1138−1152)

28 Fan J L, Shen X H, Wu Y. Scribble tracker: A matting-based approach for robust tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(8): 1633−1644

29 Godec M, Roth P M, Bischof H. Hough-based tracking of non-rigid objects. *Computer Vision and Image Understanding*, 2013, **117**(10): 1245−1256

30 Bibby C, Reid I. Robust real-time visual tracking using pixel-wise posteriors. In: Proceedings of the 2008 European Conference on Computer Vision. Marseille, France: Springer, 2008. 831−844

31 Oron S, Bar-Hille A, Avidan S. Extended lucas-kanade tracking. In: Proceedings of the 2014 European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 142−156

32 Possegger H, Mauthner T, Bischof H. In defense of color-based model-free tracking. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 2113−2120

33 Son J, Jung I, Park K, Han B. Tracking-by-segmentation with online gradient boosting decision tree. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3056–3064

34 Duffner S, Garcia C. Fast pixelwise adaptive visual tracking of non-rigid objects. *IEEE Transactions on Image Processing*, 2017, **26**(5): 2368–2380

35 Aeschliman C, Park J, Kak A C. A probabilistic framework for joint segmentation and tracking. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1371–1378

36 Papoutsakis K E, Argyros A A. Integrating tracking with fine object segmentation. *Image and Vision Computing*, 2013, **31**(10): 771–785

37 Vo B T, Vo B N. Labeled random finite sets and multi-object conjugate priors. *IEEE Transactions on Signal Processing*, 2013, **61**(13): 3460–3475

38 Wong S C, Stamatescu V, Gatt A, Kearney D, Lee I, McDonnell M D. Track everything: Limiting prior knowledge in online multi-object recognition. *IEEE Transactions on Image Processing*, 2017, **26**(10): 4669–4683

39 Punchihewa Y G, Vo B T, Vo B N, Kim D Y. Multiple object tracking in unknown backgrounds with labeled random finite sets. *IEEE Transactions on Signal Processing*, 2018, **66**(11): 3040–3055

40 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**(11): 1254–1259

41 Hou X D, Zhang L Q. Saliency detection: A spectral residual approach. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, Minnesota, USA: IEEE, 2007. 1–8

42 Wei Y C, Wen F, Zhu W J, Sun J. Geodesic saliency using background priors. In: Proceedings of the 2012 European Conference on Computer Vision. Florence, Italy: Springer, 2012. 29–42

43 Ciesielski K C, Strand R, Malmberg F, Saha P K. Efficient algorithm for finding the exact minimum barrier distance. *Computer Vision and Image Understanding*, 2014, **123**: 53–64

44 Zhang J M, Sclaroff S, Lin Z, Shen X H, Price B, Mech R. Minimum barrier salient object detection at 80 FPS. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1404–1412

45 Zhao D W, Xiao L, Fu H, Wu T, Xu X, Dai B. Augmenting cascaded correlation filters with spatial — temporal saliency for visual tracking. *Information Sciences*, 2019, **470**: 78–93

46 Hong S, You T, Kwak S, Han B. Online tracking by learning discriminative saliency map with convolutional neural network. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015. 597–606

47 Ma C, Huang J B, Yang X K, Yang M H. Hierarchical convolutional features for visual tracking. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3074–3082

48 Choi J, Chang H J, Fischer T, Yun S, Lee K, Jeong J, et al. Context-aware deep feature compression for high-speed visual tracking. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 479–488

49 Gladh S, Danelljan M, Khan F S, Felsberg M. Deep motion features for visual tracking. In: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico: IEEE, 2016. 1243–1248

50 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE, 2001. I-511–I-518

51 Danelljan M, Hager G, Khan F S, Felsberg M. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(8): 1561–1575

52 Kroeger T, Timofte R, Dai D, van Gool L. Fast optical flow using dense inverse search. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 471–488

53 Torr P H S, Zisserman A. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 2000, **78**(1): 138–156

54 Bhat G, Johnander J, Danelljan M, Khan F A, Felsberg M. Unveiling the power of deep tracking. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 493–509

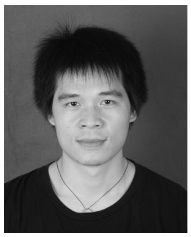
55 Lee H, Kim D. Salient region-based online object tracking. In: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, NV, USA: IEEE, 2018. 1170–1177

56 Wang N, Zhou W G, Tian Q, Hong R C, Wang M, Li H Q. Multi-cue correlation filters for robust visual tracking. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 4844–4853

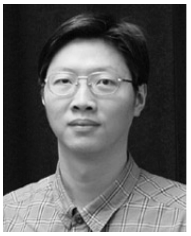
57 Lukezic A, Vojir T, Zajc L C, Matas J, Kristan M. Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 4847–4856

张伟俊 华中科技大学人工智能与自动化学院博士研究生. 2012 年获得华中科技大学电子信息工程系学士学位. 主要研究方向为计算机视觉,模式识别.本文通信作者.
E-mail: starfire.zhang@gmail.com
(**ZHANG Wei-Jun** Ph. D. candidate at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST). He received his bachelor degree from Huazhong University of Science and Technology in 2012. His research interest covers computer vision and pattern recognition. Corresponding author of this paper.)

钟 胜 华中科技大学人工智能与自动化学院教授. 2005 年获得华中科技大学模式识别与智能系统博士学位. 主要研究方向为模式识别, 图像处理, 实时嵌入式系统.
E-mail: zhongsheng@hust.edu.cn
(**ZHONG Sheng** Professor at the School of Automation, Huazhong University of Science and Technology (HUST). He received the Ph. D. degree in pattern recognition and intelligent systems in 2005 from Huazhong University of Science and Technology. His research interest covers pattern recognition, image processing, and real-time embedded systems.)



徐文辉 华中科技大学人工智能与自动化学院博士研究生. 2006 年获得吉林大学大学学士学位. 主要研究方向为计算机视觉, 算法加速.
E-mail: xuwenhui@hust.edu.cn
(**XU Wen-Hui** Ph. D. candidate at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. He received his bachelor degree from Jilin University in 2006. His research interest covers computer vision and accelerated computing.)



WU Ying 美国西北大学电子工程与计算机系终身正教授. 2005 年获得美国伊利诺伊大学厄巴纳–香槟分校电子与计算工程博士学位. 主要研究方向为计算视觉与图形学, 图像与视频处理, 多媒体, 机器学习, 人体运动, 人机智能交互, 虚拟现实.
E-mail: yingwu@ece.northwestern.edu
(**WU Ying** Professor (tenured) in the Department of Electrical Engineering and Computer Science, Northwestern University, USA. He received his Ph. D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC) in 2001. His research interest covers computer vision/graphics, image/video processing, multi-media, machine learning, human motion, human-computer intelligent interaction, and virtual environments.)