马丁, 邬向前. 基于自然语言查询的视觉目标跟踪方法综述[J]. 智能计算机与应用,2024,14(12):195-199. DOI:10.20169/j.issn.2095-2163.20231128

基于自然语言查询的视觉目标跟踪方法综述

马 丁, 邬向前
(哈尔滨工业大学 计算机科学与技术学院, 哈尔滨 150001)

摘 要: 基于自然语言查询的视觉目标跟踪方法是一个新兴的研究热点,旨在利用自然语言查询来锁定目标在视频帧中的位置。与需要手工标注矩形框的视觉目标跟踪方法不同,基于自然语言查询的视觉目标跟踪方法通过高级语义信息来指导跟踪器,旨在消除包含歧义性的手工标注矩形框,并将本地搜索与全局搜索有机地结合起来。因此,基于自然语言查询的视觉目标跟踪方法能够在实际场景中带来更灵活、稳健和准确的跟踪性能。综上所述,本文对基于自然语言查询的视觉目标跟踪方法进行综述,概述相关原理和模型改进的关键技术,总结不同网络结构的优缺点。

关键词: 自然语言查询; 视觉目标跟踪; 异质特征融合
中图分类号: TP399 **文献标志码**: A **文章编号**: 2095-2163(2024)12-0195-05

A review of visual object tracking by natural language specification

MA Ding, WU Xiangqian
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Visual object tracking by natural language specification is an emerging research hotspot, which aims to use natural language specification to locate the position of the target in video frames. Unlike visual object tracking methods that require manual annotation of rectangular boxes, visual object tracking method based on natural language query guides the tracker through advanced semantic information, aiming to eliminate ambiguous manual annotation of rectangular boxes and combine local search with global search. Therefore, tracking by natural language specification can bring more flexible, robust and accurate tracking performance in practical scenarios. In summary, this article reviews visual object tracking by natural language specification, outlines the key technologies of related principles and model improvements, and summarizes the advantages and disadvantages of different network structures.

Key words: natural language specification; visual object tracking; heterogeneous feature fusion

0 引 言

单目标跟踪是计算机视觉领域最重要的任务之一,该任务已广泛应用于诸多领域,如视频监控、机器人视觉感知和自动驾驶等。对于单目标视觉跟踪而言,给定第一帧的图像,需要根据手工标定的矩形框来初始化跟踪器。并且,目前大多数的单目标跟踪方法[1-4]都沿用了这种设置。虽然,以上这些单目标跟踪方法取得了不错的性能,但是,使用手工标定的矩形框初始化跟踪器仍存在以下问题:

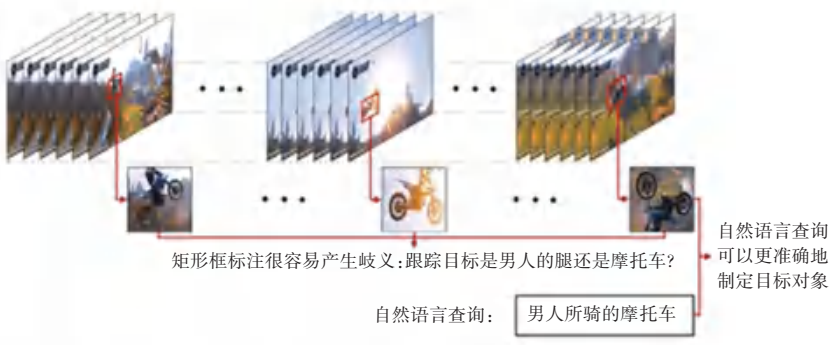
(1)初始化需要使用手工标定的矩形框限制了这些方法在实际场景中的应用。

(2)自然语言查询在目标跟踪任务中的优势举例如图 1 所示。由图 1(a)可知,使用手工标定的矩形框有时并不能准确勾勒目标的有效区域,甚至有可能导致歧义。此外,由图 1 (b) 可知,由于第一帧初始化的外观特征和跟踪过程中的对象有很大的不同,当前基于标定框的跟踪器在面对目标对象的突然外观变化时可能表现不佳。

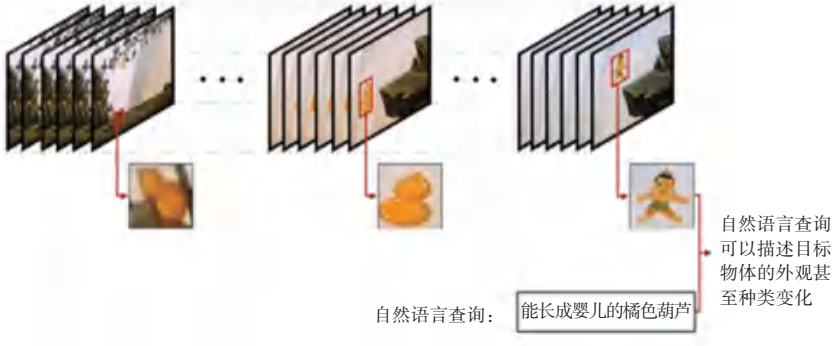
综上可知,最近一些研究正尝试引入自然语言查询以替代手工标注的矩形框来初始化跟踪器,这种新颖的人机交互形式,使得基于自然语言查询的跟踪更加贴合实际场景的应用需求。具体而言,引入自然语言查询有助于增强现有基于手工标注初始

基金项目: 国家自然科学基金青年科学基金(20230197)。
作者简介: 马 丁(1988—),男,博士,助理研究员,主要研究方向:基于自然语言查询的目标跟踪,视觉目标跟踪。
通信作者: 邬向前(1973—),男,博士,教授,主要研究方向:数字图像处理,医学图像分析。Email: xqwu@hit.edu.cn。
收稿日期: 2023-11-28

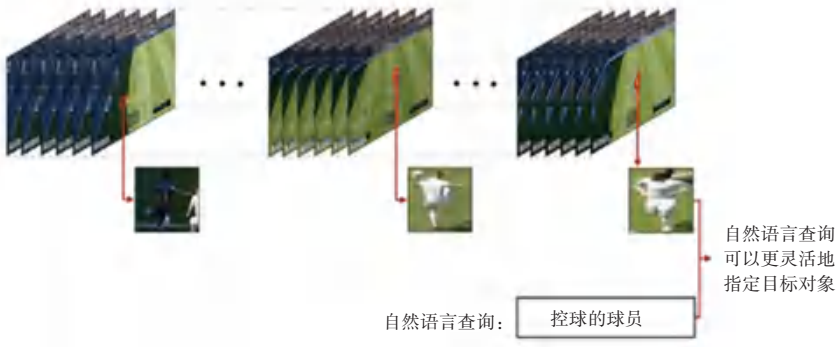
化的跟踪器的性能,提高跟踪器对于模型的任务鲁棒性,同时还可以拓展到多目标跟踪。此外,相比于手工标注的矩形框,自然语言查询表达起来更加直观。同时,自然语言查询可以从高级语义信息又可以从空间位置来更加精确地描述目标对象,如目标的属性、类别、外观形状以及与场景中其它目标的相对位置关系。以上这些信息在目标经历剧烈外观变化时,对于跟踪器准确定位目标尤为重要。同时,自然语言查询为灵活地切换目标提供了便捷,例如在图 1(c)中需要跟踪“控球的人”,理想的跟踪器应能够灵活地切换不同的控球人,现有的基于手工标注矩形框的跟踪器显然是无法做到的,而引入自然语言查询恰恰能够应对此类场景。正是发现了以上自然语言查询的技术优势,研究者们设计了一系列的基于自然语言查询的视觉目标跟踪方法。因此,本文针对基于自然语言查询的视觉目标跟踪方法进行综述,概述相关原理和模型改进的关键技术,总结不同网络结构的优缺点。



(a) 自然语言查询男人所骑的摩托车



(b) 自然语言查询能长成婴儿的橘色葫芦



(c) 自然语言查询控球的球员

图 1 自然语言查询在目标跟踪任务中的优势举例

Fig. 1 Examples of advantages of natural language specification in tracking tasks

1 基于自然语言查询的视觉目标跟踪方法

基于自然语言查询的视觉目标跟踪方法按不同的主干网络可分为基于卷积神经网络、基于胶囊网络、以及基于 Transformer 结构 3 大类。下面,本文对每一类中的方法进行了概述,对尚存在的不足等方面进行总结。

1.1 基于卷积神经网络的方法

作为引入自然语言查询的首个工作,TNL[5] 提出了一个动态卷积层用于目标在视频帧中的定位,

旨在通过 LSTM 最后一个隐藏层获得自然语言查询的编码，并通过该编码获得目标的初始定位。TNL 结构示意如图 2 所示。具体而言，该方法通过 VGG-16 预训练模型的第 6、第 7、和第 8 个卷积层获得视频帧表示。所有的 LSTM 隐藏层单元包含 1 000 维的状态向量。

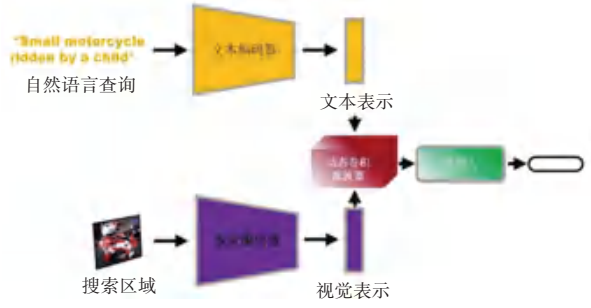


图 2　TNL 结构示意图

Fig. 2　Framework of TNL

通过给定的自然语言查询，该方法生成 1×1 大小的动态卷积滤波器，该滤波器对视频帧表示进行逐像素的滤波，从而获得目标位置的高响应区域。作为这一研究方向的开山之作，虽然已证明了引入自然语言查询的有效性，但生成的动态卷积滤波器并不稳定，因为查询中的语言变异（Linguistic Variation）会严重影响语言表示。例如，“runner in the middle with white shirt”和“runner with white shirt in the middle”，以上 2 个意思相同的自然语言表达可能会产生不同的动态卷积滤波器，从而导致次优的跟踪结果。

NLTNL[6] 提出了由 2 个阶段构成的跟踪器。第一个阶段负责检测，其中包含了深度卷积神经网络和 RNN 两个模块。第二个阶段是跟踪，同样使用了 RNN。整个流程首先通过深度卷积神经网络提出候选检测区域，然后使用 RNN 计算与输入语言的相似性，最后根据相似性进行排序，并更新跟踪器。与 TNL 类似，在度量视觉与自然语言相似度时，该方法仍然使用了不稳定的动态卷积滤波器。

在 SNLT[6] 中，对于每个视频，模型要接收 3 个输入，分别是：一个视觉样本、一个视觉搜索区域和一个语言查询。为了提取视觉样本和视觉搜索区域的视觉特征，该方法使用卷积神经网络（CNN），比如 AlexNet 和 ResNet-50。同时，利用预训练语言模型来计算自然语言描述的句子嵌入。对于预训练语言模型，SNLT 采用了基于 GloVe、HGLMM 和 BERT 的模型。接着，将以上三者传递给提出的孪生自然语言区域建议网络（SNL-RPN），该网络将预测边界框分类得分，并在一组预定义的锚点上对视觉和语言模态进行回归。最后，在获得 SNL-RPN 的预测后，动态聚合模块将视觉和语言模态的预测结合起来。

1.2　基于胶囊网络的方法

上述方法中的异质特征融合策略并不能保证视觉分支能够感知到语言查询所关注的上下文区域，而上下文区域的感知对于跟踪任务至关重要。CapsuleTNL[7] 提出了一种基于胶囊网络并结合自然语言查询的回归跟踪方法。CapsuleTNL 结构示意如图 3 所示。

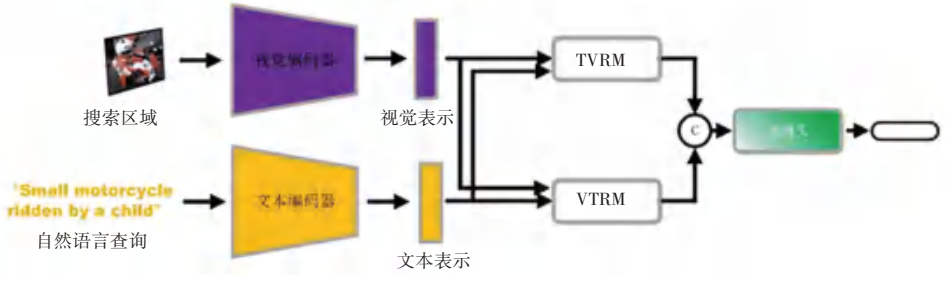


图 3　CapsuleTNL 结构示意图

Fig. 3　Framework of CapsuleTNL

首先，该方法通过文本和视觉编码器分别获得词级查询和搜索区域的特征表示。其次，为了能够促进视频帧和自然语言的双向交互，CapsuleTNL 设计了视觉-文本路由模块（VTRM）和文本-视觉路由模块（TVRM）来同时关注自然语言查询到视频帧和视频帧到自然语言查询的特征嵌入。其中，VTRM 用来减少语言变异的干扰，并提高视觉和文本表示的相关性。同时，为了引导视觉表示专注于关键的上下文信息，TVRM 致力于让视觉表示捕获自然语言刻画的上下文信息，来提高跟踪性能。以上两者都

通过路由协议来捕捉实体关系,从而对齐 2 种异质的特征嵌入。输入视频帧和自然语言查询对,CapsuleTNL 能够进行端到端的训练,并将相似的视觉和文本表示聚合以形成二者的联合特征空间。

1.3 基于 Transformer 的方法

现有的方法通常将跟踪的流程分解为两步,即视觉定位和跟踪,并相应地提出视觉定位模型和跟踪模型,相对独立地实现这 2 个步骤。这种相对独立的结构忽略了视觉定位和跟踪之间的关联关系,也就是说,自然语言查询在这 2 个步骤中均提供了全局语义信息。针对该问题,JointNLT[8] 提出了一个联合视觉定位和跟踪框架,该框架将定位和跟踪重新定义为统一的任务:基于给定的视觉语言查询定位所关注的目标。具体而言,JointNLT[8] 提出了一个多源关系建模模块来有效地建立视觉语言查询和视频帧之间的关系。此外,利用时序建模模块,在全局语义信息的引导下为模型提供时序线索,有效提高了模型对目标外观变化的适应性。

基于自然语言查询的目标跟踪方法的挑战主要来自于通过融合 2 种异构信息来预测目标的位移变化。对于这 2 种异构信息,一种是文本查询中包含的视频主要特征的静态描述、即长时上下文,另一种是从当前视频帧裁剪获得的包含目标与其附近环境的图像块、即搜索区域。目前,大多数方法仅仅是将两者进行简单的融合,并未考虑融合方式的合理性。原因在于,自然语言查询中的文本信息和搜索区域中的视觉信息有时可能是不一致的,在这种情况下直接将两者融合可能会引起冲突,从而导致对于目标跟踪位置的错误估计。为此,DecoupleTNL[9] 设计了 2 个联合优化任务,即短时上下文匹配任务和长时上下文感知任务。其中,短时上下文匹配任务用于采集一段时期内的动态上下文信息,而长时上下文感知任务用于收集全局的静态上下文信息。此外,该方法设计了一种长短时调制模块来有效融合 2 种不同的上下文信息。DecoupleTNL 结构示意如图 4 所示。

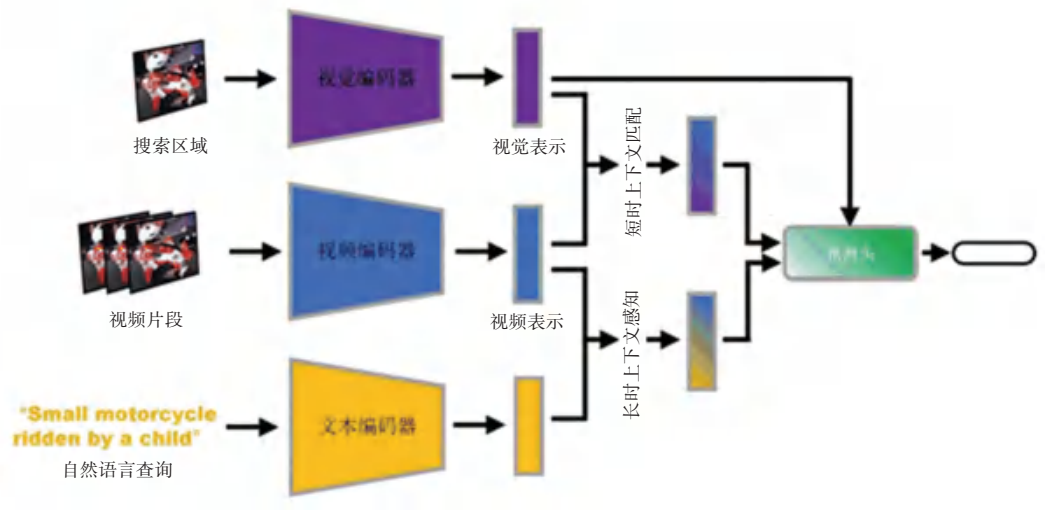


图 4 DecoupleTNL 结构示意图

Fig. 4 Framework of DecoupleTNL

2 实验数据分析

在评估基于自然语言查询的跟踪方法时,本文选择了 3 个带有自然语言查询标注的跟踪数据集。其中,OTB-lang[5] 数据集选取了 OTB100[10] 中 99 个视频序列并进行了自然语言标注。LaSOT[11] 数据集包含 1 400 个视频序列和与其对应的自然语言标注。TNL2k[12] 数据集收集了来自 YouTube 的 2 000 个视频片段,并对每个视频增加了自然语言标注。以上 3 个数据集均采用跟踪成功率(*Success*)评价指标进行性能评估。

考虑到比较的全面性,文本采用 2 种不同的方式来进行性能评估,即在有和无手工标定矩形框两种不同的条件下初始化跟踪器。

2.1 仅使用自然语言查询初始化跟踪器

在 OTB-Lang. LaSOT 和 TNL 2k 数据集上的跟踪性能比较见表 1。由表 1 可知,TNL[5] 在 OTB-lang 数据集上仅获得了 0.25 的成功率得分,而 RTNL[6] 在该数据集上获得了 0.54 的性能得分。当利用长短时上下文解耦的方式消除语言歧义时,DecoupleTNL[9] 在该数据集上获得了目前最优的性能。在 LaSOT 和 TNL2k 数据集中,CapsuleTNL[7] 分

别获得了 0.67 和 0.57 的得分，由于考虑了异质特征的上下文关系建模，所以 CapsuleTNL 的性能要优于 JointNLT[8]，即 0.59 和 0.57。

表 1　在 OTB-lang、LaSOT 和 TNL2k 数据集上的跟踪性能比较

Table 1　Performance comparison on OTB-lang, LaSOT and TNL2k

算法	仅使用自然语言查询初始化						同时使用自然语言查询和矩形框初始化					
	TNL[5]	RTNL[6]	SNLT[6]	CapsuleTNL[7]	JointNLT[8]	DecoupleTNL[9]	TNL[5]	RTNL[6]	SNLT[6]	CapsuleTNL[7]	JointNLT[8]	DecoupleTNL[9]
OTB-lang[5]	0.252	0.542	–	0.672	0.592	0.695	0.553	0.613	0.666	0.711	0.653	0.738
Lasot[10]	–	0.284	0.473	0.572	0.569	0.649	–	0.353	0.540	0.615	0.636	0.712
TNL2k[11]	–	–	–	–	0.546	0.407	–	0.250	0.248	–	0.569	0.567

2.2　同时使用自然语言查询初始化跟踪器

由表 1 可知，使用自然语言查询和矩形框初始化的跟踪器共有 6 个。具体而言，TNL[5] 在 OTB-lang 数据集上获得了 0.553 的成功率，而 RTNL[6] 则达到了 0.613。JointNLT[8] 结合了视觉定位和跟踪，在 OTB-lang 和 LaSOT 数据集上分别获得了 0.65 和 0.64 的成功率得分。相比之下，DeocupleTNL 在 OTB-lang 上取得了 0.74 的成功率得分，在 LaSOT 数据集上获得了 0.71 的成功率得分，并在 TNL2k 数据集上获得了 0.57 的最优性能。由此可见，在融合视觉和自然语言特征表示时，有效消除语言歧义是十分必要的。

3　结束语

在本文中，将基于自然语言查询的视觉目标跟踪方法按不同的主干网络分为基于卷积神经网络、基于胶囊网络、以及基于 Transformer 结构 3 大类，并对每一类中的方法进行了概述，对尚存在的不足等方面进行总结。其次，在性能比较方面可以看出，目前，在仅使用自然语言查询初始化跟踪器时，各方法的性能要远低于同时使用自然语言查询和矩形框的情况。因此，准确地初始化对于使用自然语言查询的跟踪器尤为关键。此外，现有的方法并未深入考虑视觉特征中的局部和全局信息与自然语言查询之间的多尺度对齐。最后，虽然 TNL2k 是目前针对该任务所提出的体量最大的数据集，但其中大部分视频来自动漫和游戏场景，与真实场景还存在一定差距，有必要提出更加贴合实际场景的数据集。

参考文献

[1] HARE S, GOLODETZ S, SAFFARI A, et al. Struck: Structured output tracking with kernels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2096-2109.

[2] HELD D, THRUN S, SAVARESE S. Learning to track at 100 fps with deep regression networks[C]//14th European Conference on Computer Vision(ECCV 2016). Cham: Springer, 2016: 749-765.

[3] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37(3): 583-596.

[4] WANG Xiao, LI Chenglong, LUO Bin, et al. Sint++: Robust visual tracking via adversarial positive instance generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4864-4873.

[5] LI Zhenyang, TAO Ran, GAVVES E, et al. Tracking by natural language specification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6495-6503.

[6] FENG Qi, ABLAVSKY V, BAI Qinxun, et al. Robust visual object tracking with natural language region proposal network[J]. arXiv preprint arXiv, 1912.02048, 2019.

[7] MA Ding, WU Xiangqian. Capsule-based object tracking with natural language specification[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 1948-1956.

[8] ZHOU Li, ZHOU Zikun, MAO Kaige, et al. Joint visual grounding and tracking with natural language specification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23151-23160.

[9] MA Ding, WU Xiangqian. Tracking by Natural Language Specification with long short-term context decoupling[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2023: 14012-14021.

[10] WU Yi, LIM J, YANG M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2411-2418.

[11] FAN Heng, LIN Liting, YANG Fan, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5374-5383.

[12] WANG Xiao, SHU Xiujun, ZHANG Zhipeng, et al. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13763-13773.