文章类型：论文|刊号（ISSN）：2737-4505(P) / 2737-4513(O)

# 面向高速公路场景的语义增强跨摄像头车辆跟踪方法

郭玮廷
长江大学
DOI:10.32629/etd.v7i2.18958

[摘　要] 针对高速公路场景下跨摄像头多目标跟踪(MTMCT)中,因光照变化、视角差异及目标遮挡导致的车辆重识别(ReID)精度下降和跨镜关联效率低下的问题,本文提出了一种基于语义增强与向量检索的跨摄像头车辆跟踪方法。首先,在单摄像头视角下,采用YOLOv11目标检测算法结合ByteTrack跟踪器,实现高精度的车辆轨迹提取。其次,为解决跨镜特征匹配的鲁棒性问题,本文改进了CLIP-ReID框架,设计了视觉上下文提示(Visual Context Prompt)模块以融合环境先验信息,并提出语义对齐损失(Semantic Alignment Loss)以缩小视觉特征与文本语义特征之间的异构差距,从而提取出更具鉴别力的车辆外观特征。最后,针对海量车辆特征跨镜匹配带来的计算瓶颈,引入ClickHouse向量数据库进行高维特征的存储与高效检索,完成跨摄像头轨迹的全局关联。在VeRi-776公开数据集与自建高速公路车辆数据集上的实验表明,本文方法在保持较高实时性的同时,显著提升了跨镜多目标跟踪的准确率与鲁棒性。
[关键词] 多目标跟踪；车辆重识别；语义增强；向量检索
中图分类号：F407.472　文献标识码：A

## Semantic-Enhanced Cross-Camera Vehicle Tracking Method for Highway Scenarios

Weiting Guo
Yangtze University

[Abstract] Aiming at the problems of degraded vehicle re-identification (ReID) accuracy and low cross-camera association efficiency caused by illumination changes, viewpoint variations, and target occlusions in Multi-Target Multi-Camera Tracking (MTMCT) for highway scenarios, this paper proposes a cross-camera vehicle tracking method based on semantic enhancement and vector retrieval. First, in the single-camera view, the YOLOv11 object detection algorithm combined with the ByteTrack tracker is used to achieve high-precision vehicle trajectory extraction. Second, to address the robustness of cross-camera feature matching, this paper improves the CLIP-ReID framework by designing a Visual Context Prompt module to integrate environmental prior information, and proposes a Semantic Alignment Loss to bridge the heterogeneous gap between visual features and text semantic features, thereby extracting more discriminative vehicle appearance features. Finally, targeting the computational bottleneck caused by massive vehicle feature matching across cameras, the ClickHouse vector database is introduced for the storage and efficient retrieval of high-dimensional features to complete the global association of cross-camera trajectories. Experiments on the public VeRi-776 dataset and a self-made highway vehicle dataset show that the proposed method significantly improves the accuracy and robustness of cross-camera multi-target tracking while maintaining high real-time performance.
[Key words] Multi-Target Multi-Camera Tracking; Vehicle Re-identification; Semantic Enhancement; Vector Retrieval

## 引言

随着智能交通系统(Intelligent Transportation Systems, ITS)的快速发展,高速公路视频监控网络的覆盖率逐年提升。跨摄像头多目标跟踪(Multi-Target Multi-Camera Tracking, MTMCT)作为ITS中的核心技术,旨在跨越多个非重叠视野的摄像头持续追踪同一车辆,在交通流量分析、肇事车辆逃逸追踪等领域具有重大的应用价值。

传统的多目标跟踪方法通常遵循“检测-跟踪”(Tracking-by-Detection)范式,但在跨镜场景下面临巨大挑战。首先,由于不同摄像头之间的物理位置、拍摄角度、光照条件存在显著差

文章类型：论文|刊号（ISSN）：2737-4505(P) / 2737-4513(O)

异,同一车辆在不同镜头下的视觉表观特征会发生剧烈变化,导致传统的车辆重识别(ReID)模型容易产生误匹配。其次,纯视觉特征在处理相似车型、同色车辆时鉴别力有限,缺乏对高层语义信息的理解。近年来,对比语言-图像预训练模型(CLIP)在跨模态表征学习上展现出强大的泛化能力,为解决上述问题提供了新思路。此外,随着监控视频流的增加,跨镜重识别阶段需要进行海量特征比对,传统的穷举匹配方法计算复杂度过高,难以满足实际工程的时效性需求。

## 1 相关工作

1.1单摄像头多目标跟踪。跨摄像头跟踪技术以传统的单摄像头多目标跟踪(MOT)为基础。MOT的任务是在单个视频序列中对多个目标进行识别与身份保持。核心技术包括目标检测、特征提取、运动预测与轨迹关联等。传统MOT很多算法首先利用目标检测器识别每一帧中的目标,再使用关联算法将检测到的目标与当前已有的轨迹进行对应。

随着深度学习的发展,现代MOT方法大多融合了深度特征学习与高效关联策略。例如近年来非常流行的ByteTrack策略通过融合高置信度和低置信度检测结果进行二次匹配,能够更好地处理/应对遮挡、交叉和短暂丢失等复杂场景,实现鲁棒跟踪。这类方法主要解决单摄像机视野下目标的连续跟踪与身份保持问题。

1.2车辆重识别与语义增强。对于跨摄像头跟踪来说,仅靠运动轨迹信息不足以满足精确匹配,因为不同摄像机的视角不重叠,目标也可能长时间消失在一个摄像头视野后才出现在另一个摄像头中。此时就必须借助视觉特征来识别同一辆车,这就是重识别(Re-ID)技术的作用。

重识别技术最初在行人追踪领域被提出,其目标是判断不同摄像机视野下是否为同一个目标。Re-ID的关键是提取具有区分性和鲁棒性的视觉特征,使得即使是在不同环境条件下,同一车辆的特征仍然能够被正确识别和匹配。Re-ID任务可以看成是一种大规模跨镜头检索问题,在监控视频中给定一辆车的查询图像,需要从不同摄像头拍摄的大量车辆候选集中检索出同一辆车对应的实例。

在传统视觉Re-ID中,大量使用卷积神经网络提取车辆视觉特征,但车辆外观变化极大,例如不同光照、遮挡、镜头分辨率差异以及看似相同的车辆外观等都使得特征区分变得困难。为了解决这个问题,越来越多研究引入了基于更强泛化能力的预训练模型、跨模态学习方法,以及引导式的语义提示机制等。

1.3多摄像头跟踪中的数据关联与跨镜头匹配。与单摄像头跟踪强调同一场景的对象连续相关性不同,跨摄像头跟踪需要依赖重识别特征和时空信息来建立不同视角之间的匹配关系。研究表明,在跨镜头匹配中,时空信息、摄像机之间的转移概率模型、场景拓扑关系等都能够提高匹配的准确性。

一个常见策略是通过构建摄像机之间的转移图或者拓扑关系来辅助匹配,这包括摄像机之间的物理相对位置、摄像机之间可能的目标移动路径和时间窗口等。这种策略的目标是将车辆在不同摄像头下出现的可能性进行量化,从而提高跨镜头匹配的准确性。

1.4向量检索系统在大规模视觉任务中的应用。在大规模监控网络中,单纯通过暴力计算方式进行跨镜头匹配将带来巨大的计算压力。因此,近年来一些研究提出借助向量数据库(如ClickHouse、Milvus等)来加速特征检索。所谓向量数据库是为高维向量匹配优化的数据库系统,它针对海量视觉特征向量数据构建高效索引,从而在实时性要求高的场景中快速完成相似度检索。

这些向量数据库系统通常采用近似最近邻(ANN)搜索技术,在保证较高准确率的同时实现亚秒级甚至更快的检索响应。这使得在大规模场景下的跨摄像头匹配问题具备可用性

## 2 本文方法

2.1单镜头车辆检测与跟踪模块。该模块解决的是如何在每一个摄像头视野中对车辆进行检测与跟踪的问题,它是跨摄像头系统最底层的组成部分。

车辆检测是该模块的首要任务。本文采用的目标检测器是工业界广泛使用的YOLOv11,其特点是速度快、检测精度高,尤其对不同尺度的车辆都具有良好的表现能力。YOLO系列检测器通过端到端神经网络结构实现输入图像的实时推理,适合实时监控场景中对目标检测性能的要求。

在检测后,需要将连续帧中检测到的同一车辆进行轨迹关联,这一步采用的是ByteTrack跟踪算法。ByteTrack强调对低置信度检测结果的再利用,通过二次关联减少因遮挡、漏检等情况导致的跟踪中断,使得轨迹求解更加完整、鲁棒。

单镜头跟踪模块的结果最终会生成每个摄像头视野中各车辆的完整轨迹列表,包括轨迹开始时间、结束时间、车辆外观特征等供跨镜头匹配使用。

2.2车辆重识别特征提取与语义增强。传统重识别任务通常仅依赖视觉特征提取网络来生成车辆的外观表征。然而在跨摄像头环境下,仅凭视觉特征往往难以克服环境变化及车辆视觉相似性问题。

为了提升重识别性能,本文提出将视觉特征与语义提示结合的策略,通过利用更高级别语义信息来辅助学习视觉表征。这种方法主要包含两个核心内容：

2.2.1视觉上下文语义提示机制。重识别任务中常用的视觉外观特征并不包含场景或背景信息,而在交通场景中,摄像头的地理位置、背景环境、时间上下文等信息往往能够为同一辆车的匹配提供辅助信息。因此,我们将车辆本身的视觉特征与该车辆在检测帧的场景上下文一并作为学习条件输入。

该处理方式先通过神经网络提取整帧图像的场景上下文特征,再与车辆的基础属性(如车型、颜色类别等)组合生成动态语义提示向量,作为重识别网络的文本或其他辅助信号输入。

这种方式类似于跨模态提示策略,它赋予模型根据场景动态调整对视觉特征关注的能力,使得在复杂环境下仍能保持较高辨识能力。

2.2.2语义对齐特征学习策略。在进行跨摄像头车辆重识别时,训练阶段通过语义对齐机制使视觉特征空间和语义提示空

间能够互相对齐。这样一来,重识别网络不仅学习车辆的视觉表征,还学习如何将这些表征映射到一个与语义提示一致的高层空间,使得不同摄像头下的同一车辆更容易被正确匹配。

这种对齐机制借助对比学习策略,对正样本对(相同车辆不同视角)进行拉近,对负样本对(不同车辆)进行区分,从而在特征空间中形成具有更强区分性和一致性的表征分布。

2.3向量检索加速跨镜头匹配。在获得每个摄像头视野下车辆提取的轨迹表征后,需要在这些轨迹之间进行全局比对来生成跨镜头匹配结果。

考虑到监控系统中摄像头数量多、轨迹数量庞大,传统基于距离度量的全部两两比对策略已经无法满足实时性需求。因此本文引入了向量检索数据库来加速匹配过程。

2.3.1向量数据库系统架构。向量数据库是一种针对高维向量相似性查询进行优化的数据库系统。与传统数据库主要处理结构化文本或数值数据不同,向量数据库专门针对深度特征向量进行索引与快速检索。通过构建高效索引结构(如层次化图索引、近似最近邻索引等),能够在大规模数据中进行近似快速搜索。

在本方法中,每条车辆轨迹的外观特征向量会被存入向量数据库,并根据该特征建立索引。在执行跨镜头匹配时,对于每个轨迹的特征向量,在数据库中进行快速检索,从而召回可能匹配的车辆轨迹候选集合,随后再结合时空约束等信息进行最终匹配决策。

2.3.2时空约束辅助匹配机制。除了视觉外观检索之外,为了进一步提升跨镜头匹配的精度,时空约束策略被引入作为匹配辅助信息。在真实交通场景中,同一辆车在一个摄像头视野消失和在另一个摄像头视野出现之间必然存在时间间隔,同时车辆需要沿交通路网移动,这些规律可以被显式地建模：

时间窗口限制：利用车辆消失后的合理时间范围作为可能出现匹配的先验。

摄像机地理位置关系：结合摄像机之间的物理空间距离与可能的路径转移来限制匹配候选。

结合以上时空约束条件,可有效排除不可能的匹配关系,减少误匹配概率。

## 3 实验与分析

3.1实验环境与数据集。实验基于硬件平台NVIDIA RTX GPU展开,深度学习框架为PyTorch,后端系统集成采用Python+FastAPI。

评测数据集包括：(1)VeRi-776：包含大量车辆跨镜重识别数据,用于验证特征提取模型的泛化性。(2)自建高速公路车辆数据集：采集自真实高速公路监控,包含复杂光照、大视角跨度及拥堵遮挡等真实工况,用于验证整套MTMCT方法的实际效能。

3.2消融实验。为验证本文提出的语义增强模块的有效性,在自建数据集上进行了四组消融实验：

Baseline：基础的原始CLIP-ReID模型作为特征提取器。

Baseline+Prompt：仅加入视觉上下文提示(Visual Context Prompt)模块。

Baseline+Loss：仅加入语义对齐损失(Semantic Alignment Loss)模块。

Ours(Both combined)：同时应用上述两种改进。

实验结果表明,Baseline模型的mAP表现一般,受限于跨镜表观差异。加入Prompt模块后,模型能感知环境背景,Rank-1指标有稳步提升；加入Loss模块后,视觉特征受到文本语义的强监督,鉴别力增强。当两者结合时(本文方法),各项指标达到最优,证明了视觉提示与语义对齐在跨模态车辆特征学习中的协同促进作用。

3.3系统整体跟踪效果对比。在整体跨镜跟踪性能上,将本文系统与主流的纯视觉MTMCT方法(如基于ResNet50+DeepSORT的基线)进行对比。得益于YOLOv11更精准的检测与改进的CLIP-ReID特征,本文方法在IDF1(反映跨镜轨迹稳定性和ID切换次数)上取得了显著优势。同时,依靠ClickHouse向量检索技术的引入,跨镜匹配阶段的处理耗时大幅降低,充分证明了本文方法在高速公路海量数据场景下的工程应用价值。

## 4 总结与展望

本文针对高速公路车辆跨镜跟踪面临的挑战,提出了一套融合深度学习与大数据检索技术的系统框架。通过YOLOv11和ByteTrack保证底层跟踪精度,创新性地提出结合视觉上下文提示与语义对齐损失的CLIP-ReID网络,极大提升了跨镜车辆特征的鲁棒性；最后利用ClickHouse实现了高效的全局关联。消融实验和对比实验充分验证了各模块的有效性。

未来的研究工作将聚焦于两个方面：一是进一步探索极端天气(如暴雨、浓雾)条件下的多模态数据(如雷达数据)融合,提升环境适应性；二是研究长尾车型(如特种作业车辆)的少样本重识别问题,以完善复杂交通流分析系统的泛化能力。

## [参考文献]

[1]E.Ristani, F.Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target,multi-camera tracking," in European Conference on Computer Vision (ECCV),Springer,2016,pp.17-35.

[2]A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in 2016 IEEE International Conference on Image Processing(ICIP),2016,pp.3464-3468.

[3]N.Wojke,A.Bewley,and D.Paulus,"Simple online and realtime tracking with a deep association metric,"in 2017 IEEE International Conference on Image Processing (ICIP),2017,pp.3645-3649.

[4]Y.Zhang,P.Sun, Y. Jiang, D. Yu, F. Weng, Z. Zehuan, Y. Yuan, P.Luo, W. Liu,and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in European Conference on Computer Vision (ECCV),Springer,2022,pp.1-21.

[5]X.Liu,W.Liu,T.Mei,and H.Ma,"A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in European Conference on Computer Vision (ECCV), Springer,2016,pp.869-884.

## 作者简介：

郭玮廷(1998--),男,汉族,山西省晋城市人,研究生,研究方向：计算机视觉。