Computer Science and Application 计算机科学与应用, 2023, 13(7), 1382-1389
Published Online July 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.137136

基于DinoPose的列车司机手比行为检测研究

李　珂，王　鹏*

山东高速轨道交通集团有限公司益羊铁路管理处，山东　潍坊

收稿日期：2023年6月13日；录用日期：2023年7月11日；发布日期：2023年7月18日

摘　要

本研究针对铁路场景列车驾驶室驾驶员监控视频图像提出了一种列车司机手势动作识别算法模型DinoPose。通过引入Transformers中的编码器－解码器结构来实现基于回归的人体骨架关键点检测，有效地将Dino网络的应用场景从目标检测扩展至人体骨架检测。通过多组列车驾驶室的视频图像所抽取的关键帧数据集测试，本文提出法在精度上优于Openpose和Yolo-pose算法，其中mAP达到了95.72%，手比项点的检测准确率达到85.74%以上，能够满足铁路局机务段机车司机室监控视频智能分析的实际业务需求。

关键词

Dino网络，DinoPose，骨架点检测，Transformer

Research on Hand Signal Behavior Detection of Train Driver Based on DinoPose

Ke Li, Peng Wang*

Yiyang Railway Management Office, Shandong Hi-Speed Rail Transportation Group Co., Ltd., Weifang Shandong

Received: Jun. 13th, 2023; accepted: Jul. 11th, 2023; published: Jul. 18th, 2023

Abstract

This study proposes a train driver gesture action recognition algorithm model DinoPose for video images of train cab driver monitoring in railroad scenes. By introducing the encoder-decoder structure in Transformers to achieve regression-based human skeleton key point detection, the application scenario of Dino network is effectively extended from target detection to human

*通讯作者。

文章引用: 李珂, 王鹏. 基于 DinoPose 的列车司机手比行为检测研究[J]. 计算机科学与应用, 2023, 13(7): 1382-1389.
DOI: 10.12677/csa.2023.137136

skeleton detection. Tested by the key frame dataset extracted from multiple sets of video images of train cabs, the proposed method in this paper outperforms Openpose and Yolo-pose algorithms in terms of accuracy, where the mAP reaches 95.72% and the detection accuracy of hand ratio item points reaches more than 85.74%, which can meet the actual business requirements of intelligent analysis of locomotive driver's cab monitoring video in the locomotive section of railroad bureau.

Keywords

Dino Network, DinoPose, Skeleton Point Detection, Transformer

1. 引言

随着我国经济的高质量发展，铁路交通的运输生产力得到了快速提高，铁路的安全问题也越来越受到重视[1]。铁路营业里程的增加和机车运营速度的提高，给铁路安全运营带来了巨大压力，对铁路安全行车的保障技术提出了更高的考验。根据相关的调查研究，机车驾驶室的工作环境比较恶劣，容易引起司机疲劳驾驶而出现操作失误，这是导致铁路安全事故的一个因素[2]。司机能否在列车开车停车、进出站等重要节点做出正确手势是衡量工作质量的重要指标。不正确的手势判断不利于列车运行安全，因此对司机手势动作的监控识别十分重要。机务段依赖人工方式对机车司机室监控视频，考核司机的违章行为，工作量繁重且耗时较长。因此需要探索出一种智能化分析机车司机室监控视频的方法，达到准确、快速地识别司机手势动作的目的，不仅可以减少监管系统的人力资源浪费，还能够帮助列车更加安全高效地运行。

近年来，基于深度卷积神经网络的计算机视觉技术在轨道交通司机行为识别领域受到了相关研究人员的广泛青睐。文献[3]首先利用 Alphapose 姿态估计算法[4]提取人体骨骼的关键点坐标，设计了行为分类器对手势行为的骨架点坐标进行预分类识别，然后通过 YOLOv5 目标检测算法检测感兴趣物体的位置，最后将骨架检测和目标检测结果进行融合决策得到行为的识别结果，大幅度提高了对各类危险驾驶行为检测的准确性和鲁棒性。文献[5]提出了一种基于 OpenPose [6]神经网络模型的肢体识别算法，结合目标检测模型和手势动作的知识库，对司机驾驶过程中的手指确认等操作进行检测和判断是否标准。文献[7]设计了一种基于区域三维卷积神经网络(Region Convolutional 3D Network, R-C3D)的司机手势识别模型。通过深度学习神经网络的训练和调优措施，凭借着 RC3D 网络中的特征提取子网络、时序提议子网络和行为分类子网络，可以实现高准确率且快速地识别和定位司机手势的动作。文献[8]提出一种多时空尺度的融合网络 RepC3D (Re-parameterization Convolutional3D)。首先通过背景消减法获取目标动作区域，然后再对目标区域进行 RGB 和光流转化，接着输入深度学习神经网络，并提取出视频时序特征和空间特征，最后经过 RepC3D 块将获取到的特征进行逻辑运算，获取到融合信息。在地铁司机手势动作的数据集中进行测试，该算法模型具有较高准确率和较低漏检率的性能表现，可以有效识别司机的手势行为。上述研究者全部采用深度卷积神经网络实现列车司机手比动作的检测。然而，基于深度卷积神经网络的姿态估计算法更加擅长提取局部特征，对于全局特征的理解具有一定的局限性，限制了检测精度的进一步提高，同时其推理时耗费计算资源较多，检测速度较慢。

Transformer 如今被广泛应用于自然语言处理和计算机视觉领域[9]。DETR (DEtection TRansformer) 作为首个使用 Transformer 做目标检测的模型，是一种端到端可学习的目标检测器，非常具有创新性且性能优越[10]。DETR 将目标检测建模成集合预测的任务，在训练过程中，使用二分图匹配预测和标签进行训练，而测试时不需要后处理即可产生所有结果。DAB-DETR [11]提出使用动态的锚框用于 DETR，直接将框坐标作为询问输出到 Transformer 的解码器中。在每一层动态更新 box。使用 box 可以加速训练收敛，同时可以使用 box 的高宽对位置注意力图建模。DN-DETR 引入去噪任务直接把带有噪声的真实框输入到解码器中，在 DAB-DETR 的基础上进一步加速了收敛。去噪任务仅在模型训练时出现，推理时并不需要，不会给模型的实际应用带来额外负担[12]。DINO 模型第一次让 DETR 类型的检测器取得了目标检测的最优性能，在 COCO 数据集上取得了 63.3 AP 的性能，相比之前该类型的检测器将模型参数和训练数据减少了十倍以上[13]。DINO 设计了训练模型识别负样本的方法，不仅要回归真实框，还需要辨别负样本。提出了混合查询选择方法，有助于改善询问的初始化。另外，DINO 引入非临近层的特征，增加了感受野的范围，提高小目标的表达能力。

本文在前人的研究基础上，提出了一种基于 DinoPose 的列车司机手势动作识别算法模型。首先采用本文所提出的 DinoPose 姿态估计算法，提取机车驾驶室监控视频中的列车司机右侧手臂和躯干关键骨架点的坐标位置信息，然后根据右臂和躯干的角度、位置和模长等空间信息，建立司机手比确认动作的识别模型，最后将识别结果和机车 LKJ 信息[14]进行关联分析，实现手比动作违章的检测和判断。根据试验结果，本文所设计的基于 DINO-POSE 的列车司机手势动作识别算法模型，通过 OKS (Object Keypoint Similarity)指标衡量人体关键骨架点的检测效果，其 mAP 达到了 95.72%，手比项点的检测准确率达到 85.74%以上。能够满足铁路局机务段机车司机室监控视频智能分析的实际业务需求，这对保障机车安全行驶，提高监管部门的工作效率有着重要意义。

2. 模型算法

2.1. DinoPose

本文在 Dino 网络的基础上提出 DinoPose，利用 Transformers 中的编码器－解码器结构来执行基于回归的行人和关键点检测，成功的将 Dino 从目标检测扩展到二维人体骨架点检测上。

DinoPose 由主干网络、Transformer 编码器、Transformer 解码器、关节点检测分支、目标框分支和置信度分支组成，如图 1 所示。输入的图像经过主干网络提取 C_3、C_4、C_5 的多尺度的特征后展平送入 Transformer 编码器，Transformer 编码器进一步融合了不同尺度的特征并提取得分最高的前 N 个特征作为初始查询(query)，和经过融合的键(key)、值(value)特征送入 Transformer 解码器后进行计算，经过检测框分支和置信度分支得到 N 个人体的检测框(x, y, w, h)及置信度，关节点检测分支预测得到人体的各个关节点相对人体检测框中心点的偏移值$(\Delta x, \Delta y)$和该关键点置信度。在推理过程中，我们首先得到大于置信度阈值的人体检测框，随后将人体检测框的中心点(x, y)与各关键点偏移值相加得到人体各个关键点$(x+\Delta x_i, y+\Delta y_i)$，只保留置信度大于 0.5 的关键点并过滤掉人体检测框之外的关键点。与基于特征图目标框回归的 2D 人体骨架点模型算法相比，DinoPose 通过利用 Transformer 范式把检测问题转换为预测集的问题，实现了对模型预测结果和图片中位置的解耦。DinoPose 避免了传统的稠密目标检测算法中多个预测值对应于一个真实样本的问题，从而去除了后处理非极大值抑制算法，更好的适配于现实拥挤场景问题，实现了真正的端到端。

2.2. 关键点去噪训练

与 DINO 相同，DinoPose 使用基于集合的匈牙利损失，对每个输出结果和真实样本进行一对一的

标签匹配。我们使用 Focal loss 损失函数作为人体置信度损失 L_{Pconf} 和关键点置信度损失 L_{Kconf}，对于人体检测框使用 L1 损失和 GIOU 损失进行回归(合为 L_{box})，对于人体关键点回归，使用 L1 损失 L_{l1} 和 OKS 损失 L_{oks}。最常用的 L1 损失对小目标姿态和大目标姿态具有不同的尺度差异，为了缓解这个问题，本文额外使用对象关键点相似性(Object Keypoint Similarity, OKS)损失进行人体关键点回归，其可以表示为下式：

$$L_{oks}\left(P,P^{*}\right)=\frac{\sum_{i}^{K}\exp\left(-\left\|P_{i}-P_{i}^{*}\right\|/2s^{2}k_{i}^{2}\right)\delta\left(\upsilon_{i}>0\right)}{\sum_{i}^{K}\delta\left(\upsilon_{i}>0\right)} \tag{1}$$

式中，$\left\|P_i-P_i^*\right\|$ 是第 i 个预测关键点和真实关键点之间的欧式距离，υ_i 是真实关键点的可见性标志，s 是对象比例，k_i 是每个关键点的常数系数。如上所示，OKS 损失通过进行归一化，使每个关键点的重要性相等，从而缓解不同的尺度姿态目标的损失差异。

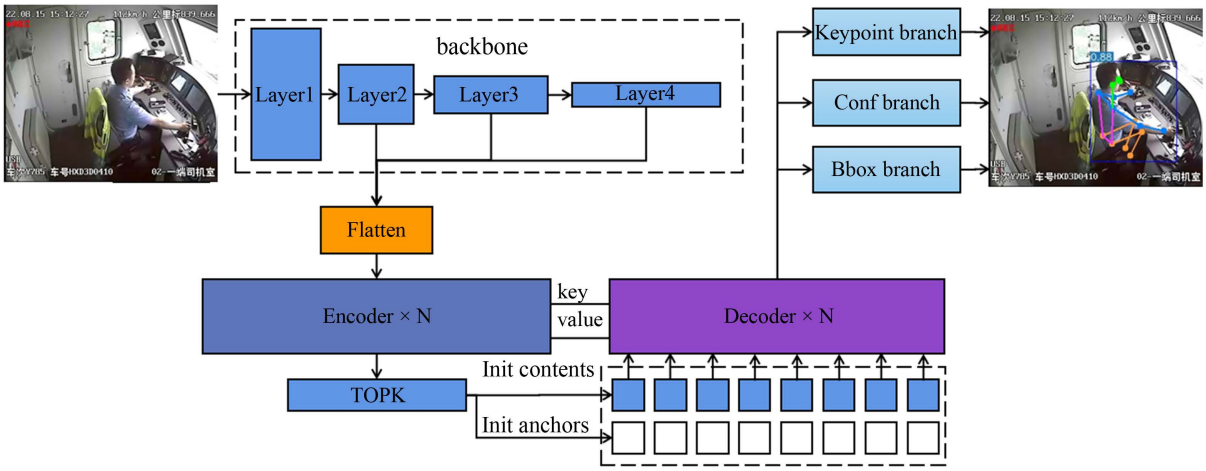

Figure 1. Structure diagram of DinoPose
图 1. DinoPose 结构图

由此，DinoPose 的一对一的标签匹配可以表示为下式：

$$C=argmin\sum_{i}^{N}L_{match}\left(y_{i},\hat{y}_{\sigma(i)}\right) \tag{2}$$

$$L_{hungarian}=\lambda_{1}L_{Pconf}+\lambda_{2}L_{box}+\lambda_{3}L_{l1}+\lambda_{4}L_{oks}+\lambda_{5}L_{Kconf}$$

式中，λ_1、λ_2、λ_3、λ_4、λ_5 分别代表损失的权重。

DINO 使用去噪训练技术，在稳定训练和加速收敛方面非常有效。我们同样在本文中将去噪思想引入到关键点训练中去，如图 2 所示。我们将添加噪声的真实目标框及其标签提供给解码器，并训练模型来重建目标框和目标关键点。

因此，DinoPose 的损失函数可由下式表述：

$$L=\sum_{i=0}^{M}L_{hungarian}+\sum_{i=0}^{M}\sum_{j=0}^{M}L_{denosing} \tag{3}$$

式中，$L_{hungarian}$ 代表基于匈牙利匹配的损失，$L_{denosing}$ 代表去噪损失，M 代表模型的解码器层数，N 代表总共有 N 组去噪损失。

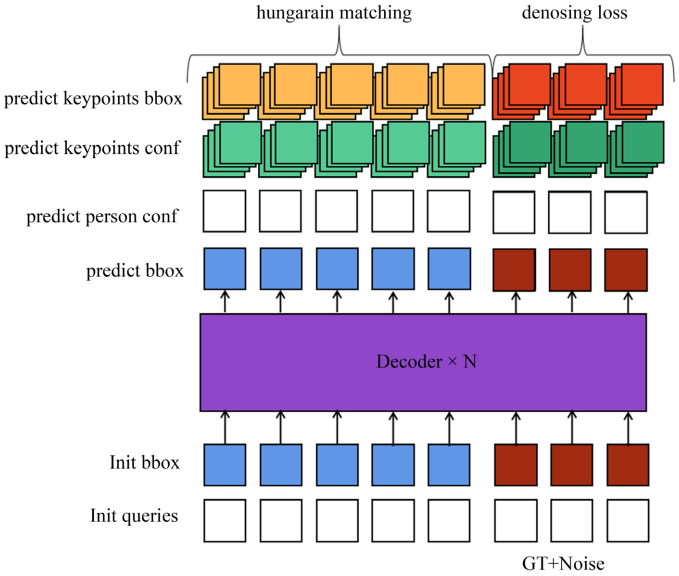

Figure 2. Denoising training of Dinopose
图 2. Dinopose 去噪训练

2.3. 后处理行为判别

当完成人体关键点估计后，我们可以通过分析人体各关节之间的角度关系得到司机的姿态和动作，通过人体各关节点坐标相减得到人体肢体的向量，通过余弦公式得到肢体之间的角度。通过计算右大臂和躯干角度 α、右大臂与右小臂角度 β 辨别司机是否正进行手势，设定如果 α 大于 90°且 β 大于 90°时，判定司机处于做手势的状态，如图 3 所示。

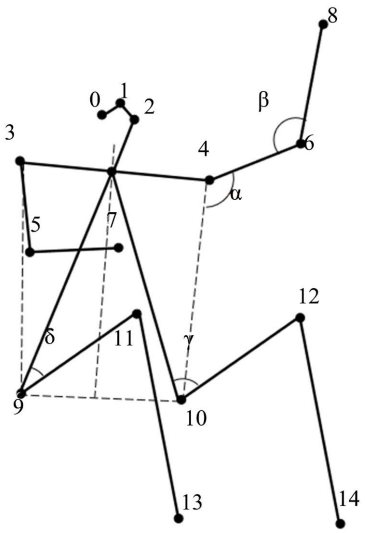

Figure 3. Diagram of human skeleton
图 3. 人体骨架示意图

3. 实验及结果

为了验证本文所提出算法的效果，本文搭建 DinoPose 网络模型进行了训练，并针对各项优化的效果进行评估。

3.1. 数据收集与参数设置

本文采集了多组列车驾驶室的数据抽取关键帧作为训练集和测试集，标记其中出现的司机人体关键点和人体目标框，标注数据采用 COCO 关键点标注格式，共标注左右眼、左右耳、鼻子、左右肩、左右肘、左右手、左右髋、左右膝、左右踝 17 个人体关键点。共标注 13,258 张图片，数据按 10:1 的比例随机分为训练集和测试集，分别为 11,933 张和 1325 张。其中测试集共有 364 个手比行为。训练集共有 4328 个手比行为。

本文在 8 块 A30 显卡上进行了训练，模型的主干网络设置为经过 Imagenet 数据集预训练的 ResNet50，Transformer 的编码层、解码层设置为 6 层，训练集和测试集图片统一输入尺寸为(384, 288)，批尺寸(batch size)设置为 16，训练步数(epoch)设置为 200，采用 Adam 梯度下降方法，初始学习率为 0.0001。本文采用 OKS 作为每张图片人体关键点识别好坏的衡量标准。统计测试集所有图片 OKS 计算 MAP (均值平均精度)作为测试集的衡量指标。

3.2. 实验结果分析

表 1 和图 4 统计了训练后 DinoPose 的精度以及与其他常规算法的对比。相比于 Openpose 和 Yolo-pose 算法，本文在精度做到了最好，在耗时上，本文的方式超越了 Openpose 方法，但略逊于 Yolo-pose 算法。DinoPose 的检测效果如图 5 所示。

Table 1. Performance comparison of DinoPose accuracy with other algorithms
表 1. DinoPose 精度与其他算法性能比较

模型	MAP@0.5::0.95	推理时间 ms
DinoPose	0.9572	23.9
Yolo-pose	0.9357	18.4
openpose	0.9536	32.7

Figure 4. Performance comparison of DinoPose accuracy with other algorithms
图 4. DinoPose 精度与其他算法性能比较图

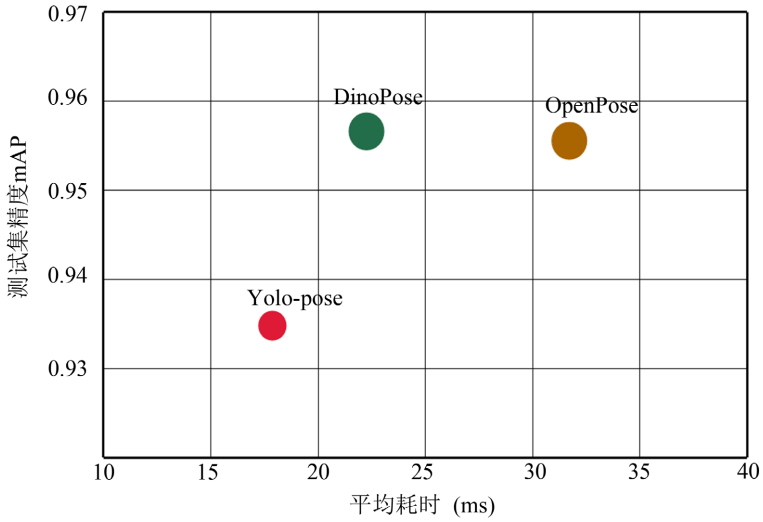

Figure 5. Detection effect picture of DinoPose
图 5. DinoPose 检测效果图

表 2 所示展示了基于角度分析的人体检测后处理的测试集手比识别情况，我们使用精确率(Precision)和召回率(Recall)衡量在测试验证集上的基于角度分析的人体检测后处理的效果，如表 3 所示。

Table 2. Test set gesture detection statistics
表 2. 测试集手势检测统计

	预测手比	预测未手比
实际手比	312	52
实际未手比	56	1269

Table 3. Effectiveness of gesture detection based on angle analysis
表 3. 基于角度分析的手势检测效果

精确率	召回率
84.78%	85.71%

4. 结论

本文利用 Transformers 中的编码器－解码器结构来实现基于回归的人体骨架关键点检测，提出了一种端到端的列车司机手势动作识别算法模型 DinoPose。DinoPose 由主干网络、Transformer 编码器、Transformer 解码器、关节点检测分支、目标框分支和置信度分支组成，将 Dino 网络从目标检测成功扩展到二维人体骨架关键点检测。通过多组列车驾驶室的视频图像所抽取的关键帧数据集测试，本文所提出的算法在精度上优于 Openpose 和 Yolo-pose 算法，能够满足铁路局机务段机车司机室监控视频智能分析的实际业务需求。

考虑到本文所提出方法的精度测试仅局限于铁路场景列车驾驶室驾驶员视频图像数据集，此后还需要在其他公开数据集上与各基准方法进行对比测试，进而验证本文所提出方法的泛化能力，扩大其应用范围。此外，由于本文所提出方法与实际铁路业务产品强相关，还需要考虑对 DinoPose 进行边缘端的应用部署。

参考文献

[1] 徐瑞. 市场经济条件下的铁路交通运输经济管理[J]. 中关村, 2022(7): 112-113.

[2] 贾子若. 铁路机车司机工作压力与安全绩效关系研究[D]: [博士学位论文]. 北京交通大学, 2013.

[3] 王永硕. 列车司机不规范行为监测系统设计[D]: [硕士学位论文]. 北京: 北京交通大学, 2022 https://doi.org/10.26944/d.cnki.gbfju.2022.002204

[4] Fang, H.S., Li, J., Tang, H., *et al*. (2022) Alphapose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 7157-7173. https://doi.org/10.1109/TPAMI.2022.3222784

[5] 叶鹏君. 基于图像识别的列车司机驾驶行为监测及关键技术研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2020. https://doi.org/10.26944/d.cnki.gbfju.2020.000792

[6] Cao, Z., Hidalgo, G., Simon, T., *et al*. (2018) Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1302-1310. https://doi.org/10.1109/CVPR.2017.143

[7] 所达. 城轨列车司机行车确认手势动作行为识别方法研究[D]: [硕士学位论文]. 北京: 北京交通大学,2021. https://doi.org/10.26944/d.cnki.gbfju.2021.001694

[8] 王涛. 基于 RepC3D 模型的地铁司机手势动作识别[D]: [硕士学位论文]. 武汉: 武汉纺织大学, 2022. https://doi.org/10.27698/d.cnki.gwhxj.2022.000116

[9] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. (Preprint)

[10] Carion, N., Massa, F., Synnaeve, G., *et al*. (2020) End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *ECCV* 2020: *Computer Vision-ECCV* 2020, *Lecture Notes in Computer Science*, Vol. 12346, Springer, Cham, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[11] Liu, S., Li, F., Zhang, H., *et al*. (2022) DAB-DETR: Dynamic Anchor Boxes Are Better Queries for DETR. (Preprint)

[12] Li, F., Zhang, H., Liu, S., *et al*. (2022) DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. 2022 *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 13609-13617. https://doi.org/10.1109/CVPR52688.2022.01325

[13] Zhang, H., Li, F., Liu, S., *et al*. (2022) DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. (Preprint)

[14] 杨志刚. LKJ2000 型列车运行监控记录装置[M]. 北京: 中国铁道出版社, 2003.