Computer Science and Application 计算机科学与应用, 2025, 15(4), 416-431
Published Online April 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.154114

Hans汉斯

基于深度学习与Mediapipe的列车司机手比行为检测方法研究

潘荣壮，田　野，刘雷新元，袁小军，李　晨，袁希文

株洲中车时代电气股份有限公司数据与智能技术中心，湖南　株洲

收稿日期：2025年3月25日；录用日期：2025年4月23日；发布日期：2025年4月30日

摘　要

列车司机的行为监测在智能交通系统中对于提高安全性和减少交通事故至关重要。针对列车司机在驾驶过程中手比行为的识别，本研究提出了一种基于深度学习与Mediapipe技术相结合的手比行为检测方法。研究重点在于提升手比行为的检测精度与实时性，尤其是在复杂环境下的应用。研究首先使用ResNet50卷积神经网络(CNN)对列车驾驶舱图像数据集进行训练，完成对掌手比与指手比的分类任务。通过对不同手势类型的数据进行训练，模型成功实现了超过85%的准确率，验证了深度学习在此类行为识别中的有效性。此外，研究采用了Mediapipe框架，通过实时的手部关键点检测与姿态估计，基于动态视频数据对智轨司机的手比行为进行了分析。该方法结合关键点之间的几何关系，准确率达到90%，能够在动态驾驶环境中实现高效的行为识别。本研究的创新性在于，结合深度学习的特征提取能力与Mediapipe的实时骨架点检测，优化了手比行为的检测精度和环境适应性。通过实验验证，提出的检测方法能够在复杂环境下稳定运行，具有显著的实时性和鲁棒性。这为智能交通系统中的司机行为监控提供了新的技术路径，尤其在提升智能驾驶舱安全性和交互效率方面具有重要应用价值。

关键词

司机行为检测，手比行为，Mediapipe，深度学习，骨架点识别

Research on Train Driver Hand Gesture Behavior Detection Method Based on Deep Learning and Mediapipe

Rongzhuang Pan, Ye Tian, Leixinyuan Liu, Xiaojun Yuan, Chen Li, Xiwen Yuan

Data & Intelligent Technology Center, Zhuzhou CRRC Times Electric Co., Ltd., Zhuzhou Hunan

Received: Mar. 25th, 2025; accepted: Apr. 23rd, 2025; published: Apr. 30th, 2025

Abstract

The monitoring of train driver behavior is crucial for enhancing safety and reducing traffic

文章引用: 潘荣壮, 田野, 刘雷新元, 袁小军, 李晨, 袁希文. 基于深度学习与 Mediapipe 的列车司机手比行为检测方法研究[J]. 计算机科学与应用, 2025, 15(4): 416-431. DOI: 10.12677/csa.2025.154114

accidents in intelligent transportation systems. This study proposes a hand gesture behavior detection method for train drivers during operation, which combines deep learning with Mediapipe technology. The focus of the research is to improve the detection accuracy and real-time performance of hand gestures, especially in complex environments. The study first uses the ResNet50 convolutional neural network (CNN) to train a dataset of train cockpit images, completing the classification task of palm gestures and finger gestures. By training on different gesture types, the model successfully achieved an accuracy rate exceeding 85%, validating the effectiveness of deep learning in such behavior recognition tasks. Additionally, the research employs the Mediapipe framework for real-time hand keypoint detection and posture estimation, analyzing the hand gesture behaviors of smart track drivers based on dynamic video data. The method, which incorporates the geometric relationships between keypoints, achieved an accuracy rate of 90%, enabling efficient behavior recognition in dynamic driving environments. The novelty of this study lies in the integration of deep learning's feature extraction capabilities with Mediapipe's real-time skeletal point detection, optimizing the detection accuracy and environmental adaptability of hand gestures. Experimental validation shows that the proposed detection method can operate stably in complex environments, demonstrating significant real-time performance and robustness. This provides a new technical pathway for driver behavior monitoring in intelligent transportation systems, with substantial application value, particularly in enhancing the safety and interaction efficiency of intelligent cockpits.

Keywords

Driver Behavior Detection, Hand Gesture Behavior, Mediapipe, Deep Learning, Skeletal Point Recognition

1. 引言

列车司机在铁路运行安全中的作用至关重要，司机的操作行为直接关系到列车的安全性。研究表明，司机的认知负荷、疲劳及情绪状态会显著影响他们的操作表现和决策能力[1]。例如，在复杂的操作环境中，认知负荷的增加可能导致决策失误，进而影响列车的安全[2]。此外，过度工作和长时间驾驶常常导致司机的疲劳，这直接影响其警觉性和反应速度[3]。在紧急情况下，司机的应急决策能力和反应速度对于确保安全尤为重要[4]。因此，如何有效地减轻司机的认知负荷，并增强其应急反应能力，已成为铁路安全研究的重要方向。

随着铁路系统的逐步实现自动化，如何平衡自动化与人工操作，尤其是在紧急情况下保持司机的判断能力，是当前铁路安全研究的重点[5]。另外，列车司机的操作确认行为，尤其是手比行为，在复杂和高压的环境中对于减少操作错误、提高安全性起到了关键作用。在信息不完全或环境干扰较大的情况下，手比行为作为一种非语言的确认方式，有助于帮助司机迅速且准确地完成操作确认[6]。因此，深入研究并优化手比行为的应用，尤其在高风险环境下，将显著提高列车的安全性和操作效率。

目前，深度学习在行为检测中的应用已经取得了显著进展，特别是卷积神经网络(CNN)和循环神经网络(RNN)。CNN 能够从数据中自动提取特征，提高行为识别的精度，而 RNN 通过时序数据的处理能力，在动作序列识别中展现出优势[7]。在手比行为的检测上，深度学习的应用显著提升了行为识别的精度和实时性。

Mediapipe 是 Google 推出的一种高效跨平台计算框架，广泛应用于姿态估计、手势识别和面部表情检测等领域，尤其适用于实时图像和视频分析[8]。Mediapipe 的高效性和轻量级特性使其成为嵌入式设备和移动端的理想选择。结合深度学习和 Mediapipe 技术，能够显著提高行为检测的准确性和实时性，尤其在多任务场景下，展示了其强大的应用潜力。

尽管深度学习和 Mediapipe 技术在行为检测领域取得了显著进展，仍面临一些挑战。首先，深度学习模型通常需要大量的标注数据和计算资源，这在实际应用中可能成为瓶颈[9]。其次，目前的研究多集中于单一场景下的行为识别，缺乏对复杂环境中行为检测的深入探讨[10]。特别是在动态和复杂的环境中，现有的深度学习和 Mediapipe 结合方法仍面临性能不稳定和环境适应性差等问题[11]。

本研究的主要目标是提高列车司机手比行为在复杂环境中的检测精度和实时性。具体来说，研究将结合深度学习和 Mediapipe 两种不同的技术方案，探索如何在复杂铁路环境中进行高效、准确的手比行为识别。

首先，利用深度学习算法，从列车司机的手比行为中提取关键特征，并提高识别精度。其次，采用 Mediapipe 技术进行实时的姿态估计和手势识别，确保系统在动态和复杂环境下的稳定性和高效性。

该研究的创新之处在于：

结合深度学习和 Mediapipe 技术：通过对比两者的优势，实现手比行为的高效检测。

环境适应性优化：特别针对复杂和动态环境，优化现有方法的适应性和稳定性。

通过这些创新方法，本研究旨在为列车司机提供更为高效的操作确认机制，提升铁路运行的安全性和效率。

2. 手比行为检测流程与背景介绍

2.1. 手比行为检测的核心问题与挑战

2.1.1. 手比行为的定义与分类

手比行为是列车在列车操作过程中通过特定的手部动作，传递信息或与其他所有权、工作人员进行协调的一种非语言沟通方式。在铁路运输中，尤其是在复杂、混乱的情况下的环境下，手比行为作为重要的沟通信号，助力传递紧急操作指令、调整列车运行状态、调整作业等[6]。司机手比行为可以按照动作方向分为横向手比及纵向手比，横向手比是指通常用于列车停运、车站或方向变更时的指示，纵向手比通常表示启动、加速或停车等重要信号。

2.1.2. 传统检测方法的不足

传统的手比行为检测方法各有其优势和局限性。模板匹配方法简单直观，但对复杂背景和姿势变化适应性差[12]；特征点检测方法能够适应一定的姿势变化，但对精度和实时性要求较高[13]；运动分析方法能够处理动态环境下的手比行为，但在复杂场景中容易发生误判[14]；机器学习方法虽然具有较强的识别能力，但需要大量的标注数据并依赖精细的特征提取[15]。随着深度学习和更先进的计算机视觉技术的不断发展，这些传统方法逐步被更为智能、灵活的算法所替代，以提升手比行为检测的准确性和效率[16]。

2.2. 手比行为检测流程概述

2.2.1. 检测流程的总体框架

图 1 为检测流程的总体框架图：

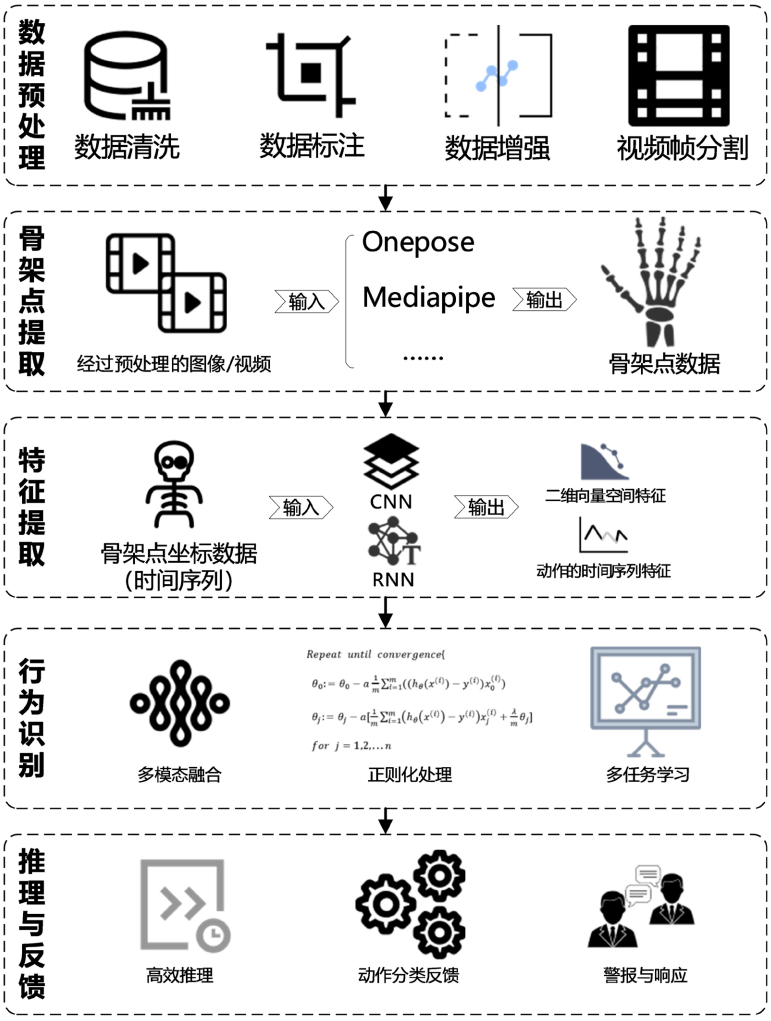

Figure 1. Overall framework of the detection process
图 1. 检测流程的总体框架

2.2.2. 关键技术与挑战

司机手比行为检测的关键技术包括骨架点提取、特征提取和行为识别。其中，骨架点提取利用工具(如 Mediapipe 和 OpenPose)从图像或视频中捕捉手部关键点；特征提取通过卷积神经网络(CNN)获取空间特征，并结合循环神经网络(RNN)或 Transformer 捕捉时间特征；行为识别通过分类模型判断手比动作类型，同时支持实时推理和多任务学习。然而，该领域仍面临数据稀缺、环境干扰(如光线、遮挡)、实时性与精度平衡以及模型鲁棒性不足等挑战[17]。未来需通过构建多样化数据集、优化模型设计和提升环境适应性来进一步推动应用[18]。

3. 基于深度学习的列车司机室手比行为检测

3.1. 骨架点提取算法

在列车司机手比行为的分析中，骨架点提取技术作为核心方法，能够有效捕捉并识别司机在操作过程中的动态行为。近年来，深度学习技术在骨架点提取领域的应用取得了显著进展，尤其是在复杂环境下对列车司机行为的实时监控和识别中，具有重要的实际意义。

骨架点提取方法通常通过深度神经网络(DNN)、卷积神经网络(CNN)以及一些专门的人体姿态估计工具，如 OpenPose 和 MediaPipe，实现对人体动作的精准捕捉。例如，OpenPose 通过多层卷积网络进行人体关节点的检测，能够有效地提取出列车司机手比行为中的关键骨架点，为后续的行为分析提供数据支持。该技术在高速动态环境下仍保持高效，能应对复杂的手比动作。MediaPipe 作为 Google 开发的实时多模态框架，进一步提高了骨架点提取的效率和精度，特别是在复杂的交通环境中，能够实时准确地捕捉司机的动作信息，保障列车操作的安全性。

骨架点提取流程一般包括数据采集、预处理、关节点检测和后续的行为分析。在列车司机手比行为的分析中，首先通过高精度摄像设备获取司机的实时视频数据，经过图像预处理后，利用深度学习模型，如 OpenPose 或 MediaPipe，提取人体的骨架点。提取的骨架点通常包括肩膀、肘部、手腕等重要关节点，能够反映出司机的手比行为[19]。接下来，通过对这些骨架点的时序数据进行进一步分析，可以从中提取出行为特征，以判断操作是否符合安全规范。

为了进一步优化骨架点提取的精度和实时性，研究者们提出了多种策略。例如，结合时序深度学习模型(如 LSTM)来处理骨架点的时间序列数据，能有效提高在快速、复杂环境下的检测精度。此外，使用算法剪枝和量化等技术对模型进行优化，能够提升实时处理能力，从而确保在列车运行过程中，手比行为的识别和确认能够及时完成[20]。

3.2. 数据增强算法

数据增强是深度学习领域中提高模型泛化能力和解决数据稀缺问题的重要手段。在列车司机手比行为的研究中，数据增强技术对骨架点数据的分析和建模起到了重要作用。通过样本扩充、数据归一化等技术，数据增强不仅能够丰富数据集的多样性，还能提升模型在复杂环境中的表现。数据增强技术的实施，包括样本扩充、数据归一化等。

3.2.1. 样本扩充

样本扩充是数据增强的核心方法之一，它通过对现有数据的变换、生成和合成，增加数据的多样性。常见的样本扩充方法包括数据旋转、翻转、裁剪、平移和缩放等。这些技术被广泛用于骨架数据的增强，例如在时间序列骨架数据中，利用数据扩展策略生成不同时间步的骨架点序列，从而提升模型对动态行为的捕捉能力[21]。

此外，生成对抗网络(GAN)等先进技术也被用于样本扩充，通过生成新的骨架点数据，进一步丰富数据集。例如，TorchIO 框架提供了强大的数据扩充和采样功能，使得模型能够更高效地利用扩充数据，优化模型训练效果[22]。

3.2.2. 数据归一化

数据归一化是提高模型稳定性的重要步骤。骨架点数据往往包含坐标和时间序列信息，通过归一化处理，可以将不同范围的特征值转换为统一的尺度，从而减少模型训练中的数值偏差[23]。例如，在 3D 骨架数据中，通过归一化将关节点坐标映射到固定范围，能够减少因数据尺度差异引起的模型性能下降问题[24]。

3.2.3. 骨架点数据增强的效果

骨架点数据增强在提升动作识别模型的性能和鲁棒性方面表现显著。以下是数据增强带来的具体效果：

1) 提升模型的泛化能力

样本扩充和归一化技术能够有效缓解过拟合问题，提高模型对未见数据的预测能力。例如，在对骨架点的动作识别任务中，利用对比学习结合极端数据增强策略，可以增强模型对稀疏数据的适应能力，显著提升模型在复杂环境中的表现[25]。

2) 增强模型对动态行为的捕捉能力

骨架点数据包含丰富的时空信息，通过时间序列数据扩展和归一化技术，能够优化模型对动态行为的检测和识别。例如，利用深度学习模型处理扩展后的骨架点时间序列，可以更好地捕捉手比行为的连续动作，从而提高分类精度。

3) 优化稀疏数据集的表现

在数据稀缺的情况下，数据增强可以有效扩展数据量，降低模型对小样本数据的依赖。例如，通过生成对抗网络生成额外的骨架点样本，能够增强模型在小数据集上的学习能力，提升对列车司机手比行为的识别精度[23]。

3.3. 深度学习模型设计

深度学习模型设计是实现高效特征提取和分类的重要环节。在列车司机行为识别等领域，合理选择与设计模型架构，优化特征提取流程及分类器，是提高识别准确性和鲁棒性的关键。

3.3.1. 模型架构选择与设计

现代深度学习模型的架构多以卷积神经网络(CNN)、YOLO (You Only Look Once)及 Transformer 为核心。每种架构均针对不同的任务特点设计，以实现高效的特征提取和分类能力。

1) YOLO 架构

YOLO 是一种端到端的目标检测模型，其显著特点是将目标检测任务简化为单次回归问题，能够在保证检测精度的同时显著提高实时性能[26]。针对复杂环境中的列车司机行为检测，YOLO 通过 DarkNet 骨干网络提取多尺度特征，并结合优化的分类器完成目标识别。此外，针对动态行为场景，可嵌入 Transformer 模块以增强全局上下文理解能力[27]。

2) CNN 架构

CNN 是深度学习中广泛应用的特征提取工具，其层次化结构能够从图像数据中提取低级和高级特征。在列车司机手势识别任务中，轻量化的 CNN 变种(如 MobileNet)可以降低计算复杂度，同时保留较高的检测精度[28]。此外，通过改进卷积核设计，可以进一步优化模型在复杂背景下的表现。

3) Transformer 架构

Transformer 架构近年来被引入到目标检测和行为识别领域，以其自注意力机制实现了全局特征建模的优势。ViT-YOLO 将 Vision Transformer 嵌入 YOLO 框架，通过多头自注意力模块捕捉复杂行为模式，在动态环境下表现出卓越的识别能力。此外，结合 CNN 的局部特征提取能力，形成混合模型进一步提高了分类性能。

3.3.2. 特征提取、分类器设计及优化策略

1) 特征提取

特征提取是深度学习模型的重要组成部分，其直接决定了模型的识别精度和泛化能力。YOLO 通过多尺度特征融合方法，在检测大尺寸目标的同时确保对小目标的捕捉能力[29]。而 Transformer 通过自注意力机制建模全局上下文信息，进一步提升特征提取效果，适用于列车司机动态行为的复杂特征识别。

2) 分类器设计

分类器是特征提取后的核心任务。改进 YOLO 的分类器可通过引入损失函数优化(如焦点损失和 IoU

损失)，增强对难分类目标的区分能力[30]。Transformer 嵌入的分类模块则通过多头自注意力对输入数据的多维信息进行加权，使得分类结果更具鲁棒性。

3) 优化策略

优化策略是深度学习模型设计的关键步骤。包括以下几个方面：

超参数调整：结合 Adam 或 SGD 优化器调整学习率，提高模型收敛速度和稳定性。

数据增强与正则化：通过数据扩充(旋转、翻转等)和正则化(Dropout、Batch Normalization)策略，提升模型泛化能力。

轻量化设计：对于嵌入式场景的行为识别任务，可以通过网络剪枝和量化技术，降低模型的计算复杂度[31]。

多模态融合：结合视觉、语音等多模态数据的特征，提升模型的识别精度。

现代深度学习模型的架构多以卷积神经网络(CNN)、YOLO (You Only Look Once)及 Transformer 为核心。每种架构均针对不同的任务特点设计，以实现高效的特征提取和分类能力。

4. 基于 Mediapipe 的智轨司机室手比行为检测

随着人工智能和计算机视觉技术的发展，Mediapipe 以其高效的骨架点提取和实时处理能力，在手势检测和行为分析中获得广泛应用[32]。列车司机手比行为检测可以通过结合 Mediapipe 的骨架点提取与深度学习分类器实现高效检测，并通过系统优化提高实时性能。

4.1. 骨架点提取与动作分类

4.1.1. Mediapipe 骨架点提取的基本原理

Mediapipe 使用基于深度学习的关键点检测网络，通过实时计算图像中的骨架点实现高效姿态估计。其内核利用图像金字塔、回归模型以及几何约束方法精确定位手部的关节点，包括指尖、手腕等关键点，形成 2D 或 3D 骨架图。这一特性为列车司机手势的准确捕捉提供了坚实基础。

4.1.2. 动作分类器的设计与实现

Mediapipe 提取的骨架点数据可以被输入到深度学习分类器中进行动作识别。常见的分类器包括：

LSTM：适用于时间序列数据，能够建模手比行为的动态特性[33]。

CNN：用以捕捉骨架点的局部几何关系，对静态手势具有较强分类能力。

Transformer：通过多头自注意力机制捕捉全局上下文信息，适用于复杂手势分类[34]。

分类器设计需要结合列车操作场景，通过引入正则化和优化损失函数提高分类器的性能。

4.2. 实时检测与系统优化

实时检测是列车司机行为监测的核心要求。Mediapipe 通过轻量化模型结构和 GPU 加速技术实现骨架点的实时检测[35]。以下是常见的优化策略：

模型压缩与剪枝：减少参数量以提高推理速度。

快速卷积算法：优化卷积计算以降低时间复杂度。

分布式计算：在多个节点间分担计算负载。

在低计算资源环境中，适配硬件和优化算法是关键。例如，使用 Tiny YOLO 或 MobileNet 等轻量化网络可在嵌入式设备上实现高效推理。进一步结合混合精度计算(FP16 和 FP32)以及硬件加速(如 NVIDIA Jetson Nano)可以在保证性能的同时降低功耗[36]。

5. 仿真实验与分析

5.1. 实验数据

5.1.1. 数据集 1 (基于深度学习的列车司机手比行为数据集)

实验数据集来源于 2017~2020 年轨道列车驾驶室的监控视频图像，本次实验选择其中 9564 张图像作为初始数据集，经过人工筛选，排除其中过曝、无效等图片后，选择了其中的 500 张图像作为本次实验的数据集，其中横向手比 163 张、纵向手比 123 张、未手比 314 张，示例如下图 2~4：

Figure 2. Horizontal hand gesture
图 2. 横向手比

Figure 3. Vertical hand gesture
图 3. 纵向手比

Figure 4. Unmeasured gesture
图 4. 未手比

5.1.2. 数据集 2 (基于 Mediapipe 的列车司机手比行为数据集)

根据智轨驾驶舱的 2 段视频进行了图像分类，从两段视频中获取了 1090 张图像，经人工挑选后，得

到 449 张质量较高图像，分为两手交叉、模拟握持方向盘、停止手势、单手大拇指、双手大拇指、未手比、握持方向盘、左手手比、右手手比 9 种类别。其中，左、右手手比是算法需要识别的项点，对于其他类别，应尽量避免误识别。示范数据如下图 5、图 6：

Figure 5. Left-hand gesture
图 5. 左手比

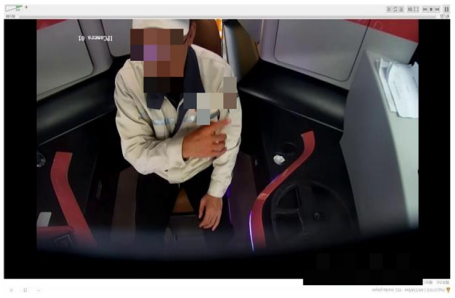

Figure 6. Right-hand gesture
图 6. 右手比

5.2. 基于深度学习的实验分析

5.2.1. 实验环境

本次研究采用 Windows64 位的计算机操作系统，CPU 采用 i5-12400F，GPU 采用 NVIDIA GeForce RTX 1650，具有 4GB 显存，模型训练以 PyTorch2.0 为框架，Python3.9 为编译环境，使用 Cuda11.7.1 进行训练加速。

5.2.2. 数据预处理步骤

图像读取与转换：使用 OpenCV 读取原始图像，转换为 RGB 格式，并统一调整为 224×224 的尺寸。

骨架点提取：利用 MediaPipe Pose 模块提取每张图片中的人体骨骼关键点。

$$\text{Keypoints} = \left\{\left(x_1, y_1, z_1\right), \left(x_2, y_2, z_2\right), \cdots, \left(x_n, y_n, z_n\right)\right\}$$

其中，$\left(x_i, y_i, z_i\right)$ 是从图像中提取出的每个关键点对应于三维坐标，共提取 33 个关键点。在提取出的坐标中，不仅包含二维坐标，还包括了深度 $\left(z-axis\right)$，有助于在三维空间进行手势识别。同时，为了避免关键点坐标受图像分辨率的影响，对提取的坐标进行了标准化处理，使得每个关键点的坐标在 0 到 1 之间。

对于骨架点检测缺失的图像，为了保证其具有一致的形状，因此需要进行骨架点填充，保证数据输入的统一性。具体公式如下：

$$\hat{K} = Pad\left(\left(K, target\_size = (33,3)\right)\right)$$

其中，K 是从图像中提取到的骨架点数据(形状为$[n,3]$，其中 n 是实际提取到的骨架点数量)，$\hat{K}$ 是经过填充后的数据，目标大小为 33 个关键点(每个关键点有三个坐标)，即填充为$(33,3)$，从而确保每个输入样本均具有 99 维(33 个关键点 × 3 个坐标值)。

5.2.3. 数据增强技术

为了增强模型的泛化能力，采用了以下几种数据增强方法：

① 水平翻转：对输入图像进行随机的水平翻转，翻转操作概率为 50%。

$$\text{image} = \text{flip}(\text{image}, \text{axis} = 1)$$

$$\text{with probability} \quad p = 0.5$$

② 随机旋转：对输入图像进行随机旋转，旋转角度为 0°，90°，180°，270°，旋转操作的概率为 50%。

$$\text{image} = \text{rotate}(\text{image}, \theta)$$

$$\text{where } \theta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$$

$$\text{withprobability} \quad p = 0.5$$

③ 随机颜色调整：对输入图像的颜色进行随机调整，包括对比度(α)和亮度(β)的变化，参数范围维：

$$\text{image} = \alpha \cdot \text{image} + \beta$$

$$\text{where } \alpha \in [0.8,\ 1.2] \text{ and } \beta \in [-20,\ 20]$$

这些增强方法有效扩展了数据集的多样性，帮助提高模型的鲁棒性和泛化能力。

5.2.4. 模型设计

本实验采用了一个简单但有效的全连接神经网络(Fully Connected Network, FCN)作为分类模型。模型的设计考虑到输入是人体骨骼的三维坐标，因此在架构上进行了优化：

1) 输入层：输入数据为 33 个关键点的三维坐标，总共 99 维特征。每个输入样本都是一个向量，包含了图像中所有关键点的位置。

2) 隐藏层：

第一个全连接层(fc1)，其输出是经过展开后的关键点数据，输出是一个 128 维的特征向量。其数学表达可以用矩阵乘法进行表示：

假设输入的骨架点数据 X 是一个大小为$(N,99)$的矩阵，其中 N 是批量大小(bitch size)，99 是每个样本的特征维度(33 个关键点 × 3 个坐标值)。因此这层的权重矩阵 W_1 的大小维$(99,128)$，偏置项 b_1 的维度为(128)，输出是一个大小为$(N,128)$的矩阵。具体公式如下：

$$Z_1 = XW_1 + b_1$$

$$\widehat{Z}_1 = ReLU(Z_1)$$

其中：

$X \in \mathbb{R}^{N\times 99}$ 是输入数据矩阵；

$W_1 \in \mathbb{R}^{99\times 128}$ 是第一层的权重矩阵；

$b_1 \in \mathbb{R}^{128}$ 是第一层的偏置项；

$Z_1 \in \mathbb{R}^{N\times 128}$ 是线性变换的结果；

$\widehat{Z}_1 \in \mathbb{R}^{N\times 128}$ 是经过 $ReLU$ 激活函数处理后的输出。

这样，第一层的输出就是一个 128 维的特征向量，作为后续网络层的输入。

第二个全连接层(fc2)：将 128 维的特征映射为 64 维，同样使用 *ReLU* 激活函数。因此，第二层计算公式可以细化为：

$$Z_2 = XW_2 + b_2$$

其中：

$X \in \mathbb{R}^{N\times 128}$ 是输入数据矩阵；

$W_2 \in \mathbb{R}^{128\times 64}$ 是第一层的权重矩阵；

$b_2 \in \mathbb{R}^{64}$ 是第一层的偏置项；

$Z_2 \in \mathbb{R}^{N\times 64}$ 是线性变换的结果。

使用 *ReLU* 激活函数处理后：

$$\widehat{Z_2} = ReLU\left(Z_2\right)$$

3) 输出层：

输出层的作用是根据模型的计算结果进行分类，在本项目中，输出层采用 3 个神经元，分别对应三类手势(普通手势、向上指、水平指)。输出层(fc3)接受前一层(fc2)的 64 维输入并输出一个大小为 $\text{len}\left(\text{label\_mapping}\right)$ 的向量。这个向量的每个元素表示对应类别的预测值。具体公式如下：

$$y = W_3 \cdot x_2 + b_3$$

其中：

$y \in \mathbb{R}^C$ 是类别分数向量，表示模型对每个类别的预测值(得分)；

$W_3 \in \mathbb{R}^{C\times 64}$ 是权重矩阵，表示每个类别与前一层 64 个神经元之间的连接权重；

$x_2 \in \mathbb{R}^{64}$ 是来自 fc2 层的输入向量，表示从隐藏层中传递下来的特征信息；

$b_3 \in \mathbb{R}^C$ 是偏置项，通常用于调节每个类别的得分。

随后，通过 Softmax 激活函数将输出得分转换为类别概率：

$$P\left(y = c\middle| x\right) = \frac{e^{y_c}}{\sum_{i=1}^{C} e^{y_i}}$$

其中，y_c 是输出向量 y 中第 c 个元素，表示类别 c 的得分，C 为类别总数。

输出采用 Softmax 激活函数，得到每个类别的预测概率。

4) 损失函数与优化器：

本实验的手势分类模型使用了交叉熵损失函数(Cross-Entropy Loss)和 Adam 优化器，适合处理多分类任务并且能够有效加速模型训练。在本实验中，通过标签映射(label_mapping)实现了清晰的类标签转换，将手势类别与具体数字标签一一对应，确保训练时的标签一致性。具体公式如下：

$$L = -\sum_{i=1}^{C} y_i \log\left(p_i\right)$$

其中，C 是类别数，y_i 是实际标签，p_i 是模型预测的类别概率。通过标签映射的方式，可以将类别转换为数值标签，从而在实际应用中快速调整和适应不同的手势识别任务。

5.2.5. 实验设置与训练过程

1) 训练过程的损失与准确率：在 1000 轮的训练过程中，模型的训练损失逐渐下降，验证集的准确率逐步提高。训练损失在第 500 轮后开始显著减小。图 7 展示了训练损失的变化趋势，表明模型在训练过程中得到了良好的优化。

2) 验证集上的分类性能：在验证集上，模型的总体准确率达到 88.15%，对于每个类别的分类准确率分别为：普通手势 90%、向上指 88%、水平指 86%。这一结果表明，模型能够有效区分不同类别的手势。同时根据模型在各类手势上的预测情况，其中大部分错误集中在“向上指”与“水平指”之间，表明这两类手势在姿态上有一定的相似性。

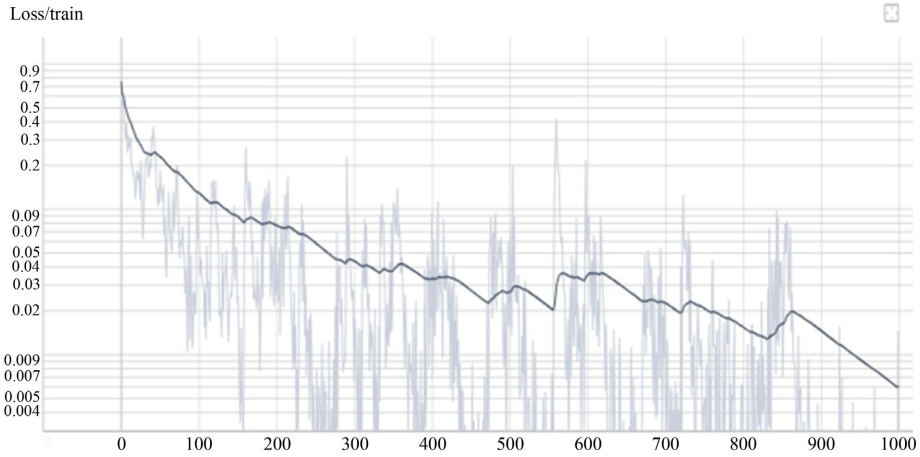

Figure 7. Loss function curve
图 7. 损失函数曲线

3) 鲁棒性与泛化能力：模型在不同环境下表现出较强的鲁棒性。在测试时，我们使用了不同背景、光照和视角的图像，模型在这些变换条件下仍然保持较高的准确率。这表明基于骨骼关键点的手势识别方法对于环境变化具有较强的适应能力。

5.3. 基于 Mediapipe 的实验分析

5.3.1. 参数配置

手势检测模型：使用 MediaPipe Hands 模型，最大检测手数为 2，设定置信度阈值为 0.5，确保手势识别的准确性与处理效率平衡。

姿势检测模型：使用 MediaPipe Pose，设置为模式 2 以提高精度，适合高精度的人体姿势分析。

几何阈值设定：在判断手势时，通过手部关键点与其它身体关键点(如肩部、肘部)的距离关系来确定手势类型，距离阈值设定为 15~25 像素。

5.3.2. 数据预处理

图像数据通过 OpenCV 读取，并转化为适合 MediaPipe 处理的 RGB 格式。在每一帧图像中，首先通过 MediaPipe Hands 模块检测手部关键点，接着使用 MediaPipe Pose 模块获取人体姿态关键点数据。这些关键点为后续手势分类和手臂检测提供了重要的空间信息。

5.3.3. 凸包算法

手势分类基于手部关键点的几何关系，通过计算手指端点与手部轮廓的相对位置来进行判定。具体方法如下：

1) 凸包的定义和数学框架

设 $\mathrm{P}=\{p_1,p_2,p_3,\cdots,p_n\}$ 为二维平面上 n 个点的集合，其中每个点 $p_i=(x_i,y_i)\in\mathbb{R}^2$，则点集 P 是所有点的最小凸多边形。形式化地，凸包 $\mathrm{H}(\mathrm{P})$ 为满足以下条件的点集：

$$\mathrm{H}(\mathrm{P})=\left\{p\in\mathbb{R}^2 \mid p=\sum_{i=1}^{n}\lambda_i p_i,\sum_{i=1}^{n}\lambda_i=1,\lambda_i\geq 0,\forall i\in\{1,2,\cdots,n\}\right\}$$

其中，λ_i 为权重系数，且满足 $\lambda_i\geq 0$ 且 $\sum_{i=1}^{n}\lambda_i=1$，即这些点通过凸组合的方式线性构造，且该构造保证了包含所有点集的最小凸多边形。

2) 凸包算法的实现

设 $\mathrm{F}=\{f_1,f_2,\cdots,f_m\}$ 为手势中 m 个关键点的集合，凸包 $\mathrm{H}(\mathrm{F})$ 可以用于判断这些关键点是否构成一个闭合的手势轮廓。通过计算这些关键点的凸包，能够有效地识别手势的类别及其内部结构。例如，在判断是否为“掌”或“指”手势时，可以依据关键点相对凸包的位置关系进行分类。

具体而言，假设 $f_i\in F$ 为手势关键点，利用点与凸包的距离测试，可以判定某点是否位于凸包外部。假设 $\mathrm{dist}(f_i,H)$ 表示点 f_i 到凸包 H 的距离，则根据点是否位于凸包外部，我们可以建立如下分类准则：

掌手势分类准则：若绝大多数关键点 $f_1,f_2,\cdots,f_m$ 位于凸包外部(即 $\mathrm{dist}(f_i,H)<0$)，则该手势可以判定为“掌”手势。形式化地，若 k 个关键点满足 $\mathrm{dist}(f_i,H)<0$，则判定为掌手势的条件为：

掌手势 if $\sum_{i=1}^{m}1_{dist(f_i,H)<0}\geq\theta$。

指手势：如果只有少数几个关键点位于凸包外部，则可以识别为指手势，分类条件为：

指手势 if $1\leq\sum_{i=1}^{m}1_{dist(f_i,H)<0}<\theta$。

无手势：若所有关键点均位于凸包内，则识别为无手势状态：

无手势 if $\sum_{i=1}^{m}1_{dist(f_i,H)<0}=0$。

5.3.4. 手臂抬起检测

手臂抬起检测的关键在于分析肩膀、肘部和手腕的相对位置。具体方法如下：

1) 通过计算肘部到肩膀的垂直距离来判断手臂是否抬起。

2) 若垂直距离小于预设阈值，则判定该手臂为抬起状态。

对于右臂和左臂，分别计算如下：

$$d_{right}=\left|y_{right\ wrist}-y_{right\ elbow}\right|$$

$$d_{left}=\left|y_{left\ wrist}-y_{left\ elbow}\right|$$

若 d_{right} 或者 d_{left} 小于阈值，则判定为右手或左手抬起。

5.3.5. 综合判断

在综合判断阶段，首先通 findDis 函数计算手腕到肩膀的距离，用于判断是左手还是右手。然后，结合抬手检测结果，输出相应的手势识别结果。最终，系统根据手势类型与抬手状态进行输出，若识别到有效的手势，则标注手势信息，并展示在图像上，如下图 8、图 9 所示：

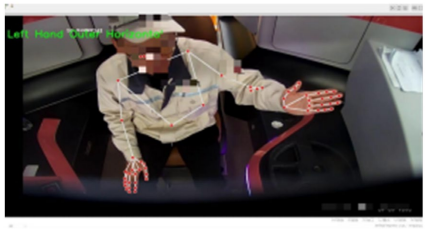

Figure 8. Left hand gesture (after recognition)
图 8. 左手比(识别后)

Figure 9. Right hand gesture (after recognition)
图 9. 右手比(识别后)

在实验中，我们使用了来自不同来源的图像数据集进行测试。每一张图像经过预处理后，输入到手势分类与抬手检测模块。表 1 为实验的输出结果：

Table 1. Hand ratio recognition test results
表 1. 手比识别检测结果

动作分类		实际数量	模糊数量	无效手比	遮挡数量	有效样本数量	检出数量	有效样本检出率
手比	左手	88	17	10	0	61	43	70%
	右手	78	14	6	43	15	14	93%
	综合	166	31	16	43	76	57	75%
行驶干扰动作		88	0	0	0	88	88	100%
未手比		192	0	0	0	192	192	100%
合计		446	0	0	0	356	337	95%

5.3.6. 结果讨论

实验结果表明，系统能够较准确地识别并分类常见的手势类型，如掌手比和指手比。此外，系统对手臂的抬起状态也能够做出较为准确的判断。然而，系统在复杂环境下(如多手重叠、快速运动等)可能会出现一定的误差，未来可以通过优化检测算法和引入更多的训练数据来进一步提升系统的鲁棒性和精度。

6. 结语

本研究提出的基于深度学习和 Mediapipe 的列车司机手比行为检测方法，成功地提升了手比行为的检测精度和实时性。实验结果表明，使用 ResNet50 卷积神经网络，手比行为的分类准确率超过 85%；通过 Mediapipe 框架的实时手部关键点检测，准确率进一步提高至 90%。这些结果证明了深度学习和计算机视觉技术在复杂铁路环境下的有效性，尤其在提升驾驶安全性和操作精度方面具有重要意义。

然而，尽管本研究取得了较好的实验效果，仍有进一步研究的空间。未来的研究可着重于以下几个方面：首先，扩展数据集，涵盖更多复杂的驾驶环境和不同类型的手比行为，以进一步提升系统的适应性；其次，优化模型的实时性，特别是在计算资源有限的场景下；最后，结合更多的传感器数据和多模态信息，进一步提高手比行为识别的准确性和鲁棒性，为智能驾驶舱的安全性提供更强有力的技术支持。

参考文献

[1] Wang, Q., Zhu, F., Dang, R., Wei, X., Han, G., Huang, J., *et al.* (2023) An Eye Tracking Investigation of Attention Mechanism in Driving Behavior under Emotional Issues and Cognitive Load. *Scientific Reports*, **13**, Article No. 16963. https://doi.org/10.1038/s41598-023-43693-8

[2] Xu, J., Fard, M., Zhang, N., Davy, J.L. and Robinson, S.R. (2024) Cognitive Load and Task Switching in Drivers: Implications for Road Safety in Semi-Autonomous Vehicles. *Transportation Research Part F*: *Traffic Psychology and Behaviour*, **107**, 1175-1197. https://doi.org/10.1016/j.trf.2024.11.005

[3] Filtness, A.J. and Naweed, A. (2017) Causes, Consequences and Countermeasures to Driver Fatigue in the Rail Industry: The Train Driver Perspective. *Applied Ergonomics*, **60**, 12-21. https://doi.org/10.1016/j.apergo.2016.10.009

[4] Tichon, J., Wallis, G. and Mildred, T. (2006) Virtual Training Environments to Improve Train Driver's Crisis Decision Making. *SimTecT* 2006 *Conference and Exhibition*, Melbourne, 29 May-1 June 2006.

[5] Cogan, B. and Milius, B. (2023) Remote Control Concept for Automated Trains as a Fallback System: Needs and Preferences of Future Operators. *Smart and Resilient Transportation*, **5**, 50-69. https://doi.org/10.1108/srt-11-2022-0018

[6] Xu, J., Tang, Z., Zhao, H. and Zhang, J. (2019) Hand Gesture-Based Virtual Reality Training Simulator for Collaboration Rescue of a Railway Accident. *Interacting with Computers*, **31**, 577-588. https://doi.org/10.1093/iwc/iwz037

[7] Chang, C., Chang, C. and Lin, Y. (2022) A Hybrid CNN and LSTM-Based Deep Learning Model for Abnormal Behavior Detection. *Multimedia Tools and Applications*, **81**, 11825-11843. https://doi.org/10.1007/s11042-021-11887-9

[8] Singh, A.K., Kumbhare, V.A. and Arthi, K. (2022) Real-Time Human Pose Detection and Recognition Using Mediapipe. In: Reddy, V.S., Prasad, V.K., Wang, J. and Reddy, K., Eds., *Soft Computing and Signal Processing*, Springer, 145-154. https://doi.org/10.1007/978-981-16-7088-6_12

[9] Watson, E., Viana, T. and Zhang, S. (2024) Machine Learning Driven Developments in Behavioral Annotation: A Recent Historical Review. *International Journal of Social Robotics*, **16**, 1605-1618. https://doi.org/10.1007/s12369-024-01117-1

[10] Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z. and Liu, Y. (2021) Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Computing Surveys*, **54**, 1-40. https://doi.org/10.1145/3447744

[11] Rahim, M.A., Miah, A.S.M., Akash, H.S., Shin, J., Hossain, M.I. and Hossain, M.N. (2024) An Advanced Deep Learning Based Three-Stream Hybrid Model for Dynamic Hand Gesture Recognition. arXiv: 2408.08035. https://doi.org/10.48550/arXiv.2408.08035

[12] Mahbub, U., Imtiaz, H., Roy, T., Rahman, M.S. and Rahman Ahad, M.A. (2013) A Template Matching Approach of One-Shot-Learning Gesture Recognition. *Pattern Recognition Letters*, **34**, 1780-1788. https://doi.org/10.1016/j.patrec.2012.09.014

[13] Oudah, M., Al-Naji, A. and Chahl, J. (2020) Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging*, **6**, Article 73. https://doi.org/10.3390/jimaging6080073

[14] Ramamoorthy, A., Vaswani, N., Chaudhury, S. and Banerjee, S. (2003) Recognition of Dynamic Hand Gestures. *Pattern Recognition*, **36**, 2069-2081. https://doi.org/10.1016/s0031-3203(03)00042-6

[15] Escalera, S., Athitsos, V. and Guyon, I. (2017) Challenges in Multi-Modal Gesture Recognition. In: Escalera, S., Guyon, I. and Athitsos, V., Eds., *Gesture Recognition*, Springer, 1-60. https://doi.org/10.1007/978-3-319-57021-1_1

[16] Kaur, A. and Bansal, S. (2022) Deep Learning for Dynamic Hand Gesture Recognition: Applications, Challenges and Future Scope. 2022 5*th International Conference on Multimedia*, *Signal Processing and Communication Technologies* (*IMPACT*), Aligarh, 26-27 November 2022, 1-6. https://doi.org/10.1109/impact55510.2022.10029100

[17] Ay, Ö. and Emel, E. (2025) Real-time Assembly Task Validation Using Deep Learning-Based Object Detection and Operator's Hand-Joints Trajectory Classification. *IEEE Access*, **13**, 57009-57029. https://doi.org/10.1109/access.2025.3554263

[18] Patel, M., Rao, S., Chauhan, S. and Kumar, B. (2024) Real-Time Hand Gesture Recognition Using Python and Web Application. 2024 1*st International Conference on Advances in Computing*, *Communication and Networking* (*ICAC2N*), Greater Noida, 16-17 December 2024, 564-570. https://doi.org/10.1109/icac2n63387.2024.10895151

[19] He, L. and Zhang, J. (2021) Railway Driver Behavior Recognition System Based on Deep Learning Algorithm. 2021 4t*h International Conference on Artificial Intelligence and Big Data* (*ICAIBD*), Chengdu, 28-31 May 2021, 398-403. https://doi.org/10.1109/icaibd51990.2021.9458983

[20] Liu, C. and Szirányi, T. (2021) Real-time Human Detection and Gesture Recognition for On-Board UAV Rescue. *Sensors*, **21**, Article 2180. https://doi.org/10.3390/s21062180

[21] Park, S., Kwon, H., Baek, J. and Chung, K. (2022) Dimensional Expansion and Time-Series Data Augmentation Policy for Skeleton-Based Pose Estimation. *IEEE Access*, **10**, 112261-112272. https://doi.org/10.1109/access.2022.3214659

[22] Pérez-García, F., Sparks, R. and Ourselin, S. (2021) Torchio: A Python Library for Efficient Loading, Preprocessing, Augmentation and Patch-Based Sampling of Medical Images in Deep Learning. *Computer Methods and Programs in Biomedicine*, **208**, Article ID: 106236. https://doi.org/10.1016/j.cmpb.2021.106236

[23] Hernández-García, A. and König, P. (2018) Data Augmentation Instead of Explicit Regularization. arXiv: 1806.03852. https://doi.org/10.48550/arXiv.1806.03852

[24] Xin, C., Kim, S., Cho, Y. and Park, K.S. (2024) Enhancing Human Action Recognition with 3D Skeleton Data: A Comprehensive Study of Deep Learning and Data Augmentation. *Electronics*, **13**, Article 747. https://doi.org/10.3390/electronics13040747

[25] Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T. and Ding, R. (2022) Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-Supervised Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 762-770. https://doi.org/10.1609/aaai.v36i1.19957

[26] Poon, Y., Kao, C., Wang, Y., Hsiao, C., Hung, M., Wang, Y., *et al*. (2021) Driver Distracted Behavior Detection Technology with Yolo-Based Deep Learning Networks. 2021 *IEEE International Symposium on Product Compliance Engineering—Asia* (*ISPCE-ASIA*), 30 November-1 December 2021, 1-5. https://doi.org/10.1109/ispce-asia53453.2021.9652435

[27] Zhang, Z., Lu, X., Cao, G., Yang, Y., Jiao, L. and Liu, F. (2021) ViT-YOLO: Transformer-Based YOLO for Object Detection. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 2799-2808. https://doi.org/10.1109/iccvw54120.2021.00314

[28] Abbass, M.A.B. and Ban, Y. (2024) MobileNet-Based Architecture for Distracted Human Driver Detection of Autonomous Cars. *Electronics*, **13**, Article 365. https://doi.org/10.3390/electronics13020365

[29] Guo, K., Li, X., Zhang, M., Bao, Q. and Yang, M. (2021) Real-Time Vehicle Object Detection Method Based on Multi-Scale Feature Fusion. *IEEE Access*, **9**, 115126-115134. https://doi.org/10.1109/access.2021.3104849

[30] Steno, P., Alsadoon, A., Prasad, P.W.C., Al-Dala'in, T. and Alsadoon, O.H. (2020) A Novel Enhanced Region Proposal Network and Modified Loss Function: Threat Object Detection in Secure Screening Using Deep Learning. *The Journal of Supercomputing*, **77**, 3840-3869. https://doi.org/10.1007/s11227-020-03418-4

[31] Shi, J., Bian, J., Richter, J., Chen, K., Rahnenführer, J., Xiong, H., *et al*. (2021) MODES: Model-Based Optimization on Distributed Embedded Systems. *Machine Learning*, **110**, 1527-1547. https://doi.org/10.1007/s10994-021-06014-6

[32] Fan, Y. (2024) The Gesture Recognition Improvement of Mediapipe Model Based on Historical Trajectory Assist Tracking, Kalman Filtering and Smooth Filtering. *Proceedings of the* 2024 7*th International Conference on Computer Information Science and Artificial Intelligence*, Shaoxing, 13-15 September 2024, 641-647. https://doi.org/10.1145/3703187.3703295

[33] Xing, Y., Lv, C., Cao, D. and Lu, C. (2020) Energy Oriented Driving Behavior Analysis and Personalized Prediction of Vehicle States with Joint Time Series Modeling. *Applied Energy*, **261**, Article ID: 114471. https://doi.org/10.1016/j.apenergy.2019.114471

[34] Tang, Y., Pan, M., Li, H. and Cao, X. (2024) A Convolutional-Transformer-Based Approach for Dynamic Gesture Recognition of Data Gloves. *IEEE Transactions on Instrumentation and Measurement*, **73**, 1-13. https://doi.org/10.1109/tim.2024.3400361

[35] Ong, A.J.S., Cabatuan, M., Tiberio, J.L.L. and Jose, J.A. (2022) LSTM-Based Traffic Gesture Recognition Using MediaPipe Pose. *TENCON* 2022—2022 *IEEE Region* 10 *Conference* (*TENCON*), Hong Kong, 1-4 November 2022, 1-5. https://doi.org/10.1109/tencon55691.2022.9977857

[36] Ma, J., Chen, L. and Gao, Z. (2018) Hardware Implementation and Optimization of Tiny-Yolo Network. In: Zhai, G., Zhou, J. and Yang, X., Eds., *Digital TV and Wireless Multimedia Communication*, Springer, 224-234. https://doi.org/10.1007/978-981-10-8108-8_21