State of the Art in Real-Time Human Pose Estimation for Guided Multimodal Human-Robot Interaction

Internship Project Report

Khalid Zouhair

Host University: University of Northern British Columbia (UNBC)

Department of Computer Science and Engineering

Supervisor: Dr. Shruti Chandra

I. Introduction

Human pose estimation (HPE) is a core area in computer vision that involves detecting and interpreting human body joints and posture from visual inputs. In recent years, HPE has become increasingly significant in the field of socially assistive robotics, where robots are required to understand, respond to, and interact with humans in a natural and context-aware manner.

This report focuses on the application of real-time HPE in guided multimodal human-robot interaction systems, particularly in scenarios involving physical exercise coaching and rehabilitation. The goal of such systems is to provide users with interactive support using a combination of vision, speech, and physiological monitoring. As part of an internship project at the University of Northern British Columbia (UNBC), this study contributes to the development of a human-robot interaction (HRI) platform where a social robot analyzes users' body posture in real time, interprets their actions, and delivers feedback in a responsive and meaningful way.

The primary aim of this review is to analyze the current state of the art in real-time pose estimation frameworks that are suitable for embedded deployment on robots such as QTrobot RD-V2, equipped with NVIDIA Jetson AGX Orin. The focus is on identifying methods that balance speed, accuracy, and hardware efficiency, and which have proven effective in applications like physical therapy, elder care, and fitness coaching. Furthermore, the review considers how these methods can be integrated with speech processing and physiological data interpretation to support truly multimodal robot behavior.

This document begins with an overview of pose estimation approaches, followed by a comparative analysis of relevant models based on both technical literature and recent experimental deployments. The goal is to inform the selection and implementation of a robust, real-time pose estimation system as a core component of the larger multimodal HRI platform under development.

II. Background and State of the Art

Human pose estimation (HPE) has developed rapidly over the past decade, moving from classical computer vision techniques (e.g., background subtraction and contour detection) to deep learning-based models capable of robust, real-time performance. This evolution has enabled new applications in human-robot interaction, especially in contexts like rehabilitation, fitness coaching, and social robotics.

Early models struggled with occlusions, variations in body shape, and real-time constraints. However, models such as OpenPose [2] and AlphaPose [3] demonstrated that multi-person 2D pose estimation could be achieved with high accuracy using convolutional neural networks. These approaches became foundational for subsequent real-time systems.

More recently, lightweight and mobile-optimized models such as MediaPipe [1] and MoveNet [5] have gained traction due to their ability to run on edge devices like smartphones and Jetson boards. These frameworks prioritize inference speed and hardware efficiency, making them suitable for embedded applications such as the QTrobot.

The current state of the art includes:

- **MediaPipe (Google)** Fast, cross-platform, and optimized for edge deployment. Supports 33 landmarks and has been integrated into HRI systems [1, 7, 8].
- **YOLO-Pose (Ultralytics)** Combines object detection and keypoint estimation, suitable for ONNX and Jetson export [4].
- **MoveNet (Google)** Extremely fast with mobile deployment support, albeit with fewer landmarks [5].
- **HRNet** High-accuracy deep model with strong benchmark results, though not yet optimized for real-time robotic applications [6].

In the context of guided multimodal HRI, current research is also exploring the integration of pose data with other inputs such as speech and physiological signals. Two recent studies [7, 8] have successfully demonstrated how pose estimation combined with social robots like NAO and Pepper can be used to provide adaptive, real-time feedback in physical therapy scenarios. These works highlight the importance of combining vision with other modalities to deliver more intelligent and personalized robot behavior.

III. Identified Research Gap

Despite the significant progress in human pose estimation and its integration into socially assistive robotics, there remains a notable gap between low-level pose detection and high-level movement understanding. Most existing systems are limited to identifying joint positions without interpreting them in the context of physical activity or therapeutic guidance. Additionally, while frameworks like MediaPipe and MoveNet offer fast inference, their deployment in embedded robotics platforms, such as the Jetson AGX Orin-powered QTrobot, is still underexplored particularly for real-time multimodal interaction. Moreover, current literature lacks standardized methods for recognizing specific physical gestures (e.g., squats, arm raises) and translating them into personalized, adaptive robot responses. Addressing this gap is essential for developing intelligent, embedded systems that not only detect human motion, but also interpret and respond to it meaningfully in physical guidance scenarios.

IV. Overview of Pose Estimation Approaches

Human pose estimation (HPE) refers to the process of detecting key skeletal points of the human body from images or video, typically including joints like elbows, knees, and shoulders. It plays a critical role in human-robot interaction (HRI), particularly for systems that provide feedback during physical tasks such as rehabilitation or fitness coaching.

Pose estimation techniques can be grouped into two main categories: **2D pose estimation**, where joint locations are predicted on the image plane, and **3D pose estimation**, which infers depth information to reconstruct joint positions in space. For embedded, real-time applications, 2D models are typically favored due to lower computational costs [1].

Modern HPE systems rely heavily on deep learning and large annotated datasets. The most widely used frameworks include:

- **MediaPipe Pose (Google)** A lightweight model using BlazePose architecture, capable of real-time inference on mobile and embedded hardware. It outputs 33 landmarks and supports upper and lower body tracking [1].
- **OpenPose (CMU)** A landmark framework that introduced Part Affinity Fields (PAFs) for multi-person 2D estimation. Known for high accuracy but has relatively high GPU requirements [2].
- **AlphaPose** Designed for accuracy, not speed. It achieves high precision on COCO and MPII benchmarks, but its large architecture limits real-time performance [3].
- **YOLO-Pose** A fast and compact model combining object detection with keypoint estimation. Built on the YOLOv8 framework and optimized for edge inference using ONNX [4].
- **MoveNet (Google)** A TensorFlow Lite model providing fast and accurate body tracking for mobile and embedded systems. Offers Lightning (speed) and Thunder (accuracy) variants [5].
- **HRNet** Maintains high-resolution features throughout the network, achieving strong benchmark accuracy. However, it is not optimized for real-time or edge deployment [6].

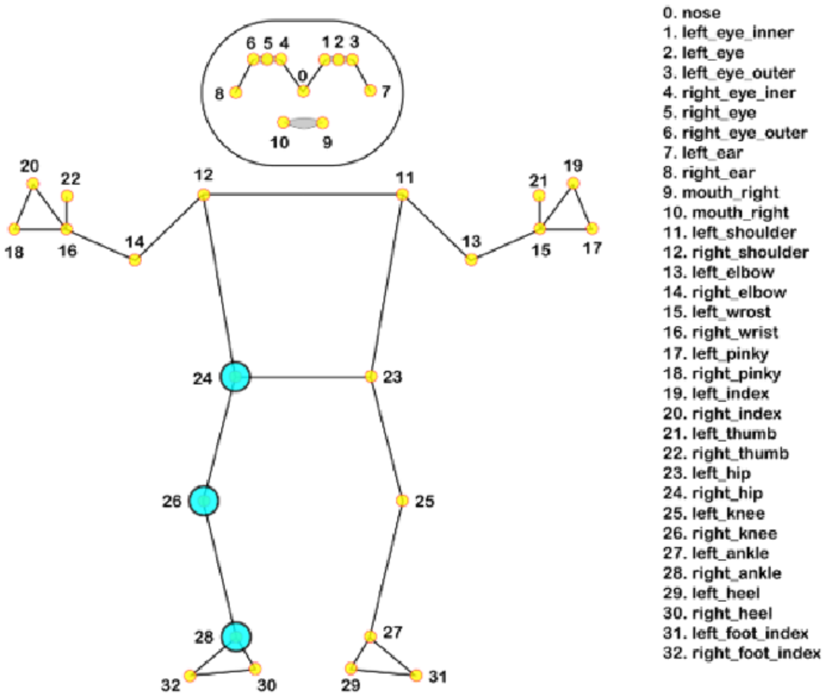

Figure 1: Illustration of 33 pose landmarks in MediaPipe Pose. Source: [1]

Given the hardware used in this project Jetson AGX Orin and Raspberry Pi — the focus is on 2D models that:

- Can run at >15 FPS on embedded GPUs.
- Provide at least upper-body landmarks for exercise tracking.
- Are compatible with Python and ROS (or easy to wrap).
- Have been used in HRI or physical feedback systems.

V. Evaluation Criteria

To choose the most suitable pose estimation framework for this project, I defined a set of evaluation criteria based on technical requirements, hardware limitations, and use case relevance. These criteria are grounded in recent literature on embedded vision systems and real-time robotics [1, 4, 5].

1. **Inference Speed (FPS)** The system must operate in real time to support interactive robot feedback. A frame rate of 15–30 FPS is generally considered the lower bound for smooth perception-action loops in HRI systems [1].

2. **Jetson AGX Orin Compatibility** Since QTrobot uses Jetson AGX Orin, the model must support GPU acceleration via CUDA or TensorRT. Models that rely only on x86-64 desktops or large server GPUs are not practical for this setup.

3. **ROS Integration or Python Compatibility** The robot's middleware is ROS-based. Models with existing ROS nodes (like OpenPose), or with simple Python APIs that can be wrapped into ROS nodes (e.g., MediaPipe), are preferred.

4. **Landmark Coverage** The number and type of detected body landmarks influence how well the robot can recognize exercise movements. Systems like MediaPipe offer 33 landmarks, including fingers, feet, and face which can be useful for future extensions [1].

5. **Accuracy vs. Efficiency Tradeoff** A highly accurate model that can't run in real time is less useful than a slightly less accurate model that responds immediately. A balance between detection quality and low latency is essential for user experience [6].

6. **Prior Use in Physical Coaching or HRI Systems** Preference is given to models that have been tested in real-time feedback systems like SARs or rehabilitation robots since these applications share similar constraints and user expectations [7, 8].

7. **Deployment Flexibility and Support** Actively maintained open-source models with ONNX export, TensorFlow Lite versions, or pre-built SDKs allow faster prototyping and smoother integration. Community support also helps in resolving bugs and improving deployment success rates.

These criteria will guide the comparison of pose estimation models in the next section, helping identify which ones are best suited for real-time, embedded, multimodal robot interaction.

VI. Comparative Analysis of Pose Estimation Models

This section compares six leading pose estimation models based on the evaluation criteria outlined earlier. The focus is on real-time performance, hardware compatibility, integration ease, and suitability for human-robot interaction (HRI), especially in exercise or rehabilitation contexts.

1. MediaPipe Pose MediaPipe is a lightweight framework developed by Google for real-time perception tasks. It uses BlazePose to track 33 landmarks and runs efficiently on embedded GPUs like Jetson AGX Orin. Studies show it achieves 30+ FPS on Jetson-class hardware [1]. It was successfully used in robot-guided therapy sessions with the NAO robot [7].

2. OpenPose OpenPose introduced Part Affinity Fields (PAFs) for multi-person 2D pose estimation [2]. It offers good accuracy and native ROS support, but is GPU-intensive. On Jetson, it runs slowly unless heavily optimized, typically under 10 FPS [9].

3. AlphaPose AlphaPose provides high accuracy on benchmark datasets like COCO [3], but is too slow for real-time use on embedded hardware without quantization. It lacks ROS support and is mainly suited for offline applications.

4. YOLO-Pose This newer model builds on YOLOv8 and combines detection and pose estimation. It supports ONNX export and performs well on Jetson boards, reaching 20–30 FPS depending on resolution [4]. However, it lacks detailed anatomical landmarks and needs custom ROS wrappers.

5. MoveNet MoveNet is an ultra-fast TensorFlow Lite model optimized for edge devices [5]. The Lightning version exceeds 50 FPS on Jetson hardware. However, it only tracks 17 landmarks and has limited real-world use in HRI systems so far.

6. HRNet HRNet maintains high-resolution features throughout its network [6], offering strong accuracy but poor speed. It is not suited for Jetson deployment or real-time robotics without major simplifications.

Model Comparison Table

Model	**FPS (Jetson)**	**ROS Support**	**Landmarks**	**Used in HRI**	**Fit**
MediaPipe	30+	Indirect	33	Yes	**High**
OpenPose	5–10	Yes	18	Yes	Medium
AlphaPose	¡10	No	17	No	Low
YOLO-Pose	20–30	Custom	17	No	Medium–High
MoveNet	50+	Custom	17	Limited	Medium
HRNet	¡5	No	17–32	No	Low

Summary MediaPipe is the most practical choice for this project. It combines fast inference, good body coverage, and successful use in rehab-focused robotics. YOLO-Pose and MoveNet are promising alternatives, especially when performance is prioritized over anatomical detail. OpenPose remains useful for prototyping but requires GPU tuning. AlphaPose and HRNet are not suitable for real-time embedded robotics.

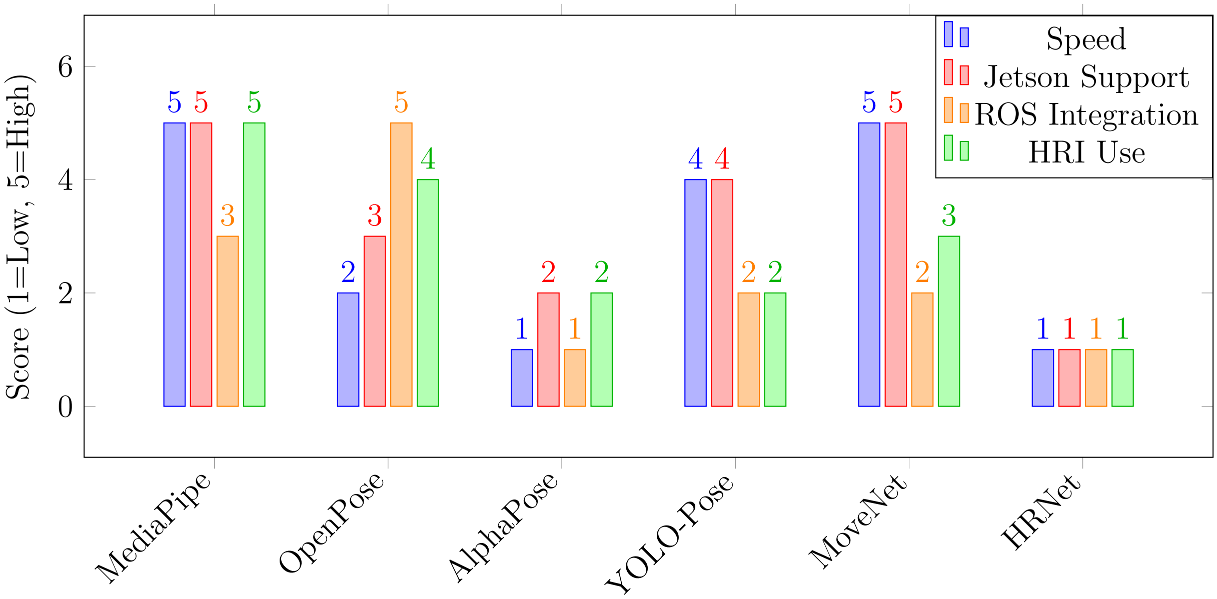

Figure 2: Comparison of models across four key evaluation dimensions.

VII. Conclusion

In this report, I explored the current state of real-time human pose estimation technologies with a focus on their applicability to embedded, multimodal human-robot interaction systems specifically for physical activity monitoring using the QTrobot platform.

The review compared six major frameworks: MediaPipe, OpenPose, AlphaPose, YOLO-Pose, MoveNet, and HRNet. The evaluation was based on key criteria such as real-time performance, Jetson AGX Orin compatibility, ROS integration, body landmark coverage, and previous use in HRI or rehabilitation settings.

Based on the analysis, **MediaPipe** stands out as the most suitable framework. It offers high frame rates, detailed full-body tracking, and proven performance in similar therapeutic and interactive systems. **YOLO-Pose** and **MoveNet** are also promising, especially for future iterations where lightweight models or faster response times may be prioritized.

This study has helped me understand how to critically compare pose estimation models not only in terms of accuracy but also in terms of deployability, responsiveness, and ease of integration into real-time robotic systems. These insights will directly guide the development and testing of the perception module in the upcoming phases of this internship project.

References

1. Lugaresi, C., et al. (2019). *MediaPipe: A Framework for Building Perception Pipelines.* arXiv preprint arXiv:1906.08172.

2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). *Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields.* Proceedings of the IEEE CVPR.

3. Fang, H.-S., Xie, S., Tai, Y.-W., & Lu, C. (2017). *RMPE: Regional Multi-person Pose Estimation.* Proceedings of the IEEE ICCV.

4. Ultralytics. (2023). *YOLOv8-Pose Documentation.* Available at: `https://docs.ultralytics.com/tasks/pose/`

5. Google AI Blog. (2021). *Pose Detection with MoveNet.* Available at: `https://blog.tensorflow.org/2021/05/pose-detection-with-movenet.html`

6. Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). *Deep High-Resolution Representation Learning for Human Pose Estimation.* Proceedings of the IEEE CVPR.

7. Pillai, D., & Abdulaziz, N. (2024). *Development and Evaluation of NAO Humanoid Robot for Rehabilitation Therapy.* Proceedings of the 2024 ICSPIS Conference, IEEE.

8. Sardinha, E. N., Sarajchi, M., Xu, K., et al. (2024). *Real-Time Feedback on Older Adults Exercise: A Socially Assistive Robot Coaching System.* Proceedings of the 2024 IEEE-RAS International Conference on Humanoid Robots (Humanoids).

9. ROS Wiki Contributors. (2023). *ros_openpose Package Benchmark Notes.* Available at: `http://wiki.ros.org/openpose_ros`