mmPose-NLP: A Natural Language Processing Approach to Precise Skeletal Pose Estimation using mmWave Radars

1. Introduction

Background

- Traditional human pose estimation relies on optical sensors like Kinect and RGB cameras, which provide accurate skeletal tracking but suffer from occlusion, poor lighting conditions, and privacy concerns.
- mmWave radars provide an alternative approach, as they are unaffected by lighting variations, **offer better privacy protection**, and can penetrate minor occlusions, making them suitable for healthcare and surveillance applications.
- This study proposes an NLP-inspired approach to enhance pose estimation using mmWave radar point clouds, leveraging deep learning techniques to improve accuracy and robustness.

Research Goals

- Develop a deep learning framework that processes mmWave radar data to accurately estimate human skeletal poses.
- Implement an NLP-based methodology that transforms radar point cloud sequences into structured skeletal representations.
- Evaluate the performance of the proposed approach against existing vision-based methods such as Kinect.

2. Related Work

Pose Estimation Techniques

- **RGB-based Methods:**
  - Require good lighting conditions and a clear sight.
  - Struggle with occlusions and background clutter.
- **Depth-based (Kinect):**
  - Provides depth data but has range limitations and is sensitive to environmental factors.
  - Effective in controlled indoor settings but not ideal for challenging environments.
- **mmWave Radar-based Methods:**
  - Robust to lighting variations and minor occlusions.
  - Lower spatial resolution compared to optical sensors, requiring advanced processing techniques.

NLP in Pose Estimation

- Natural language processing (NLP) techniques, particularly sequence-to-sequence learning and attention mechanisms, have shown promise in modeling temporal dependencies in radar point cloud sequences.
- By leveraging recurrent neural networks (RNNs) such as GRUs and attention-based decoding, radar data can be effectively translated into structured pose representations.

3. Methodology

System Overview

- The system consists of two mmWave radar sensors that capture dynamic point cloud data of human motion from different perspectives.
- The raw point cloud data undergoes preprocessing, including noise reduction, voxelization, and coordinate transformation, to ensure consistency in skeletal pose estimation.
- A gated recurrent unit (GRU)-based encoder-decoder model with an attention mechanism processes the structured data to extract skeletal information.
- The system outputs precise skeletal joint positions, which are later evaluated against ground truth obtained from a Kinect sensor.

Key Components

1. Radar Point Cloud Processing

- **Point Cloud Acquisition:** The mmWave radar sensors detect reflected signals from human motion, capturing **N frames** of radar point cloud data.
- **Point Filtering:** Initial preprocessing removes noise and irrelevant reflections using DBSCAN.
- **Voxelization:** The filtered point cloud is mapped into a structured voxel grid, where each voxel represents a small unit of space. This process standardizes the input format for deep learning models.
- **Normalization and Embedding:** The voxelized data is normalized and embedded into a latent space that preserves spatial relationships between detected reflections and human body structure.

2. Deep Learning Model

- **GRU-based Encoder-Decoder Architecture:**
  - **Encoder Stage:**
    - A sequence of GRUs processes the radar frames to capture temporal dependencies between consecutive poses.
    - The encoded feature representation is passed to the decoder for skeletal reconstruction.
  - **Decoder Stage:**
    - The decoder iteratively reconstructs skeletal joint positions by predicting voxel representations frame by frame.
    - An attention mechanism selectively enhances key temporal and spatial features relevant for pose estimation.

- **GRU-based Encoder-Decoder Architecture:**
    - **Attention Mechanism:**
        - The attention mechanism helps focus on informative regions within the point cloud, improving joint localization accuracy and robustness.
- **Output Layer:**
    - The final prediction generates a **3D skeletal pose with 25 joints**, each represented as **(x, y, z) coordinates**.
    - The predicted pose undergoes **de-voxelization**, where the learned mapping transforms voxel indices back into continuous space for refined skeletal estimation.

4. Experiments & Results

Experimental Setup

- **Dual Radar Configuration:** Two mmWave radars are positioned to provide complementary viewpoints and reduce occlusion effects.
- **Ground Truth:** A Kinect v1 system is used as the reference to compare the estimated skeletal poses against a well-established motion capture system.
- **Data Collection:** Human motion sequences are recorded across different poses and movements to evaluate the robustness of the proposed method.
- **Synchronization:** The radar and Kinect systems are time-synchronized to ensure accurate ground truth alignment with radar-based estimations.

Data Fusion from Two Radars

- Each radar independently estimates skeletal joints using its respective point cloud data.
- **Coordinate Transformation:** To align the skeletal estimates from both radars, coordinate transformations are applied, mapping each radar’s coordinate frame to a global reference frame.

- **Joint Association:** A nearest-neighbor approach is used to match corresponding joint predictions from both radars, ensuring consistency.
- **Weighted Fusion:** The final skeletal joint position is determined by taking a weighted average of the two radar estimates, with weights assigned based on signal confidence and positional consistency.
- **Refinement:** The fused skeletal pose undergoes a final refinement step, where joint positions are adjusted using temporal smoothing techniques to reduce fluctuations.

Evaluation Metrics

- **Mean Per Joint Position Error (MPJPE):**
  - Measures the Euclidean distance between predicted joint positions and the ground truth Kinect skeleton.
  - Lower MPJPE values indicate better skeletal estimation accuracy.
- **Comparison with Optical Methods:**
  - The proposed mmWave-based approach is benchmarked against traditional vision-based methods to assess its robustness under occlusion and poor lighting conditions.

5. Applications & Future Work

Potential Applications

- **Fall Detection for Elderly Patients:**
  - Enables real-time monitoring of elderly individuals in home or healthcare environments.
  - Provides immediate alerts in case of detected falls, improving emergency response times.
- **Gesture Recognition and Human-Computer Interaction:**
  - Enables intuitive control of smart devices and systems through radar-based gesture inputs.
  - Offers enhanced privacy compared to camera-based recognition systems.
- **Smart Healthcare Monitoring Systems:**
  - Non-intrusive monitoring for posture assessment and rehabilitation tracking.

Future Improvements

- **Enhanced Radar Resolution:**
  - Deploying multi-view radar setups can significantly improve spatial resolution and pose estimation accuracy.
- **Advanced Data Fusion Techniques:**
  - Implementing deep learning-based sensor fusion models could further refine skeletal predictions.
- **Real-time Processing for Live Applications:**
  - Optimizing the model for real-time execution enables deployment in hospitals, assisted living facilities, and smart home environments.

Conclusion

- This study presents an **NLP-inspired approach** for estimating human skeletal poses using mmWave radar data.
- By leveraging **deep learning, recurrent architectures, and attention mechanisms**, the system achieves **competitive accuracy compared to traditional Kinect-based methods**.
- The method demonstrates **strong potential for real-world applications** such as fall detection, gesture recognition, and healthcare monitoring.
- Future work aims to refine data fusion techniques, enhance resolution, and improve real-time processing capabilities.