Temporal Dynamics in Human Pose Estimation Models

Monitoring people without cameras: Privacy is important!

Dan Teodor Savastre[1]

Supervisor(s): Marco Zuñiga Zamalloa [1]**, Girish Vaidya**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Dan Teodor Savastre
Final project course: CSE3000 Research Project
Thesis committee: Marco Zuñiga Zamalloa, Girish Vaidya, Michael Weinmann

Abstract

Human Pose Estimation using Millimeter Wave radars has emerged as a promising alternative to traditional camera-based systems, addressing privacy and deployment constraints. While state-of-the-art Deep Learning models predominantly focus on spatial feature extraction to determine the positions of key points in the human body, this research investigates the effects of incorporating temporal dynamics in such models. It focuses of modifying an existing state-of-the-art spatial model to account for temporal dynamics and compares the performance of the two models. Long Short-Term Memory networks are used to capture temporal dependencies between frames of point clouds which significantly boosts the precision of key point detection. The proposed temporal model demonstrates a 53% reduction in Mean Absolute Error and a 45% reduction in Root Mean Squared Error compared to state-of-the-art model. Moreover, these improvements were achieved with a less complex model architecture and similar training times. The robustness of the model was further validated on a different dataset, showcasing its potential for broad application in fields such as healthcare, sports analysis, traffic monitoring and robotics. This study underscores the efficacy of temporal dynamics in pose estimation, and showcases the advantages of accounting for temporal dependencies when evaluating more complex movements.

1 Introduction

Human Pose Estimation (HPE) refers to algorithms that determine the spatial positions of key points in the human body and using these points to reconstruct the digital skeleton of the person being analyzed. The applications of such algorithms are extended across a diverse range of fields, including healthcare, for monitoring patients posture during rehabilitation exercises, gesture recognition, sports analysis, robotics, military applications and traffic monitoring systems [1–5].

There are three main technologies that are used for human pose estimation: Computer Vision (CV), Wearables and Wireless sensing. Over the years, the number of cameras in public places has been increasing, with uses from security and surveillance, to entertainment. The presence of cameras in public spaces is something that people are growing increasingly wary about, and the European Union is taking action to limit camera deployments in public spaces. This is why Computer Vision systems, while effective, present significant drawbacks such as privacy concerns, and stringent deployment conditions regarding lighting and placement [1].

The second commercially available solution are wearable sensors, which pose an inconvenience to the users due to the necessity of regular charging, while also limiting the user's flexibility [6]. These wearable devices, or inertial sensors, suffer from higher noise levels, making it more difficult to accurately classify activities [1].

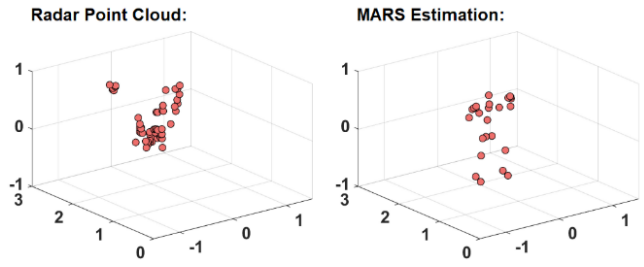

Figure 1: From left to right the figure shows the mmWave point cloud representation and estimated pose for one frame [1].

To address the drawbacks posed by current solutions, wireless sensing systems are gaining popularity, particularly systems using Millimeter Wave (mmWave), with state-of-the-art systems performing on par with, if not superior to, camera-based solutions. The benefit of using these radar-based systems is that the data collected is in the form of point clouds, from which personally identifiable information cannot be retrieved [7]. Moreover, radar-based systems also require less storage compared to a camera-based solutions, because point clouds retain less information compared to an image, as can be seen in Figure 1.

State-of-the-art systems for Human Pose Estimation using mmWave radars investigate either spatial dynamics or temporal dynamics, but not the effect of extending a spatial model by including temporal data. This paper aims to contribute to the existing body of research by analyzing the impact of incorporating temporal information into human pose estimation models. Specifically, investigating how accounting for the temporal domain influences the performance and complexity of state-of-the-art spatial HPE models. By addressing this aspect, this paper aims to provide insights that can inform the development of more robust and efficient human pose estimation systems.

This research builds upon an existing Human Pose Estimation model, referred to as MARS: mmWave-based Assistive Rehabilitation System[1] [1]. This model was proposed to assist in rehabilitation of patients with motor disorders and provide feedback on their movement. The model was extended in this research by incorporating Long Short-Term Memory (LSTM) Networks to capture temporal dynamics. This new temporal model was then trained and compared with the MARS model, with the experimental results demonstrating a significant overall reduction of the localization errors. Importantly, this precision improvement was achieved with negligible impact on computational complexity.

[1]MARS model available at https://github.com/SizheAn/MARS

2 Related Research

Human Pose Estimation is the task that involves detecting and predicting the spatial positions of key points, or landmarks, in the human body from images, videos, or sensor data. These key points typically include joints such as the elbows, wrists, hips, knees, etc. and are used to reconstruct the digital skeleton of the person being analyzed.

The evolution of Human Pose Estimation systems has been marked by significant advancements in both hardware and software technologies, leading to more accurate and efficient methods. The success of deep learning in CV-based HPE models is mainly due to the availability of big data, superior representation capability of deep neural networks, and high-performance hardware [9]. Comparing the CV-based HPE solution seen in Figure 2 with the point cloud data from Figure 1, the increased difficulty of performing pose estimation from mmWave radar data becomes more apparent.

For the purpose of this research, HPE models can be divided in two main categories: spatial models and temporal models. mmWave radar solutions for Human Pose Estimation emerged with spatial models employing Deep Learning techniques that predict 15 key points in the human body with similar performance to camera-based solutions [5, 10]. Following this, the MARS system introduced in 2021 further advanced the field by integrating more complex algorithms and processing techniques, such as Convolutional Neural Networks, and extending the number of key points estimated to 19 [1]. At the same time, researchers started publishing temporal models with comparable performance, leveraging Recurrent Neural Network architectures, such as LSTM or GRU, to capture temporal dynamics [3, 6, 11]. Because the architectures of these temporal models were significantly different from those of the spatial models, the impact of temporal dynamics is not immediately apparent. Currently, research is being done into a third category of HPE models, which is the so called Fusion Models that employ a spatial and a temporal model, along with a "fusion" between the two, resulting in more complex architectures [12, 13].

Figure 2: Human Pose Estimation using Computer Vision [8]

The field of wireless sensing for human pose estimation is relatively new, which means there are no major open-source datasets widely used by researchers. This lack of standardized datasets makes it difficult to compare existing spatial and temporal models to establish the importance of temporal data. The absence of benchmark datasets hampers the ability to validate and replicate findings, highlighting the need for collaborative efforts to develop and share comprehensive datasets in this emerging field. Addressing these challenges is essential for advancing the state of the art in HPE and enabling more consistent and comparable evaluations of different methodologies [7, 14].

As mentioned before, both spatial and temporal models for HPE have been proposed by researchers with promising results. Despite this a critical knowledge gap remains: understanding the impact of temporal data on spatial models, specifically, if LSTM can be used to enhance the performance of existing spatial models by capturing temporal dynamics. Temporal models, which incorporate sequential data, have demonstrated the ability to enhance model performance by capturing the dynamics of human motion over time. However, the specific influence of temporal data on the accuracy and computational complexity of spatial HPE models has not been thoroughly investigated. This research aims to fill this gap by analyzing how incorporating temporal information, particularly through LSTM networks, affects the performance and computational complexity of a state-of-the-art spatial model.

3 Problem Statement

As mentioned before, this research aims to address the following main question: How does incorporating Temporal Dynamics in mmWave radar-based Human Pose Estimation models impact the performance and complexity of the model? To address this overarching question, the research is guided by two sub-questions:

- How does the inclusion of temporal data enhance the **precision** of HPE in comparison to models that do not use temporal data?
- How does the **complexity** of HPE models change when temporal data is incorporated?

By systematically investigating these sub-questions, the research seeks to provide a comprehensive understanding of the effectiveness of temporal dynamics in improving pose estimation accuracy and the associated computational costs.

4 Methodology

This chapter outlines the approach taken to investigate the impact of temporal dynamics on mmWave-radar based Human Pose Estimation models. This study integrates LSTM networks into an existing state-of-the-art model which focuses solely on spatial feature extraction. By leveraging temporal dependencies inherent in sequential data, the modified model aims to achieve superior accuracy in predicting the spatial positions of key points in the human

body. In the following sections the datasets used in this research are introduced, followed by a description of the MARS model, alongside the model performance that forms the baseline for the comparisons in section 6.

4.1 Datasets

mmWave radars collect data in the form of 5D time-series point clouds, as seen in the left side of Figure 1, meaning that at each time step the radar registers points with 3D coordinates(x, y, z), and additionally it records the Doppler Velocity and Signal Intensity for each point, resulting in 5 dimensions [15]. The additional dimensions provide information about the direction and speed of each point, enabling more accurate and robust models.

The first dataset that is used in this research is the one from the MARS paper, that uses a Microsoft Kinect V2 for the ground truth reference and the Texas Instruments (TI) IWR1443 Boost mmWave radar [16] for the radar processing. The dataset contains ten gestures: Left upper limb extension, Right upper limb extension, Both upper limbs extension, Left front lunge, Right front lunge, Squat, Left side lunge, Right side lunge, Left limb extension and Right limb extension. Four different users performed each movement for 2 minutes resulting in a dataset with 2.28 million reference data points from Kinect V2 and 3.81 million data points from mmWave data.

The second dataset that was used in this research is one gathered by a master student from TU Delft[2]. This dataset focuses on five different movements performed by 15 participants: Static Waving, Normal Walking, Combined Walking and Waving, Static Movement and Free Movement. The benefit of including this dataset is that it provides more variety to the data and provides tests for moving targets, which the MARS dataset does not include.

4.2 MARS Model

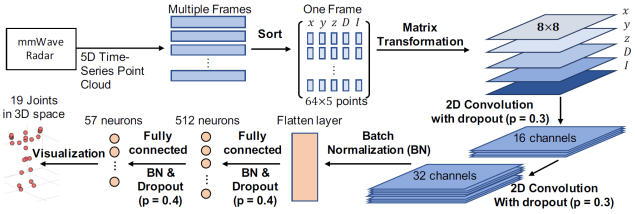

Figure 3: Overview of MARS model architecture [1]

The MARS model, a system designed for assistive rehabilitation in patients with motor disorders, serves as the baseline state-of-the-art spatial model for this research. It uses point cloud data to estimate the 3D coordinates of 19 key points in the human body, corresponding to 57 outputs. This model primarily focuses on data pre-processing and feature extraction using a CNN architecture. The performance metrics used to evaluate the MARS model are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The MARS Model has a total of 3,255,469 parameters, of which 1,084,793 are trainable.

Data Pre-processing: At each time step, the mmWave radar stores at most 64 points to form a data frame. Since the points are stored in random order, one way of standardizing between frames is to sort the points in ascending order of their x, y and z coordinates. This means the points are sorted in ascending order of their x coordinates, then the points with the same x coordinate are sorted by their y coordinates, and finally, the points with the same x and y coordinates are sorted by their z coordinates. The shape of the resulting data frame is 64x5, which through the matrix transformation in Figure 3 is rearranged in 5 channels, each with an 8x8 feature map.

CNN Architecture for Feature Extraction: The model employs a CNN architecture that incorporates Batch Normalization (BN) and Dropout layers to enhance performance and prevent overfitting. The Convolutional layers in Figure 3 extract features from the input mmWave point cloud data. The Batch Normalization layers are used to stabilize and accelerate the training by normalizing the input of each layer, improving the convergence and preventing overfitting. Dropout layers are used to randomly set a fraction of the input units to zero during training to prevent overfitting by reducing reliance on specific neurons. Finally, the output is passed through the fully connected layers to output the predicted coordinates of the 19 key points at that frame.

Performance Metrics: The performance of the MARS model is evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Figure 4 shows the baseline MAE that the temporal model will be compared to. To evaluate the quality of a prediction the absolute error for each key point is calculated by taking the distance between the ground truth and the predicted coordinates. The sum of these errors is then divided by the number of key points giving the MAE for that prediction. MAE treats all errors equally, providing a balanced view of the average error. On the other hand, RMSE is more sensitive to outliers because it squares the errors before averaging them. This means that if the RMSE is much

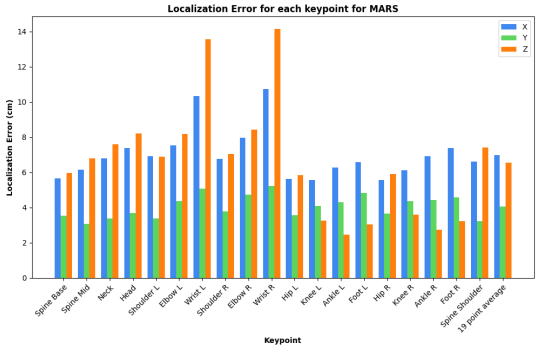

Figure 4: Mean Absolute Error for each key point in MARS model

[2]This dataset is currently private. A reference to it will be added once it becomes publicly available.

higher than the MAE, it suggests that there are some large errors in the predictions. These metrics along with the number of parameters and computational complexity provide a baseline for comparing improvements with the incorporation of temporal dynamics.

5 Temporal Human Pose Estimation Model

5.1 Modifying MARS to account for Temporal Dynamics

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that was designed to address the vanishing gradient problem in traditional RNNs. The LSTM model, proposed by Hochreiter and Schmidhuber in 1997 [17], includes a special unit known as the memory cell that serves as an accumulator or a gated leaky neuron. This memory cell can learn long-term relationships and has the unique ability to remember or forget information, making it highly effective for sequence prediction tasks.

Initially developed for sequence data such as word and sentence prediction, LSTM networks have been adapted to handle time-varying data, aligning well with the continuous and time-dependent nature of the pose estimation problem. By extracting the dynamics from the data, the LSTM network can effectively track the temporal dependencies between different poses thereby improving the precision of Human Pose Estimation.

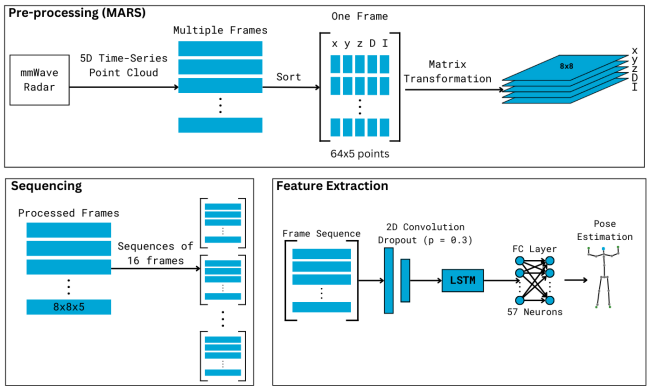

Figure 5: Overview of modified MARS architecture incorporating temporal dynamics

The Temporal Model builds on top of the existing MARS model to incorporate temporal dynamics by making some modifications to the model architecture. Figure 5 shows the architecture of the temporal model, with the pre-processing strategy used for MARS, alongside two main changes. Firstly, after the pre-processing, a sequencing step is added. This step divides the data into sequences of frames, with the length of the sequence representing the amount of memory the LSTM has. The second change comes in the feature extraction step, where after the Convolutional layers of MARS, an LSTM layer is added to capture the temporal dynamics of the data, followed by the fully connected layer as output. The modifications to the MARS model result in a temporal model with 1,208,248 total parameters, of which 402,386 are trainable. In comparison, the original MARS model contains 3,255,469 total parameters with 1,084,793 being trainable. This represents a substantial reduction in model complexity, with a 62.88% decrease in total parameters and a 62.90% decrease in trainable parameters. Despite this reduction, the Temporal Model significantly improves the pose estimation performance, demonstrating its efficiency and effectiveness in capturing temporal dependencies.

To accurately compare with the MARS model, the Temporal Model must be optimized. From the modifications brought to MARS there are two main parameters that need to be fine-tuned:

- **Sequence Length:** This determines how many frames the LSTM can remember. It is essential to find the optimal sequence length that balances memory capacity and computational efficiency.
- **Number of LSTM Units:** This defines the output dimensionality of the LSTM layer. For instance, 32 units means the LSTM layer output will have 32 neurons. The optimal number of units needs to be determined to ensure the model's effectiveness.

By systematically exploring these parameters and configurations, this research aims to determine the most effective way to integrate temporal data into HPE models, thereby enhancing their performance and understanding the trade-offs in computational complexity. Once an optimal temporal model is found, section 6 shows the comparison of this model with respect to MARS.

5.2 Optimization of LSTM Parameters

The LSTM model's performance is subject to the optimization of two parameters: sequence length and the number of LSTM units. The sequence length determines the number of frames the LSTM layer processes at a time, which must be optimized to balance memory capacity and computational efficiency. The number of LSTM units, which defines the output dimensionality of the LSTM layer is also subject to fine-tuning to ensure optimal performance.

With the goal of optimizing the parameters for LSTM, a series of tests was conducted to ascertain the most effective options. The sequence lengths were varied between 2 frames

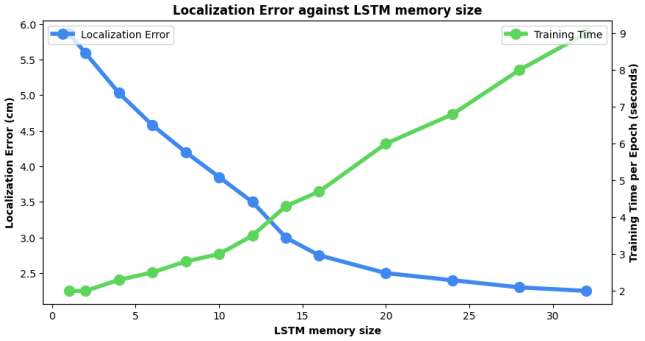

Figure 6: Localization error and training time plotted against **LSTM memory size**

and 32 frames, with the results of this experiment shown in Figure 6, proving that the optimal sequence length is approximately 16 frames of memory. With memory sizes larger than 16 frames, the computational complexity increase outweighs the improvement in precision. The underlying hypothesis for these tests is that the sequence length does not need to extend beyond half of a gesture, and that any additional information will result in increased computational complexity and negligible reduction of the error, diminishing returns. As for the number of LSTM units, the values tested were 32, 57, 128 and 256, the results being shown in Figure 7. From this, it can be seen that 32 units results in a loss of information, while more than 128 units results in increased computational complexity. Analyzing the figure, the hypothesis that increasing the dimension may result in increased computational complexity and training time, while having a small dimension for the LSTM output may result in loss of information is confirmed, and the optimal number of LSTM units for the temporal model was chosen to be 57. This value was used as 57 is the number of outputs of the model as well, meaning that the LSTM layer can learn dependencies of the 3 coordinates of the 19 key points throughout the movements.

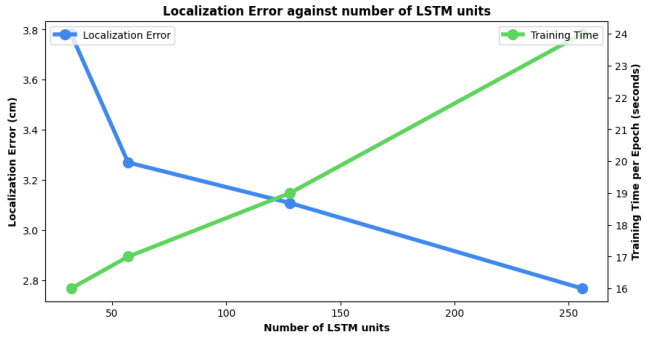

Figure 7: Localization error and training time plotted against **LSTM units**

For these tests, the performance of the temporal model was evaluated using the Mean Absolute Error (MAE). Additionally, the training time, number of parameters and computational complexity were also take into account, as these factors are crucial in practical applications. In section 6, the optimized temporal model is compared to MARS on the MARS dataset, as well as the second dataset, compiled by a master student at TU Delft.

6 Experimental Setup and Results

6.1 Experiment Setup

The programming language chosen for this research was Python, utilizing TensorFlow and Keras as the primary frameworks. The datasets used when conducting the experiments were split into three subsets: 60% of the data was used for training, 20% was used for validation, and 20% for testing. To ensure the robustness of the results, the implemented training loop trained 10 models for 150 epochs, using batches of size 128. The accuracy for each model was saved, and the average of the 10 models was taken to eliminate any outliers and improve reproducibility. This section evaluates the optimized Temporal Model against the state-of-the-art spatial model MARS on the dataset used in MARS, underlining the benefits of temporal dynamics for human pose estimation. Following this, a second analysis is performed, on the Moving Target Dataset of a TU Delft student, to evaluate the model's performance on more complex movements.

6.2 Testing Temporal Model on MARS dataset

Figure 8 compares the localization error of the state-of-the-art spatial model with that of the optimized Temporal Model using the two metrics: MAE and RMSE. It can be seen that the Temporal Model performs better on the MARS Dataset, with an average MAE of 3.04, 2.37 and 2.85cm for the x-, y- and z-axes respectively. The overall MAE of the Temporal Model of 2.75cm shows a 53% improvement over the 5.87cm MAE of the MARS model. Moreover, looking at the Root Mean Squared Error box plots, it can be seen that the Temporal Model represents an improvement over the MARS model, with the overall RMSE reduced by 45%, from 8.10cm to 4.44cm.

The temporal model not only improves the overall accuracy, as evidenced by the 53% reduction in MAE, but also significantly reduces larger errors, as shown by a 45% reduction in RMSE. This indicated that the temporal model is effective in minimizing both typical and more substantial errors in human pose estimation.

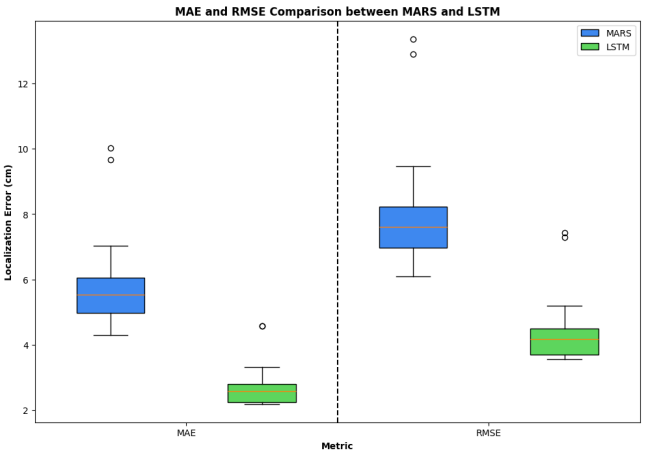

Figure 8: Box plot comparing MARS and Temporal Model on MARS Dataset

6.3 Testing Temporal Model on Moving Target Dataset

To evaluate the generalizability of the optimized model, it was tested on a more complex dataset gathered by a master student at TU Delft. This dataset, is comprised of 5 distinct movements: static waving, normal walking, combined walking and waving, static movements and free movement. This data was instrumental in assessing the model's

robustness and adaptability to previously unseen scenarios. By applying the optimized model to this dataset, the study aimed to verify whether the improvements achieved during the optimization phase could be consistently reproduced across varied and untrained data samples.

The accuracy of the two models evaluated on the Moving Target Dataset is found in Figure 9. From the box plots it can be seen that the optimized Temporal Model achieves a 34% reduction in MAE, representative of an overall prediction improvement. Moreover, MARS performs much worse in terms of RMSE, with outliers of over 100cm from the ground truth location, accentuated by the logarithmic scale of the y-axis in Figure 9. The temporal model offers an 82% improvement over the RMSE of the MARS model, from which it can be concluded that the Temporal Model is much more effective at handling and reducing large errors in pose estimation.

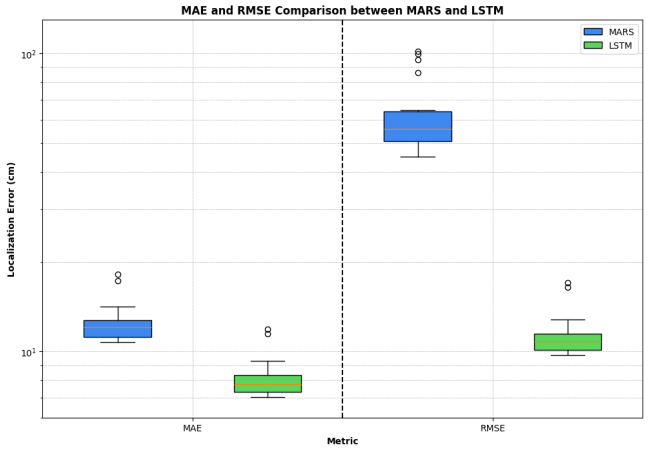

Figure 9: Box plot comparing MARS and Temporal Model on Moving Target Dataset

The results of the two models on the Moving Target Dataset emphasize the potential of Temporal Dynamics in Human Pose Estimation. They show that by incorporating this aspect of the data, more accurate and versatile models can be created, while keeping the architecture simple and the training times low. The increased error MARS achieved on this second dataset, underlines the importance of fully utilizing the different aspects of the data, and in the case of time series, temporal dynamics are an important one.

7 Responsible Research

The MARS dataset used in this research is open source. The second dataset, Moving Target Dataset, is currently under development, therefore it has not been published at the time of writing. The ethical and consent considerations, as well as the data collection procedures associated with these datasets are addressed in the respective research papers referenced in this report.

To ensure that the experiments have been conducted transparently and that they are reproducible, the experimental setup is detailed in section 6 of this report. The code used for this research is publicly available on GitHub[3]. The repository includes the models, the MARS dataset, training loops and other necessary components to replicate the experiments conducted in this study.

In conducting this research, care was taken to avoid introducing or perpetuating bias. The datasets used are diverse and cover a range of human poses and activities, helping to ensure the generalizability of the models. Validation and testing were performed rigorously to assess the fairness and accuracy of the models, minimizing the risk of biased outcomes.

The computational resources used for this research were optimized to balance performance and environmental impact. By leveraging efficient algorithms and model architectures, the research aimed to minimize energy consumption while maintaining high standards of precision and efficiency. The results of this study have the potential to benefit various fields, including healthcare, sports analysis, robotics and traffic monitoring, by providing more accurate and efficient human pose estimation models.

8 Conclusions and Future Work

The results of this research demonstrate that incorporating temporal dynamics into human pose estimation models significantly improves precision while also decreasing the model complexity. By leveraging LSTM networks to capture temporal dynamics, the enhanced model exhibited a reduction in MAE and RMSE of 53% and 45% respectively, when compared with the state-of-the-art spatial model on the MARS dataset. This improvement underscores the value of temporal information in enhancing the accuracy of pose estimation, making the approach viable for applications in healthcare, sports analysis, and many others. Additionally, the model's performance on the Moving Target dataset validated its robustness and adaptability, reinforcing its potential for broader application. The performance gap between the two models when evaluated on the Moving Target Dataset, proves that Temporal Dynamics significantly improve the model performance in more complex applications.

Future work in this field can focus on exploring alternative architectures that further capture temporal dynamics, such as Temporal Convolutional Networks. This exploration would aim to balance precision and computational efficiency while potentially offering improvements over the current LSTM-based approach. Moreover, the integration of spatial and temporal models into a unified framework, or "fusion models", represents a promising direction for future research. These models could combine the strengths of both domains, providing even more accurate and efficient human pose estimation systems. Another avenue for future research includes the development and sharing of standardized, open-source mmWave-radar datasets to facilitate the validation and replication of findings across the research community.

[3]https://github.com/dansavastre/MARS_LSTM

References

[1] Sizhe An and Umit Y. Ogras. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems*, 20(5s):1–22, 2021.

[2] G. Bhat, N. Tran, H. Shill, and U. Y. Ogras. w-har: An activity recognition dataset and framework using low-power wearable devices. *Sensors*, 20(18):26, 2020.

[3] D. K. Bila, M. Unel, L. T. Tunc, and Ieee. Improving vision based pose estimation using lstm neural networks. In *46th Annual Conference of the IEEE-Industrial-Electronics-Society (IECON)*, IEEE Industrial Electronics Society, pages 483–488, 2020.

[4] B. Solongontuya, K. J. Cheoi, and M. H. Kim. Novel side pose classification model of stretching gestures using three-layer lstm. *Journal of Supercomputing*, 77(9):10424–10440, 2021.

[5] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20(17):10032–10044, 2020.

[6] S. Setiyadi, H. Mukhtar, W. A. Cahyadi, C. C. Lee, W. T. Hong, and Ieee. Human activity detection employing full-type 2d blazepose estimation with lstm. In *IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pages 200–206, 2022.

[7] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):849–856, 2020.

[8] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. B. Chen, C. X. Huang, and C. H. Yang. The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. *Ieee Access*, 8:133330–133348, 2020.

[9] G. J. Lan, Y. Wu, F. Hu, and Q. Hao. Vision-based human pose estimation via deep learning: A survey. *Ieee Transactions on Human-Machine Systems*, 53(1):253–268, 2023.

[10] A. Adhikari, S. Sur, and Machinery Assoc Comp. Millipose: Facilitating full body silhouette imaging from millimeter-wave device. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) / ACM International Symposium on Wearable Computers (ISWC)*, pages 1–3, NEW YORK, 2021. Assoc Computing Machinery.

[11] X. T. Shi, T. Ohtsuki, and Ieee. A robust multi-frame mmwave radar point cloud-based human skeleton estimation approach with point cloud reliability assessment. In *IEEE Sensors Conference*, IEEE Sensors, NEW YORK, 2023. Ieee.

[12] G. Mei and et al. Radar-based 3d skeleton estimation enhanced with joint temporal-spatial constraints. *Lecture Notes in Electrical Engineering*, 2023.

[13] S. Z. An, U. Y. Ogras, and Acm. Fast and scalable human pose estimation using mmwave point cloud. In *59th ACM/IEEE Design Automation Conference (DAC) - From Chips to Systems - Learn Today, Create Tomorrow*, pages 889–894, NEW YORK, 2022. Assoc Computing Machinery.

[14] A. J. Chen, X. Y. Wang, S. H. Zhu, Y. X. Li, J. M. Chen, and Q. Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *30th ACM International Conference on Multimedia (MM)*, pages 3501–3510, NEW YORK, 2022. Assoc Computing Machinery.

[15] Cesar Iovescu Rao and Sandeep. The fundamentals of millimeter wave radar sensors, July 2020.

[16] Texas Instruments. IWR1443BOOST, 2014.

[17] S Hochreiter and J Schmidhuber. Long short-term memory. *NEURAL COMPUTATION*, 9(8):1735–1780, NOV 15 1997.