Multi-modal pose estimation in XR applications leveraging integrated sensing and communication

Nabeel Nisar Bhat
University of Antwerp-imec, Belgium

Javad Sameri
Ghent University-imec, Belgium

Jakob Struye
University of Antwerp-imec, Belgium

Maria Torres Vega
KU Leuven, Belgium

Rafael Berkvens
University of Antwerp-imec, Belgium

Jeroen Famey
University of Antwerp-imec, Belgium

Abstract

Mobile extended reality (XR) applications are anticipated to generate substantial traffic for 6G. Such applications not only require high data rate and low-latency transmissions, but also accurate and real-time pose estimation to enable interactive and immersive experiences. While sub-6 GHz signals have been exploited for pose estimation, they cannot cope up with multi-gigabit data rates required by XR applications. Instead, mobile communications at mmWave frequencies can potentially support data rates up to several giga-bits per second (Gbps) and, therefore, can be used to deliver XR content wirelessly to the Head-Mounted Display (HMD). Moreover, mmWave frequencies can offer improved sensing due to the large available bandwidth. Therefore, mmWave communications can play a crucial role in enabling device-free interactivity by offering both high-speed communication and accurate sensing capabilities. However, mmWave propagation characteristics are different from sub-6 GHz. Path loss plays a significant role, and can lead to degraded sensing performance. Therefore, our proposal supplements wireless sensing at mmWave frequencies with wireless electromyography (EMG) armbands. By capturing patterns of muscle activities, we can counteract the limitations of mmWave-based pose estimation, thereby enriching the granularity and precision of pose estimation. This paper proposes a conceptual architecture to achieve multi-modal pose estimation for XR applications. Early results highlight the shortcomings of mmWave-based sensing, and we identify future steps and opportunities on integration of both approaches.

ImmerCom '23, October 6, 2023, Madrid, Spain

ACM ISBN 979-8-4007-0339-3/23/10…$15.00
https://doi.org/10.1145/3615452.3617944

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Machine learning*.

Keywords: Extended reality, integrated sensing and communication, mmWave, channel state information, electromyography, pose estimation

ACM Reference Format:

Nabeel Nisar Bhat, Javad Sameri, Jakob Struye, Maria Torres Vega, Rafael Berkvens, and Jeroen Famey. 2023. Multi-modal pose estimation in XR applications leveraging integrated sensing and communication. In *The 1st ACM Workshop on Mobile Immersive Computing, Networking, and Systems (ImmerCom '23), October 6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3615452.3617944

1 Introduction

In recent years, extended reality (XR), encompassing virtual reality (VR), augmented reality (AR), and mixed reality (MR), has emerged as a cutting-edge application for future wireless networks. XR applications have disrupted various fields such as education, industry, health care, and entertainment. In such applications, pose estimation is pivotal to enable immersive and realistic experiences.

Pose estimation involves predicting the location and orientation of joints or keypoints of the human body. The goal is to determine position and orientation of a human relative to a given scene. Pose estimation finds applications in gaming, healthcare, sports, and extended reality (XR). However, our focus is on XR, as it requires not only accurate but also real-time pose. This pose information is crucial in enabling a realistic and intuitive experience between the user and the XR world. State-of-the-art solutions for pose estimation typically rely on cameras. For instance, OpenPose [4] can estimate the 2D pose of multiple people in an image. It can jointly detect body, foot, hand, and facial keypoints from a single image. Some commercial products such as Kinect [29] can also predict and track body keypoints with high accuracy. Kinect can output a 2D or 3D skeleton of a human body from an image for a total of 25 keypoints. However, camera-based methods work well only in line-of-sight (LOS) scenarios.

Moreover, these methods can infringe on the privacy of the user [28].

Specifically for VR, current solutions require users to wear handheld controllers to interact with VR content. However, these controllers fully occupy users' hands and are unintuitive to many users. To overcome these limitations, we plan to leverage existing communication infrastructure to enable device-free interactivity. We envision a generic scenario where mobile communication signals can be used to wirelessly stream XR content to the head-mounted display (HMD), and the same signals can also be used for sensing (pose estimation). These poses will be used as inputs to enable users to control and manipulate virtual objects and environments. The approach of using communication signals also for sensing is known as integrated sensing and communication (ISAC) [5]. However, delivering XR requires high data rates and low latency communication for an immersive experience. Mobile communications at sub-6 GHz cannot handle XR content delivery due to limited bandwidth. Instead, mmWave and sub-THz frequencies (30-300 GHz) [6, 21, 25] can stream XR content seamlessly due to the large available bandwidth, enabling a data rate up to several Gbps. Additionally, the large bandwidth can also improve sensing. For example, utilizing a 2 GHz bandwidth at 60 GHz can lead to 15 cm of raw resolution, which can greatly benefit localization applications [14]. Moreover, a high number of antennas can be packed at higher frequencies leading to accurate angle estimation. Overall, improved sensing and communication is possible with mmWave frequencies.

In this work, we propose a proof-of-concept multi-modal pose estimation for XR applications that leverages mmWave communication signals. For this reason, we extract channel state information (CSI) from commercial-off-the-shelf (COTS) mmWave Wi-Fi access points (APs). CSI has been used for pose estimation at sub-6 GHz frequencies with reliable accuracy. However, utilization of CSI for pose estimation at mmWave with COTS has not been studied yet. This work aims to demonstrate the effectiveness of ISAC-based pose estimation leveraging CSI from mmWave Wi-Fi. However, mmWave signals have some inherent challenges. They are highly directional and suffer from high path loss. Also, such signals are highly sensitive to obstructions and blockage, making them highly dependent on the position of devices (mmWave sensors) or users. Moreover, mmWave signals have a short range that can restrict coverage area for sensing applications. We propose to tackle this issue by using electromyography signals (EMG) [12]. EMG signals are robust to obstacles and can capture detailed muscle movements. In particular, surface electromyography (sEMG) can be used for detecting fine-grained muscle activities, e.g., finger movements [24]. By combining communication signals with EMG signals, our goal is to infer the overall pose of the human body while countering the limitations of mmWave and enhancing the granularity of pose estimation. With this setup, we believe we can accurately track the user's poses and translate their physical movements into the virtual world, allowing a virtual avatar to replicate the same poses. As a preliminary step, we investigate the effectiveness of pose estimation only with mmWave ISAC signals. We go beyond simple classification and predict skeleton-based poses from mmWave data. Our experiments with three users performing a set of 8 XR-related gestures in a single environment validate the potential of mmWave-based pose estimation. However, the challenges encountered highlight the need for an additional sensing modality.

2 Related work

2.1 Integrated sensing and communication (ISAC)

Rapid advances in wireless communication technologies and the growing demand for sensing applications have led to the emergence of ISAC [5]. ISAC is gaining significant traction, as evidenced by recent standardization efforts (e.g., IEEE 802.11bf [17, 20]) and its identification as one of the main ground-breaking innovations in 6G by leading telecom vendors. The standout advantage of ISAC is that most of the network infrastructure is already in place for communication purposes anyway, which facilitates a multi-sensory mesh that can offer sensing capabilities at a limited additional cost. Most of the focus in ISAC has been on Wi-Fi signals. Wi-Fi signals have opened new avenues for sensing applications giving stiff competition to vision-based approaches. Wi-Fi signals have been used for a plethora of applications ranging from coarse activity detection [17, 18] to fine-grained gesture recognition [16, 31]. Moreover, Wi-Fi signals have also been used for a full skeletal-based pose estimation [19, 30, 32]. However, most of these works utilize sub-6 GHz signals. On the other hand, there exists a notable gap for COTS sensing at mmWave. Very few works have explored the potential of COTS devices for sensing in the mmWave spectrum. At mmWave frequencies, high bandwidth and massive number of antennas allow for unprecedented accuracy in wireless sensing.

Toshiaki et al. [13] have made pioneering contributions in mmWave-based sensing using COTS devices. In particular, the authors use spatial beam signal-to-noise ratios (SNRs) for fingerprinting-based indoor localization. TP-Link Talon AD7200 routers are used for data collection. The effectiveness of experiments is evaluated over 7 different locations and a total of 28 orientations; 4 orientations at each location in a regular office space. Deep learning is used to extract features from beam SNRs and to predict different locations. The results show 100% accuracy for location classification, 99% for simultaneous location and orientation classification. In addition, an average root mean-square error (RMSE) of 11.1 cm is achieved when predicting the coordinates. In another

study, Yu et al. [27] used beam SNR from the same mmWave APs (TALON AD7200) for human pose and seat occupancy classification. The validity of pose estimation is conducted along five different sessions for 8 distinct poses in a single environment. Deep learning is used to extract features from different poses. The method achieves an overall accuracy of 91.2%. Furthermore, Blanco et al [3] investigate the design of joint sub-6 GHz and mmWave localization system. CSI-based angle estimation is used to predict location. MikroTik wAP 60Gx3 devices are used for data collection. The system achieves 18 cm median location error with COTS devices. In our previous work [2], we used beam SNR from mmWave COTS APs for gesture recognition in multiple environments and compared it with CSI collected from sub-6 GHz APs. With deep neural network, we were able to achieve 96.7% accuracy in a single environment in classifying 10 XR-related poses. However, as the task became more complex (e.g., when adding more users, changing environments or altering orientations of gestures), the accuracy dropped significantly. This is because mmWave signals are highly directional and susceptible to blockage, and sensitive to the shape of obstacles, making it difficult for neural network to extract relevant information.

In this work, we go beyond classification and aim at regressing full body pose skeletons from mmWave CSI. Compared to classification, skeleton-based pose provides fine-grained information and can be used for device-free interactivity. Skeleton information allows us to estimate the positions and orientations of various body joints, such as the head, shoulders, elbows, wrists, and knees. This enables us to reconstruct the user's complete body pose in a virtual environment. However, the problems with mmWave encourage the need for a multi-modal approach. To counter the limitations of mmWave-based sensing, we propose the multi-modal approach based on ISAC and EMG signals, making the overall system less sensitive to the blockage and shape of obstacles. This integration will allow an enhanced understanding of user movements and improve the accuracy of pose classification as the task becomes complex.

2.2 Surface Electromyography (sEMG)

Surface Electromyography (sEMG) has been developed as a non-invasive, inexpensive, and easy-to-set-up method for studying muscle activity [22]. Engineering developments have extended electromyography (EMG) beyond medical and rehabilitation applications to areas like Human-Computer interactions (HCI). Applications like gesture command, sign language translation, and finally, mapping user gestures in virtual environments[15].

Furthermore, sEMG exhibits substantial promise as a tool for delivering accurate hand gesture recognition and tracking estimation. The recorded data contains information regarding hand posture. Nonetheless, these signals are not directly usable due to the contamination with noise from various undesired sources[22, 26]. sEMG signals' significant inter-subject variability complicates interpretation, necessitating robust algorithms for reliable gesture recognition. Recent research employed a CNN and Bi-LSTM hybrid model on sEMG signals, achieving a 98.33% accuracy by training subject-specific models [11]. Another study mapping hand gestures to a virtual environment achieved up to 80±2.36% accuracy in subject agnostic setting, underscoring the inter-subject variability issue [15]. Models demonstrated high accuracy for a small number of gestures but showed limitations when the number increased, affecting the continuous decoding of hand and finger positions [22].

To further advance the potential of sEMG signals, many researchers have incorporated IMU sensors alongside these signals [10]. A comparative study between sEMG and combined sEMG-IMU systems demonstrated a 7% improvement in classification resulting in 97.5% accuracy and a 1% improvement in recognition resulting in 88.15% accuracy. Furthermore, the study indicated a reduction in uncertainty, as evidenced by the decreased standard deviation of the classification results when employing sEMG-IMU signals instead of solely relying on sEMG signals[26]. In another innovative approach, researchers employed explainable AI (XAI) to reduce sensor requirements and attempted to elucidate the operational principles of such a multi-modal system[10].

Despite the promising potential of sEMG, there is limited research exploring the fusion of sEMG data with sensor inputs other than IMU. Jiang et al. [8] proposed a fusion-based wristband combining force myography (FMG) and EMG for hand gesture recognition, an approach designed to address the sweat vulnerability of sEMG and the low information density associated with FMG. Furthermore, a multi-modal system integrating sEMG and ultrasound devices was developed in [7]. Its performance was compared with systems relying only on these individual physiological signals for predicting continuous finger angles [7]. However, to the best of our knowledge, no multi-modal system has been introduced, leveraging both CSI and sEMG measurements.

3 Proposed Conceptual Framework

In this section, we will begin by presenting a conceptual multi-modal framework that integrates CSI and sEMG information for body pose estimation. Subsequently, we will delve into our methodology for processing CSI data, which serves as a first step towards a proof of concept for our approach.

Our proposed framework for body-pose estimation involves utilizing two wireless devices to capture CSI, two sets of EMG wristbands for recording sEMG signals, and a camera to collect ground truth data. The ground truth data captured with the camera is only required during the training phase of

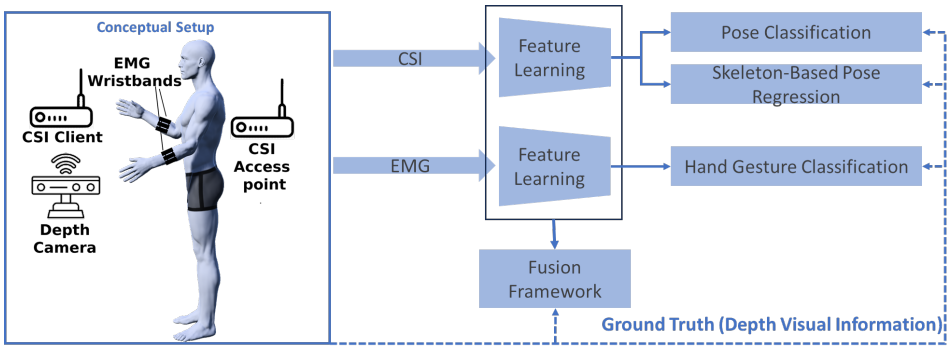

Figure 1. Conceptual Framework

the system. Each signal type requires a dedicated model to extract and interpret the features. Subsequently, a higher-level model is trained to integrate the feature sets extracted by the primary models. Figure 1 depicts the conceptual diagram of this framework.

Considering the distinct nature and sampling rates of EMG and CSI, we hypothesize that convolutional neural networks (CNNs) would be an effective approach for feature fusion [9]. In particular, multi-modal late fusion is deemed suitable. Late fusion is the predominant approach for information fusion in academic research due to its simplicity in integrating data from various sources. Moreover, late fusion enables the utilization of well-established architectures and methods that have been extensively researched and refined over time. This approach even allows for the incorporation of transfer learning and pretrained models, enhancing its versatility. Additionally, late fusion is particularly suitable for our specific setup, where each modality possesses distinct dimensions [9]. Consequently, late fusion eliminates the need to address data misalignment issues, making it an ideal choice for our fusion process. This approach entails training specific models for each modality, enabling us to leverage the strengths of individual models that have been developed specifically to extract features from specific sensors in the initial feature learning phase. By combining these features using a CNN at higher levels, we can effectively fuse information from the sensors.

Furthermore, for simultaneous utilization of CSI and sEMG information, a significant challenge arises from the fact that the sEMG wristbands primarily provide hand position and orientation information, whereas CSI provides information about the entire environment. Consequently, determining the optimal parameters becomes challenging due to the large search space and meager overlap of information. To address this, a novel approach can be adopted by dividing the problem into two subproblems. The first subproblem involves estimating the general body posture using CSI signals, while the second subproblem focuses on fusing information from both CSI and sEMG in hand areas to enhance the resolution of the body pose estimation framework. This localized approach narrows the search space, making it easier for the model to identify optimal parameters for hand gesture recognition.

Finally, the merit of this framework lies in effectively utilizing the advantages associated with the two sensor types deployed. Both methods offer privacy-preserving measurements, avoiding the privacy concerns frequently associated with vision-based sensing systems. Additionally, these non-invasive methods enable rapid inference, thus facilitating real-time applications. These attributes render them an optimal choice for a multi-modal fusion system for gesture recognition. Furthermore, sEMG signals offer fine-grained information on hand gestures , thus supplementing CSI-based body pose estimation and enhancing system resolution. However, the potential applications of such a comprehensive setup remain unexplored, signifying a significant opportunity for further research and development.

3.1 Methodology

In this early work, we use only CSI from mmWave Wi-Fi APs for pose estimation and leave integration of EMG signals for future work. Since phase measurements of CSI tend to be noisy, we use only the amplitude part. To enhance the accuracy of our measurements, we employ background subtraction to eliminate static reflections or clutter from the CSI. The resulting CSI, with clutter removed, is then inputted into learning-based algorithms. Our analysis covers both classification and regression-based learning. In the case of regression, our objective is to predict skeletal-based poses from the mmWave CSI data.

We employ common classification methods such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Decision Trees. These methods are considered standard and serve as a useful reference point in evaluating the performance of more advanced deep learning algorithms. With regard to deep learning, we experimented with several architectures, such as fully connected layers (FCN), convolutional

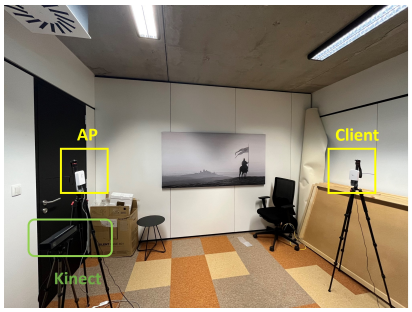

Figure 2. Environment and Experimental Setup

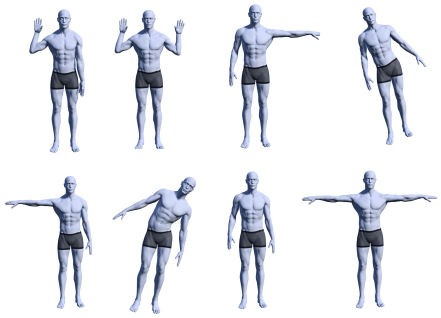

Figure 3. Poses

neural networks (CNN), and a combination of both. In particular, we explored three architectures: one fully connected layer (FCN), four fully connected layers (4 FCN), and a CNN inspired by the work Szegedy et al.[23]. After careful experimentation, we found that a CNN with a single inception module and overall depth of 5, performs well on both the training and test data. The architecture is kept simple and efficient yet achieves satisfactory results. The simple architecture implies a lower inference time. Moreover, the learning process also involves tuning several hyperparameters such as learning rate, weight decay (regularizer), and learning rate scheduler, which have a significant impact on the training process.

4 Experiments

4.1 Hardware and Experimental Set-up

In the current setup, we use two COTS MikroTik wAP 60Gx3 devices, one acting as AP and the other as a client. These devices implement the IEEE 802.11ad standard and operate at 60 GHz frequency. The MikroTik wAP 60Gx3 uses three of 6x6 antenna arrays arranged at different angles to provide a combined aperture of 180°. We follow the workof Blanco et al. [3] and install OpenWRT on these devices and modify device drivers to gain access to CSI. However, as of now, CSI is always extracted from the middle antenna array. Moreover, Qualcomm uses only 32 bits to define beam patterns of these devices and therefore, only 32 antennas can be addressed. The other four antennas are unused, and two of the remaining 32 only report noise [3]. The final CSI obtained corresponds to 30 antennas. Each CSI measurement contains an amplitude and phase value. However, phase measurements are noisy, and calibration is needed. Instead, in this work, we take the simplest approach and only use the amplitude part for further processing. To record the ground truth for poses, we use a Kinect V2, which can track the 3D coordinates of 25 body keypoints. We perform our experiments in a regular office space, shown in Figure 2. The room measures 3.3 m, 3.4 m and 3 m in length, width, and height, respectively. The yellow boxes indicate Mikrotik wAP 60Gx3 AP and client devices. The green box shows Kinect. A user performs a set of poses in between the two MikroTik devices. These poses are motivated by XR applications and state-of-the-art [2, 27]. The poses include Arms up, Left hand up, Right lean, Right hand up, Left lean, Empty, and Arms wide, shown in Figure 3.

Table 1. Accuracy of machine learning and deep learning-based methods

Users	KNN	SVM	Tree	1 FCN	4 FCN	CNN
1	88%	87.6%	75 %	86%	90%	**95.5%**
2	76%	87.6%	86%	81%	92%	**95.9%**
3	62.6%	69.7%	51%	79%	82%	**91%**
1+2+3	81.2%	77%	47%	66%	85%	**93.6%**

4.2 Data acquisition

Three users (two males and one female, with different body shapes and heights) performed a series of eight poses between the mmWave AP and the client. Each pose was held for a duration of 15 seconds, and the entire session involved around 20 rounds of performing the set of eight poses. By the end of the session, each pose had data corresponding to five minutes. Additionally, background data was collected during each session to remove any clutter or interference. To ensure a balanced evaluation, we split the data into a standard 75:25 ratio for training and test datasets.

4.3 Classification

Table 1 presents the results (overall accuracy) of machine learning and deep learning-based methods on the test data for users 1,2, and 3, and the final case (1+2+3) involves combining the data of 3 users. For deep learning, after several experiments, we used learning rate of 3e-4, weight decay of 8e-2, batch size of 64, a learning rate scheduler with a patience of 10 and 200 number of epochs. We can clearly see that, in general, deep learning-based methods 4 FCN and CNN outperform machine learning-based methods. The CNN always achieves the best accuracy for each case. Despite achieving very good accuracy, we can see in the case of user 3 there is a 4% drop in accuracy. This is because user

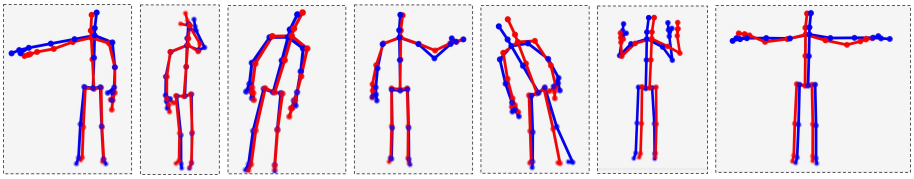

Figure 4. Pose Performance: Ground truth from Kinect (red) and pose from mmWave (blue)

3 has a considerable difference in body shape and height with respect to the other two users. mmWave signals are known to be particularly sensitive to these physical characteristics [1]. Also, beamforming employed to focus the signal to the intended receiver is significantly impacted by the user's body shape and orientation. Therefore, CNN encounters challenges in effectively extracting pose-specific features for user 3. Also, when the data is combined from all users, CNN achieves an accuracy of 93.6% which is basically a mean accuracy of 3 users. Moreover, we anticipate having multiple people in the room at the same time and increased number of poses may lead to a degraded performance which is unwanted, especially in XR applications, where accurate pose estimation is of utmost importance. To enhance the reliability and address the inherent subject and environmental variability issues associated with the mmWave channel, we have proposed the fusion of sEMG signals with mmWave sensing. Incorporating sEMG signals is beneficial due to their low susceptibility to environmental variations, thereby adding a layer of robustness to the system. Additionally, the utilization of transfer learning techniques can effectively mitigate subject-specific variations in sEMG signals by learning general morphology and templates. Transfer learning has been extensively explored and established as a well-regarded approach in this field.

4.4 Regressing pose skeletons

In contrast to the previous task, we demonstrate the capability to regress full body pose skeletons using mmWave CSI data. For this, we use the CNN described above with a depth of 7 to handle the complexity of the task. We first synchronize the data from the two sensors, ground truth from Kinect and CSI from mmWave Wi-Fi. Then, we train the CNN with labelled data of skeletons obtained from the Kinect sensor. During the testing phase, we solely rely on the CSI data from the mmWave for predictions. To evaluate the accuracy of our CNN model, we utilize the mean square error (MSE) metric, which quantifies the similarity between the predicted skeleton and the ground truth. Our results demonstrate a MSE of 0.0048, 0.0055, and 0.0058 for user 1, user 2, and user 3 respectively on the test data. These results further validate the performance of mmWave signals for pose estimation. To illustrate the performance, we present the test results corresponding to user 2 in Figure 4. The red colored skeletons in Figure 4 represent the ground truth obtained from the Kinect,

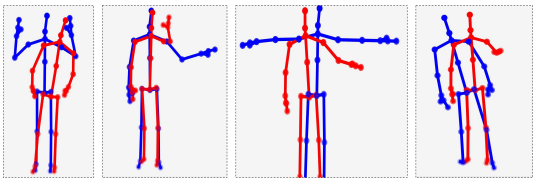

Figure 5. Pose Anomalies: Ground truth from Kinect (red) and pose from mmWave (blue)

while the blue ones correspond to the poses predicted from the mmWave Wi-Fi. It is evident that blue colored skeletons closely follow red ones, meaning the pose estimated is quite accurate. However, in some cases, we recorded highly inaccurate poses with a mean squared error (MSE) of 0.043. Figure 5 showcases the incorrect pose skeletons predicted by the neural network. These anomalies occur when the network struggles to accurately extract pose-specific features, leading to erroneous pose outputs. These results highlight limitations of the mmWave and the challenges involved in pose estimation. Even, in accurately predicted poses, there is room for improvement as can be seen in Figure 4 and this is where EMG signals come into the picture. These results from classification and regression, on one hand validate the performance of ISAC for pose estimation, on the other, highlight the challenges and need for additional modality to further improve the results.

5 Conclusion and Future Work

In this work, we presented a multi-modal approach for pose estimation leveraging ISAC and EMG signals. We proposed an architecture that benefits from the fusion of ISAC and EMG information in body pose estimation. As an initial experiment, we investigated the performance of ISAC for pose estimation considering both classification and regression tasks. From our experiments, we concluded that ISAC signals at mmWave can be very useful for pose estimation, however, need additional robustness to counter the limitations and to improve accuracy for accurate XR pose estimation. In future, we plan to integrate the EMG signals into the existing ISAC-based system to further improve accuracy.

6 Acknowledgments

This research is partly funded by the FWO WaveVR project (Grant number: G034322N).

References

[1] Tianyang Bai and Robert W Heath. 2014. Analysis of self-body blocking effects in millimeter wave cellular networks. In *2014 48th Asilomar conference on signals, systems and computers*. IEEE, 1921–1925.

[2] Nabeel Nisar Bhat, Rafael Berkvens, and Jeroen Famaey. 2023. Gesture Recognition with mmWave Wi-Fi Access Points: Lessons Learned. In *2023 IEEE 24th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 127–136.

[3] Alejandro Blanco, Pablo Jiménez Mateo, Francesco Gringoli, and Joerg Widmer. 2022. Augmenting mmWave localization accuracy through sub-6 GHz on off-the-shelf devices. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 477–490.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[5] Yuanhao Cui, Fan Liu, Xiaojun Jing, and Junsheng Mu. 2021. Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges. *IEEE Network* 35, 5 (2021), 158–167.

[6] Carlos De Lima, Didier Belot, Rafael Berkvens, Andre Bourdoux, Davide Dardari, Maxime Guillaud, Minna Isomursu, Elena-Simona Lohan, Yang Miao, Andre Noll Barreto, et al. 2021. Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges. *IEEE Access* 9 (2021), 26902–26925.

[7] Youjia Huang, Xingchen Yang, Yuefeng Li, Dalin Zhou, Keshi He, and Honghai Liu. 2017. Ultrasound-based sensing models for finger motion classification. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1395–1405.

[8] Shuo Jiang, Qinghua Gao, Huaiyang Liu, and Peter B Shull. 2020. A novel, co-located EMG-FMG-sensing wearable armband for hand gesture recognition. *Sensors and Actuators A: Physical* 301 (2020), 111738.

[9] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. 2020. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13289–13299.

[10] Peiqi Kang, Jinxuan Li, Shuo Jiang, and Peter B Shull. 2022. Reduce System Redundancy and Optimize Sensor Disposition for EMG-IMU Multi-modal Fusion Human-Machine Interfaces with XAI. *IEEE Transactions on Instrumentation and Measurement* (2022).

[11] Naveen Kumar Karnam, Shiv Ram Dubey, Anish Chand Turlapaty, and Balakrishna Gokaraju. 2022. EMGHandNet: A hybrid CNN and Bi-LSTM architecture for hand activity classification using surface EMG signals. *Biocybernetics and Biomedical Engineering* 42, 1 (2022), 325–340.

[12] EunSu Kim, JaeWook Shin, YongSung Kwon, and BumYong Park. 2023. EMG-Based Dynamic Hand Gesture Recognition Using Edge AI for Human–Robot Interaction. *Electronics* 12, 7 (2023), 1541.

[13] Toshiaki Koike-Akino, Pu Wang, Milutin Pajovic, Haijian Sun, and Philip V Orlik. 2020. Fingerprinting-based indoor localization with commercial MMWave Wi-Fi: A deep learning approach. *IEEE Access* 8 (2020), 84879–84892.

[14] Filip Lemic, James Martin, Christopher Yarp, Douglas Chan, Vlado Handziski, Robert Brodersen, Gerhard Fettweis, Adam Wolisz, and John Wawrzynek. 2016. Localization as a feature of mmWave communication. In *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 1033–1038.

[15] Iliana Loi, Angeliki Grammatikaki, Panagiotis Tsinganos, Efe Bozkir, Dimitris Ampeliotis, Konstantinos Moustakas, Enkelejda Kasneci, and Athanassios Skodras. 2022. Proportional Myoelectric Control in a Virtual Reality Environment. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.

[16] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. Signfi: Sign language recognition using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–21.

[17] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi. 2022. SHARP: Environment and Person Independent Activity Recognition with Commodity IEEE 802.11 Access Points. *IEEE Transactions on Mobile Computing* (2022).

[18] Sameera Palipana, David Rojas, Piyush Agrawal, and Dirk Pesch. 2018. FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–25.

[19] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. Gopose: 3D human pose estimation using Wi-Fi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.

[20] Francesco Restuccia. 2021. IEEE 802.11 bf: Toward ubiquitous Wi-Fi sensing. *arXiv preprint arXiv:2103.14918* (2021).

[21] Jakob Struye, Filip Lemic, and Jeroen Famaey. 2020. Towards ultra-low-latency mmwave Wi-Fi for multi-user interactive virtual reality. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[22] Afroza Sultana, Farruk Ahmed, and Md Shafiul Alam. 2022. A systematic review on surface electromyography-based classification system for identifying hand and finger movements. *Healthcare Analytics* (2022), 100126.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[24] Turker Tuncer, Sengul Dogan, and Abdulhamit Subasi. 2022. Novel finger movement classification method based on multi-centered binary pattern using surface electromyogram signals. *Biomedical Signal Processing and Control* 71 (2022), 103153.

[25] Bram van Berlo, Amany Elkelany, Tanir Ozcelebi, and Nirvana Meratnia. 2021. Millimeter wave sensing: A review of application pipelines and building blocks. *IEEE Sensors Journal* 21, 9 (2021), 10332–10368.

[26] Juan Pablo Vásconez, Lorena Isabel Barona López, Ángel Leonardo Valdivieso Caraguay, and Marco E Benalcázar. 2022. Hand Gesture Recognition Using EMG-IMU Signals and Deep Q-Networks. *Sensors* 22, 24 (2022), 9613.

[27] Jianyuan Yu, Pu Wang, Toshiaki Koike-Akino, Ye Wang, Philip V Orlik, and Haijian Sun. 2020. Human pose and seat occupancy classification with commercial MMWave Wi-Fi. In *2020 IEEE Globecom Workshops*. IEEE, 1–6.

[28] Tao Zhang, Tingyu Song, Daolin Chen, Tian Zhang, and Jie Zhuang. 2019. WiGrus: A Wi-Fi-based gesture recognition system using software-defined radio. *IEEE Access* 7 (2019), 131102–131113.

[29] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia* 19, 2 (2012), 4–10.

[30] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[31] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the 17th annual international conference on mobile systems, applications, and services*. 313–325.

[32] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. 2023. MetaFi++: Wi-Fi-Enabled Transformer-based Human Pose Estimation for Metaverse Avatar Simulation. *IEEE Internet of Things Journal* (2023).