Combining Vision and EMG-Based Hand Tracking for Extended Reality Musical Instruments*

Max Graf[1] and Mathieu Barthet[1]

Centre for Digital Music, Queen Mary University of London
{max.graf, m.barthet}@qmul.ac.uk

Abstract. Hand tracking is a critical component of natural user interactions in extended reality (XR) environments, including extended reality musical instruments (XRMIs). However, self-occlusion remains a significant challenge for vision-based hand tracking systems, leading to inaccurate results and degraded user experiences. In this paper, we propose a multimodal hand tracking system that combines vision-based hand tracking with surface electromyography (sEMG) data for finger joint angle estimation. We validate the effectiveness of our system through a series of hand pose tasks designed to cover a wide range of gestures, including those prone to self-occlusion. By comparing the performance of our multimodal system to a baseline vision-based tracking method, we demonstrate that our multimodal approach significantly improves tracking accuracy for several finger joints prone to self-occlusion. These findings suggest that our system has the potential to enhance XR experiences by providing more accurate and robust hand tracking, even in the presence of self-occlusion.

Keywords: Extended reality, extended reality musical instruments, hand tracking, surface electromyography, deep learning

1 Introduction

Extended reality (XR) is an umbrella term encompassing virtual, augmented and mixed reality (VR/AR/MR). In recent years, the increased popularity of XR technology has seen the establishment of extended reality musical instruments (*XRMIs*) as a research field [23]. Milgram et al. described the reality-virtuality continuum [21], along which digital applications can be placed. It stretches from real-world environments to fully virtual environments. Head-mounted XR devices bridge this continuum. They are capable of rendering three-dimensional imagery onto screens, removing the necessity for separate monitors or mobile displays, blending the real and virtual worlds together. The rapid development of XR technologies has opened up new possibilities for musical creation, performance, and interaction, with the emergence of various XR-based musical instruments and applications. Many XRMIs follow an embodied interaction paradigm. These instruments offer novel opportunities for artists to experiment with embodied interaction techniques, spatial sound design, and immersive performances, thus expanding the boundaries of traditional music making. XRMIs fall within the larger category

* This work was supported by the UKRI Centre for Doctoral Training in AI & Music [grant number EP/S022694/1].

of digital musical instruments (DMIs). [27] suggest that the control of DMIs can be made intimate (personal and familiar) by using appropriate control metaphors, low latency action-to-sound, and continuous gesture recognition. This study is part of a larger project that aims to support control intimacy in XRMIs.

Based on the current state of XR technology and prior work in XRMIs [3, 9], we highlight that gesture sensing errors on XR devices are a bottlenek for intimate musical control. Head-mounted XR devices (HMDs) rely on a set of sensors to record data and provide embodied control interfaces for users, e.g., head-tracking, hand tracking, and body pose detection. The transduction of these real-world sensor data to digital representations depends on computational methods. In this work we focus on the problem of hand tracking, more specifically, accurate tracking of finger joints. Hand-tracking algorithms often use visual information from camera sensors in conjunction with machine learning techniques, for example, in the Oculus Quest 2 device [11]. The accuracy of vision-based hand-tracking algorithms may be high [26], but current recognition rates do not reach 100%. Self-occlusion - the occlusion of finger joints by other parts of the hand - as well as challenging lighting situations lead to failure cases in vision-based tracking systems. Such error cases may produce instances of jitter, tracking loss, or glitches in the virtual representation of the hands, which can have detrimental effects on the usability and user experience in XRMIs, as shown in a previous study [9].

This work aims to address such sensing-related issues through the use of surface electromyography (sEMG) sensors. EMG sensors measure the electrical potential produced during muscle contractions in the body. Surface electromyograms can be obtained through electrodes that are positioned on the surface of the skin, above muscle tissue regions. We present an investigation into the potential of sEMG sensors and deep learning models to enhance hand-tracking accuracy in XRMIs. Our approach combines sEMG data and vision-based tracking methods to address sensing-related issues commonly encountered in XRMI performance. Thereby, we aim to improve the tracking accuracy and responsiveness of XR musical instruments, especially in situations where vision-based tracking falls short.

The scope of this paper is limited to the exploration of sEMG and deep learning techniques for hand-tracking in XR musical instruments. While our findings may have broader applications in other areas of XR interaction, the primary focus is on the improvement of XRMI design and user experience. Through our work, we aim to contribute to the ongoing development of more accurate, intuitive, and expressive extended reality musical instruments.

2 Background

The development of XRMIs has attracted growing interest as VR, AR, and MR technologies continue to advance. Early studies in this domain focused on the design and evaluation of virtual interfaces for musical performance and interaction [19, 23, 8, 22]. Several works investigated user experience [6, 5], interaction techniques [2] and collaborative music making [20, 10]. More recent studies have explored the creation of novel instruments and control schemes [5, 3, 4, 9]. While there is no gold standard for XRMI design, many XRMIs rely on hand-tracking to facilitate embodied interaction with the

instrument [22, 8, 4, 9]. Various vision-based tracking methods are employed, including depth-sensing cameras [22, 8, 4], and machine learning-based approaches [11]. Despite the progress in hand tracking research, limitations such as occlusion, lighting issues, and computational complexity continue to pose challenges for hand-controlled XRMI applications.

Surface electromyography (sEMG) has emerged as a promising alternative to vision-based tracking methods for capturing user input in various applications, including XR. Several studies have explored the use of sEMG data and deep learning architectures for hand gesture recognition [18, 17, 1, 16]. These works have reported promising results, highlighting the potential of employing sEMG and deep learning models for precise finger movement estimation. However, some of these works depend on complex tracking setups [1] or leverage low-resolution tracking data for training [16].

A notable limitation of these studies is the lack of shared training data and code, hindering the reproducibility and comparability of the results across different research efforts. Several datasets on the topic of finger joint angle estimation through sEMG data have been published. However, they either require specialised sEMG measuring equipment [13, 15, 14], making the reproduction of results an expensive endeavour, or introduce temporal biases into the dataset due to lack of synchronisation during the recording procedure [12]. The absence of implementation details ([18, 17, 1]) makes it difficult for other researchers to build upon these works, potentially slowing down the progress in the field.

The potential benefits of integrating sEMG data and deep learning models with vision-based tracking methods has not been thoroughly investigated in the context of XRMIs. While machine learning has found its way into the NIME community [24], the use of machine learning approaches to improve XRMI control remains an under-explored area. This study aims to develop a multimodal hand tracking approach that leverages sEMG data, deep learning models, and vision-based tracking techniques. We share the training data and code[1], fostering further research and innovation in the domain of sEMG-based neural interfaces.

3 Deep Learning Model for Finger Joint Angle Estimation

We have developed a software pipeline for data collection, feature extraction and modelling of sEMG data. We focus on eight finger joints that are prone to self-occlusion: the metacarpophalangeal and proximal interphalangeal joints of the index, middle, ring and pinky fingers. Specifically, we want to model the rotations of the finger bones connected to these joints relative to the hand. The thumb is excluded from our investigation. Modelling thumb rotations with sEMG data is a hard problem, since the majority of muscles related to thumb movements are located in the hand, rather than the forearm. With that in mind, the goal of the model is to estimate the eight finger joint angles from a window of sEMG data.

[1] https://github.com/maxgraf96/sEMG-myo-unity

3.1 Data Collection

We collect surface EMG signal measurements using the Thalmic Labs Myo armband[2] and vision-based hand tracking data from the Oculus Quest 2 XR headset. Both devices are employed simultaneously to capture the muscle activity and finger joint rotations, respectively. This approach allows users to capture data without the need for external tracking devices.

The Myo armband is a non-invasive wearable device that features eight sEMG sensors. The armband is worn on the forearm, with the sensors evenly distributed around the circumference of the arm, allowing it to capture the activity of the forearm muscles during finger movements. We obtain sEMG data from the Myo armband using the *PyoMyo* Python framework [25], extracting rectified and smoothed signals at a sampling frequency of 50Hz. The Oculus Quest 2 XR headset is equipped with four monochrome cameras that provide a wide field of view, enabling it to capture hand positions and movements. The built-in hand tracking algorithm [11] processes the camera data and estimates the 3D rotations of the user's hand joints in real time. In our system, we sample the hand joint rotations from the XR device at 50 Hz and synchronize them with the sEMG data from the Myo armband.

One researcher recorded hand gestures and movements in a controlled environment, with a focus on gestures relevant to XRMI interaction. This study should be seen as a proof-of-concept for our methodology. Hence, for this study, we focused on data from the right hand only. The gestures included various finger flexions and extensions, as well as combinations of multiple finger movements. They were performed at different speeds, forces, and orientations. We conducted three data collection sessions across three days to account for the natural variability in sEMG readings, ensuring a more robust dataset. The armband was fitted on the right forearm, covering the right flexor carpi radialis, flexor digitorum superficialis and the right extensor carpi radialis longus, as described in [7]. During the data collection process, the XR headset was strategically positioned in diverse locations and orientations to minimise self-occlusion of the hand. Data collection sessions lasted between ten and fifteen minutes, resulting in a substantial amount of synchronized sEMG and hand tracking data.

3.2 Feature Extraction

Figure 1 shows the data flow in our pipeline. The selection of features was informed by both the literature and a series of experiments.We use Python to extract both time domain and frequency domain features from the sEMG data. The pipeline takes 2D windows of sEMG samples, with N number of samples and C channels. We then compute the following features per channel: in the time domain, mean absolute value (MAV), root mean square (RMS), and variance (VAR); in the frequency domain, median frequency bin (MDF), mean frequency bin (MNF), and peak frequency bin (PF). Additionally, wavelet coefficients at the fourth level are extracted to provide further information about the signal's characteristics, as reported in [1]. Wavelet analysis provides a multiresolution representation of the sEMG signal, capturing both the time and frequency

[2] `https://xinreality.com/wiki/Myo`

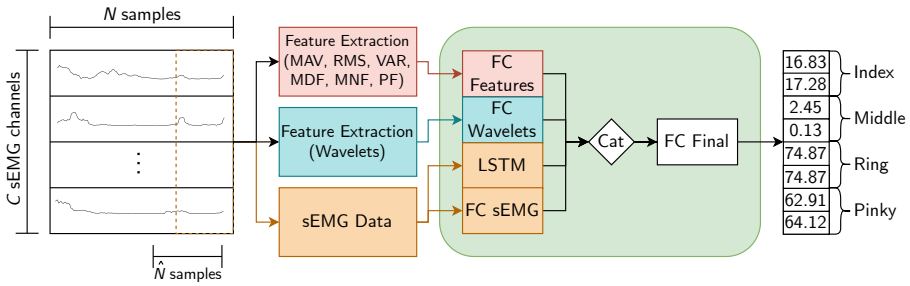

Fig. 1: Data preprocessing, feature extraction, and model pipeline

characteristics of the data. This comprehensive feature representation aims to capture essential information from the sEMG signals both locally and globally, enabling holistic representation of the data.

3.3 Model Architecture

The model architecture is a combination of a Long Short-Term Memory (LSTM) network and multiple separate sets of fully connected (FC) layers, aiming to capture both general trends in the sEMG signal data provided by the features and high-frequency characteristics of the signal. Formally, our model learns a mapping

$$F : \mathbb{R}^{N \times C} \rightarrow \mathbb{R}^{M} \tag{1}$$

where N denotes the number of sEMG samples, C represents the number of sEMG channels and M denotes the number of predicted joint angle values at every time step. More accurately, we learn a mapping

$$F(\mathbf{s}) = \phi_{\text{final}}\Big(\phi_{\text{lstm}}(\mathbf{s}_{\hat{N}}) \oplus \phi_{\text{feat}}(\psi_{\text{time-freq}}(\mathbf{s}_N)) \oplus \phi_{\text{wav}}(\psi_{\text{wavelet}}(\mathbf{s}_N)) \oplus \phi_{\text{filt}}(\mathbf{s}_{\hat{N}})\Big) \tag{2}$$

where

- $F(\mathbf{s})$ represents the mapping function that takes $N \times C$ sEMG samples ($\mathbf{s}$) and outputs M finger joint angles.
- $\mathbf{s}_N$ denotes all $N \times C$ sEMG samples.
- $\mathbf{s}_{\hat{N}}$ denotes the last $\hat{N} \times C$ sEMG samples in the data point.
- ϕ_{lstm} represent the LSTM layers, and ϕ_{feat}, ϕ_{wav} and ϕ_{filt} represent the fully connected layers processing time/frequency domain features, wavelet features and filtered EMG data respectively.
- ϕ_{final} denotes the final set of fully connected layers.
- $\psi_{\text{time-freq}}$ and ψ_{wavelet} denote the feature extraction functions for time-frequency domain features and wavelet features, respectively.
- $\oplus$ represents the tensor concatenation operation.

The LSTM network, a type of recurrent neural network, operates on the subset $\mathbf{s}_{\hat{N}}$, which contains the last $\hat{N}$ samples of the EMG data, capturing the temporal dependencies within the most recent portion of the sEMG signal. LSTMs can effectively learn medium-to-long-range dependencies in time-series data and retain information across multiple time steps. Additionally, $\mathbf{s}_{\hat{N}}$ is fed into a separate fully connected layer, which was empirically found to improve the model's performance. The time and frequency domain features are computed over all N samples and processed by a set of fully connected layers. These layers are designed to extract higher-level representations from the sEMG features, capturing general patterns and trends in the data. The wavelet features are fed into a separate set of fully connected layers, allowing the model to learn distinct patterns associated with the wavelet coefficients. This additional information can help the model to better discriminate between different types of hand movements and gestures.

The outputs of the LSTM layers and the three sets of fully connected layers (time-frequency domain features, wavelet features, and $\hat{N}$ sEMG samples) are concatenated and passed to a final set of fully connected layers. This combination of network components aims to capture a comprehensive representation of the sEMG signal, taking into account both general trends and high-frequency changes. The final output of the model is an estimation of the eight finger joint angles described above.

3.4 Model Implementation and Training

We selected $N = 150$, which gives a sampling window size of three seconds. Related works use window sizes of five seconds [18, 17]. During our experiments with the model architecture, we found that a smaller window size of 150 samples did not reduce the quality of the predicted finger joint angles, while yielding a performance gain in the data processing pipeline. $\hat{N} = 50$ was selected as a trade-off between LSTM accuracy and performance. Higher values for $\hat{N}$ produced slightly better results, but incurred a performance degradation, slowing down the model at inference time.

The collected data comprises approximately 80000 sEMG samples with corresponding finger joint angle measurements. The sEMG samples were separated into training and validation sets using a 90/10 ratio. Our deep learning model is built using the PyTorch framework. The model was trained on a single NVidia RTX 2080 Ti GPU for approximately 500000 steps with a batch size of 256. The mean squared error function was employed as a loss metric. We applied an exponentially decreasing learning rate, starting at 0.003 and reducing to 0.0003 over the first 10000 steps. During training, we tracked the mean average difference between the predicted joint angles and the ground truth angles for both training and validation sets. Our criterion for stopping the training procedure was the moment of obtaining a mean joint angle difference value of less than 1° across the validation set.

4 Multimodal XR Hand Tracking with sEMG and Vision-Based Tracking

In this section, we present our approach to multimodal XR hand tracking by combining sEMG and vision-based tracking techniques. In our system, the vision-based tracking

data provides information about the overall hand position and orientation in 3D space, while the sEMG-based model produces granular information about individual finger joint rotations. The sEMG data are continuously sampled, preprocessed, and passed to the trained deep learning model to estimate the eight finger joint angles. The hand position and orientation data are combined with the estimated finger joint angles to generate a complete hand pose representation.

The deep learning model is optimized for real-time performance, ensuring that the sEMG data can be processed with minimal latency. Our system operates at 50Hz, which incurs a latency of 20ms. For this work, data processing and model inference took place on a consumer notebook. The notebook concurrently runs a python server, responsible for sEMG data aggregation, preprocessing and model inference, and a 3D XR environment on the Unity platform. At every time step, the estimated finger joint angles are transferred from Python to Unity through the low-latency ZeroMQ framework[3]. A video demonstration of our system is available online[4]. It shows a side-by-side comparison of the vision-based tracking system and our multimodal approach.

4.1 Evaluation

To validate the effectiveness of our multimodal hand tracking system, we conducted an experimental evaluation, which compared the multimodal tracking to the baseline vision-based hand tracking system provided by the XR device. We simultaneously collected tracking data from the vision-based tracking system and the multimodal tracking system. A Leap Motion sensor was used to acquire ground truth labels for finger joint angles. The Leap Motion is a high-precision vision-based hand tracking device that captures finger joint angles and hand position in 3D space. The ground truth labels were then compared to the results from both the vision-based and the multimodal tracking system.

The experimental setup included a series of hand pose tasks, designed to cover a wide range of hand movements, including gestures prone to occlusion. The tasks were selected with regard to their utility in playing a keyboard-inspired XRMI. The tasks were performed while wearing the Oculus Quest 2 headset and the Myo armband. The Leap Motion sensor was placed on a table to record ground truth data. We recorded tasks under two conditions: 1) Full view of the hand - here, the XR headset was positioned at a 50cm distance, 45° above the hand to ensure optimal visual tracking conditions. 2) Self-occlusion of the hand - in this condition, the distance was kept identical, but the angle of the XR headset was lowered, such that the back of the hand occluded the fingers. Figure 2 illustrates the six hand pose tasks devised: (i) extending all fingers and making a fist; individual flexion and extension of the (ii) index, (iii) middle, (iv) ring and (v) pinky fingers; (vi) sequential flexion and extension of pinky, ring, middle and index fingers (similar to the gesture of drumming on a table while waiting for something). Each task involves the execution of the gesture at three different speeds: slow, over approximately two seconds, moderate (one second), and fast (half a second). To account for variability

[3] https://zeromq.org/

[4] https://www.youtube.com/watch?v=ivl2g2t2oaI

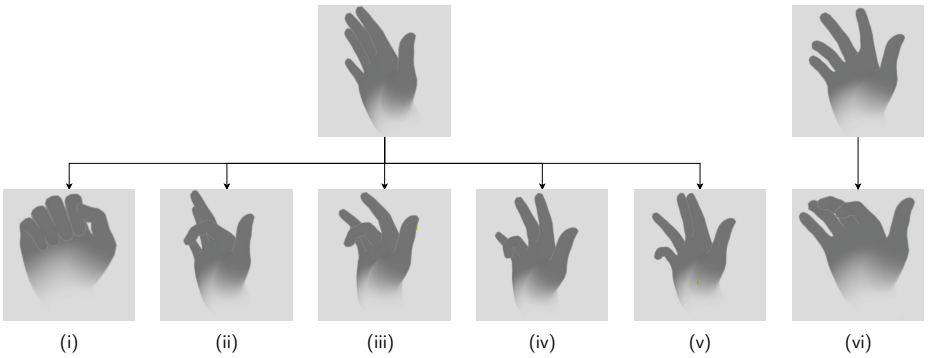

Fig. 2: Finger movements in the six hand pose tasks

in the sEMG measurements, all tasks were executed three times, over two days, under identical lighting conditions.

To measure the degree of finger occlusion, we integrated a ray-casting system with the XR application. We cast rays from the XR headset's 3D position to the eight finger bones whose rotations we measured every time a sample was taken. Rays intersecting other parts of the hand, e.g., the back of the hand, were used to mark the respective finger bones as occluded. This allowed us to quantify the level of occlusion per finger for every recording.

4.2 Analysis Methods & Results

To evaluate the performance of the vision-based and multimodal tracking systems across every task, we obtained matrices of difference values between the estimated joint angles and the ground truth angles for both systems at every time step. The matrices were aggregated across the three sessions. We assessed the normality of the difference matrices using the Shapiro-Wilk test for each of the six hand pose tasks. We then applied the Wilcoxon signed-rank test to see whether there was a significant difference between the results produced by the vision-based and multimodal tracking systems.

Across tasks, the results of the Shapiro-Wilk test showed p-values $< .001$, indicating that the data in the difference value matrices did not fit a normal distribution. Therefore, we proceeded with the non-parametric Wilcoxon signed-rank test for further analysis. Figure 3 shows the results obtained from the mean joint angle differences across all finger joints per task, for both occlusion conditions (full view and occluded). Under the occluded condition, our model produces significantly lower deviations from the ground truth data across all tasks, compared to the vision-based tracking system. On average, it improves the finger joint angle tracking accuracy by five to 15 degrees across all fingers.

Table 1 shows the results obtained from the Wilcoxon signed-rank tests, aggregated across all eight tracked joints per task for both conditions. The p-values indicate significant differences between the difference value matrices. Additionally, the table lists the average finger occlusion results obtained through the raycasting occlusion measure

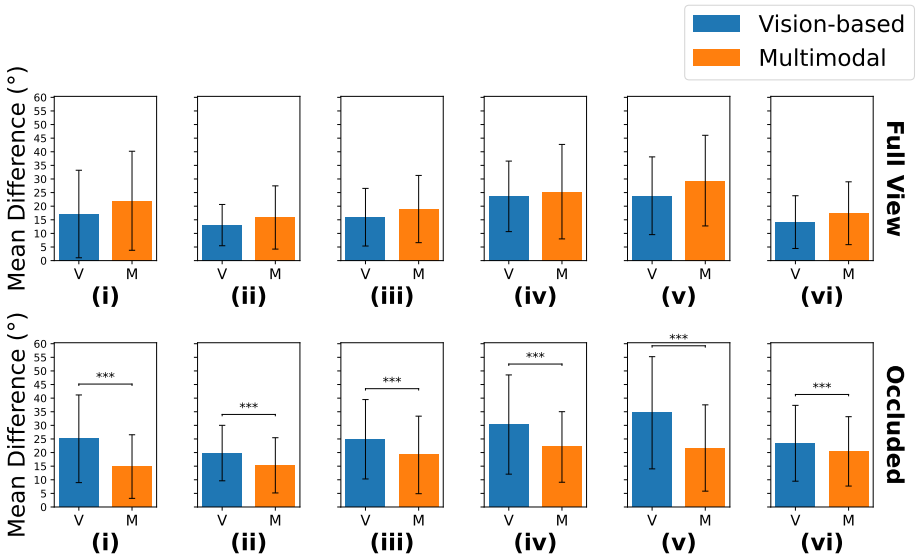

Fig. 3: Average deviation in degrees between the finger joint angles generated by the vision-based (V) and multimodal (M) tracking systems and the ground truth data for each task. Error bars show standard deviation; three asterisks indicate a significant difference between the V and M values ($p < .001$)

described above. The occlusion results describe the mean portion per task, in which the fingers were occluded by another part of the hand.

5 Discussion

The results of the evaluation showed that the multimodal hand tracking system outperformed the pure vision-based hand tracking system across all tasks under the occluded condition. These findings support our hypothesis that the integration of sEMG-based finger joint angle estimation can help overcome occlusion-related limitations in vision-based hand tracking, resulting in more accurate and reliable XRMI interactions. Under the full view condition, the vision-based hand tracking produced fewer errors in all tasks. This was expected, as the vision-based hand tracking system operates optimally under full view of the hand.

Despite the promising results, our study has several limitations. Due to the nature of sEMG data, the tracking performance of the multimodal approach is unlikely to extend to other users without fine-tuning the deep learning model. Surface EMG signals differ substantially between individuals and can be influenced by factors such as muscle fatigue, electrode placement, and individual anatomical differences. It will be valuable to investigate the system's performance across different users and under varying conditions. The identification of sEMG data representations that allow for generalisation under consideration of these factors without requiring extensive amounts of data is still

Table 1: P-values and average occlusion measurement results across tasks under both conditions

	Full view		**Occluded**	
Tasks	P-value	Occlusion (%)	P-value	Occlusion (%)
(i)	1.0	16.78	$<.001$	93.00
(ii)	1.0	2.65	$<.001$	70.58
(iii)	1.0	15.25	$<.001$	63.11
(iv)	1.0	16.78	$<.001$	53.02
(v)	1.0	27.94	$<.001$	78.82
(vi)	1.0	10.36	$<.001$	59.48

an ongoing research topic. However, our work allows XR users with access to sEMG devices to train their own models using our pipeline and code.
The performance of our multimodal hand tracking system was evaluated using a single type of XR headset and sEMG armband. Future research should explore more complex occlusion scenarios, as well as test the system's performance across different hardware setups and sEMG devices, to better understand the generalisability of our findings.
With that in mind, we see numerous avenues for further research. The integration of additional tracking modalities, such as depth sensing or inertial measurement units (IMUs), could further enhance the robustness and accuracy of the multimodal hand tracking system by enabling stronger representations of the underlying data. A future study will explore the impact of our multimodal hand tracking system on usability, user experience and task performance in XRMI interactions, and provide insights into the practical implications of our findings. By conducting user studies with tasks that require precise hand movements and are susceptible to occlusion, the benefits of our system for real-world applications could be better understood.
Our study provides evidence that the combination of vision-based tracking and sEMG-based finger joint angle estimation can effectively address occlusion issues in hand tracking for XRMI interactions. The findings suggest that the multimodal hand tracking system has the potential to enhance user experiences and enable more immersive and natural interactions in virtual environments.

6 Conclusion

In this paper, we introduced a multimodal hand tracking system designed to address occlusion issues in XRMI interactions by combining vision-based tracking with sEMG-based finger joint angle estimation. The goal of this study was to demonstrate the potential of our proposed system to improve hand tracking accuracy and robustness, even when the hand is partially occluded.

While our results show promise, the experimental setup was relatively simple, and further research should explore more complex scenarios and investigate the system's performance across different hardware and user conditions. Future work could also in-

tegrate additional tracking modalities and machine learning techniques to enhance the robustness and accuracy of the system.

Our multimodal hand tracking system demonstrates the potential to improve XRMI interactions by addressing occlusion issues in vision-based hand tracking. As XR technologies continue to evolve, the integration of complementary tracking modalities, such as sEMG and vision-based tracking, will likely play a crucial role in enhancing user experiences and enabling more immersive and natural interactions in virtual environments.

References

[1] C. Avian et al. "Estimating Finger Joint Angles on Surface EMG Using Manifold Learning and Long Short-Term Memory with Attention Mechanism". In: *Biomedical Signal Processing and Control* 71 (Jan. 1, 2022), p. 103099.

[2] F. Berthaut. "3D Interaction Techniques for Musical Expression". In: *Journal of New Music Research* 49.1 (Jan. 1, 2020), pp. 60–72.

[3] S. Bilbow. "Developing Multisensory Augmented Reality As A Medium For Computational Artists". In: *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 72. New York, NY, USA: Association for Computing Machinery, Feb. 14, 2021, pp. 1–7.

[4] S. Bilbow. "Evaluating Polaris˜ - An Audiovisual Augmented Reality Experience Built on Open-Source Hardware and Software". In: *NIME 2022*. June 16, 2022.

[5] A. Çamcı, M. Vilaplana, and R. Wang. "Exploring the Affordances of VR for Musical Interaction Design with VIMEs". In: *Proceedings of the International Conference on New Interfaces for Musical Expression*. NIME. Birmingham, UK, June 1, 2020, pp. 121–126.

[6] T. Deacon, T. Stockman, and M. Barthet. "User Experience in an Interactive Music Virtual Reality System: An Exploratory Study". In: *Bridging People and Sound*. Ed. by M. Aramaki, R. Kronland-Martinet, and S. Ystad. Vol. 10525. Cham: Springer International Publishing, 2017, pp. 192–216.

[7] *Delsys EMG Sensor Placement Technical Note 101*. Accessed on 15.08.2023.

[8] J. Fillwalk. "ChromaChord: A Virtual Musical Instrument". In: *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. 2015 IEEE Symposium on 3D User Interfaces (3DUI). Mar. 2015, pp. 201–202.

[9] M. Graf and M. Barthet. "Mixed Reality Musical Interface: Exploring Ergonomics and Adaptive Hand Pose Recognition for Gestural Control". In: *International Conference on New Interfaces for Musical Expression*. NIME 2022. June 28, 2022.

[10] R. Hamilton and C. Platz. "Gesture-Based Collaborative Virtual Reality Performance in Carillon". In: *Proceedings of the 2016 International Computer Music Conference*. International Computer Music Conference. Utrecht, Netherlands, 2016, p. 5.

[11] S. Han et al. "MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality". In: *ACM Transactions on Graphics* 39.4 (Aug. 12, 2020), 87:87:1–87:87:13.

[12] X. Hu et al. “Finger Movement Recognition via High-Density Electromyography of Intrinsic and Extrinsic Hand Muscles”. In: *Scientific Data* 9.1 (1 June 29, 2022), p. 373.

[13] N. J. Jarque-Bou et al. “A Calibrated Database of Kinematics and EMG of the Forearm and Hand during Activities of Daily Living”. In: *Scientific Data* 6.1 (1 Nov. 11, 2019), p. 270.

[14] N. Jiang. *Gesture Recognition and Biometrics Electromyography (GRABMyo) Dataset*. Jan. 4, 2022.

[15] P. Kaczmarek, T. Mańkowski, and J. Tomczyński. “putEMG—A Surface Electromyography Hand Gesture Recognition Dataset”. In: *Sensors* 19.16 (16 Jan. 2019), p. 3548.

[16] H. Lee, D. Kim, and Y.-L. Park. “Explainable Deep Learning Model for EMG-Based Finger Angle Estimation Using Attention”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 1877–1886.

[17] Y. Liu, C. Lin, and Z. Li. “WR-Hand: Wearable Armband Can Track User’s Hand”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (Sept. 14, 2021), 118:1–118:27.

[18] Y. Liu, S. Zhang, and M. Gowda. “NeuroPose: 3D Hand Pose Tracking Using EMG Wearables”. In: *Proceedings of the Web Conference 2021*. WWW ’21. New York, NY, USA: Association for Computing Machinery, June 3, 2021, pp. 1471–1482.

[19] T. Mäki-Patola et al. “Experiments with Virtual Reality Instruments”. In: *Proceedings of the 2005 Conference on New Interfaces for Musical Expression*. NIME ’05. SGP: National University of Singapore, May 1, 2005, pp. 11–16.

[20] L. Men and N. Bryan-Kinns. “LeMo: Supporting Collaborative Music Making in Virtual Reality”. In: *IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*. Apr. 5, 2018.

[21] P. Milgram et al. “Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum”. In: *Telemanipulator and Telepresence Technologies* 2351 (Jan. 1, 1994).

[22] A. G. Moore et al. “Wedge: A Musical Interface for Building and Playing Composition-Appropriate Immersive Environments”. In: *2015 IEEE Symposium on 3D User Interfaces (3DUI)*. 2015 IEEE Symposium on 3D User Interfaces (3DUI). Arles, France: IEEE, Mar. 2015, pp. 205–206.

[23] S. Serafin et al. “Virtual Reality Musical Instruments: Guidelines for Multisensory Interaction Design”. In: *Proceedings of the Audio Mostly 2016*. AM ’16: Audio Mostly 2016. Norrköping Sweden: ACM, Oct. 4, 2016, pp. 266–271.

[24] Théo Jourdan and Baptiste Caramiaux. “Machine Learning for Musical Expression: A Systematic Literature Review”. In: NIME. 2023.

[25] P. Walkington. *PyoMyo*. Version 0.0.5. Nov. 2021.

[26] F. Weichert et al. “Analysis of the Accuracy and Robustness of the Leap Motion Controller”. In: *Sensors* 13.5 (5 May 2013), pp. 6380–6393.

[27] D. Wessel and M. Wright. “Problems and Prospects for Intimate Musical Control of Computers”. In: *Computer Music Journal* 26.3 (Sept. 2002), pp. 11–22.