Leveraging Ultrasound Sensing for Virtual Object Manipulation in Immersive Environments

Keshav Bimbraw
Robotics Engineering
Worcester Polytechnic Institute
Worcester, USA
kbimbraw@wpi.edu

Jack Rothenberg
Robotics Engineering
Biomedical Engineering
Worcester Polytechnic Institute
Worcester, USA
jarothenberg@wpi.edu

Haichong Zhang
Robotics Engineering
Biomedical Engineering
Worcester Polytechnic Institute
Worcester, USA
hzhang10@wpi.edu

***Abstract*—Hand gesture recognition is a fundamental component of intuitive and immersive user interfaces in virtual reality (VR) applications. This paper presents a data-driven approach utilizing ultrasound data and deep learning techniques for hand gesture recognition in VR interfaces. The proposed methodology involves acquiring data from a subject, training a model using the acquired data, and evaluating the model's performance on both the training data and during real-time inference. The evaluation metrics primarily focus on accuracy percentage: measuring the classifier's performance in correctly classifying hand gestures. 4 hand gestures were primarily considered for the study and demonstration. For offline evaluation with a 20% test-train split, an accuracy percentage of 91% was observed. For online evaluation, an accuracy percentage of 92% was achieved. Results on the classification of 7 hand gestures were also analyzed for both online and online evaluation due to the promising results from the 4 gesture classification. The latency of the pipeline, from ultrasound data acquisition using screenshots to sending commands for VR object manipulation, was measured to be 59.48 milliseconds. The results demonstrate the effectiveness of the approach in accurately recognizing hand gestures, both during training and in real-time inference. We supplement our results with a video of the forearm ultrasound data being used to control a custom-designed VR game in a low-latency fashion. This research provides valuable insights into the performance and applicability of ultrasound-based hand gesture recognition techniques in VR interfaces. By employing deep learning and leveraging real-time data acquisition, this approach paves the way for intuitive and immersive interactions in various VR applications. The study contributes to the field of body sensor networks, highlighting the potential of forearm ultrasound based data-driven techniques for enhancing user interaction and immersion in VR environments.**

***Index Terms*—Ultrasound, Machine Learning with Biosignal Processing, Body Sensor Networks, Virtual Reality (VR)**

I. Introduction

The field of human-computer interaction has seen increasing interest in developing smart and intuitive upper limb interaction for augmented/virtual reality (AR/VR) applications [1]. The human hands play a crucial role in these interactions, necessitating accurate estimation of different hand gestures. Various technologies, such as surface electromyography (sEMG), force myography (FMG), vision-based approaches, resistive hand gloves, depth information based approaches and WiFi sensing have been previously investigated [2]. However, these methods have limitations. Placing sensors directly on the fingers can restrict hand mobility, while vision-based methods are sensitive to various factors such as poor resolution, frame rate, illumination conditions, and occlusions [3], [4].

Biosignals from the forearm offer a viable alternative for comprehending hand movements while ensuring unhindered user experience and mitigating the challenges associated with complex and cumbersome hand motions [5]. sEMG has been used for the recognition of hand gestures for controlling AR/VR interfaces [6]. sEMG's sensitivity to muscle fatigue, especially with prolonged muscle movements can hinder natural control of AR/VR interfaces. Additionally, signal contamination from motion artifacts, electromagnetic and environmental interference can introduce variations in signal properties, impeding its seamless adoption [7]. To overcome these challenges, ultrasound imaging of the forearm, known as sonomyography, has emerged as an alternative sensing modality [8]–[12]. Sonomyography provides a 2-dimensional visualization of the musculoskeletal structure of the forearm, enabling the identification of hand gestures and finger movements through image processing and classification algorithms. Studies have demonstrated the classification of multiple hand motions with high accuracy using ultrasound data [8], [10]–[12]. Researchers have also explored the combination of ultrasound imaging with deep learning techniques. For instance, an average classification accuracy of 83% has been reported for offline analysis for 11 hand gestures using a convolutional neural network (CNN) [12]. Comparisons between ultrasound and sEMG data for finger motion classification have shown the potential of ultrasound in capturing spatial features from B-mode ultrasound images [13].

With forearm ultrasound's capability to be used for hand gesture classification, it is important to demonstrate its efficacy for use as a modality for controlling virtual objects in AR/VR interfaces. The proposed pipeline offers a solution to utilizing a wireless ultrasound probe for manipulating virtual objects leveraging a deep learning based approach for hand gesture classification. Through evaluation, both offline and online, the effectiveness of the approach is demonstrated. The results are supplemented by demonstration videos of the low-latency pipeline for virtual object manipulation. By harnessing the

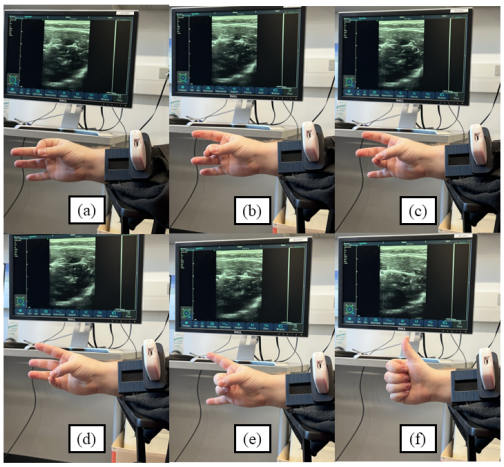


Fig. 1. The different hand gestures and their corresponding ultrasound images (except open hand): (a) Index Pinch, (b) Middle Pinch, (c) Ring Pinch, (d) Pinky Pinch, (e) Hand Horns, and (f) Thumbs Up.

power of ultrasound-based hand gesture recognition techniques and real-time data processing, this study paves the way for improved user experiences and opens new avenues for the development of sophisticated VR interfaces.

II. Materials and Methods

Forearm ultrasound data, captured using a wireless ultrasound probe, was seamlessly streamed to a Windows 10 system with an NVIDIA GeForce RTX 2070 SUPER and an AMD Ryzen 7 2700X eight-core processor. This data was used for a comprehensive evaluation of the system's performance in both offline and online settings. The pipeline's capabilities were demonstrated by the subject's manipulation of virtual objects in an immersive environment. The system components are shown in Figure 2.

A. Forearm ultrasound based gesture estimation

To acquire ultrasound data, a SonoQue L5 linear probe was used. The data was acquired from one subject (IRB-23-0634). The ultrasound probe was strapped to the subject's forearm using a custom-designed 3D-printed wearable armband.

1) Hand gestures: Two studies were conducted for this research. One focused on a total of 4 hand gestures which included open hand, index pinch, hand horns, and thumbs-up gestures. For the second study, a total of 7 hand gestures were used. These included the gestures in the 4-hand gesture study, in addition to the middle pinch, ring pinch, and pinky pinch (shown in Figure 1). The 4 hand gestures were mapped to positive X, Y, and Z in the VR space. The 7-hand gestures were mapped to the 3-D movements of an object in the VR space (positive and negative X, Y, and Z). The open hand gesture for both of these led to no movement in the VR space. The subject alternated between the open hand and the rest of the hand gestures.

2) Data acquisition: The data from the probe was streamed to a Windows system. A Python script was developed to take screenshots of the ultrasound data and save it in a folder, similar to the one described in [14]. These images measured 640x640 pixels. During training, the subject switched between the different hand movements, using audio signals as cues. This audio-based cue is based on [11], [12]. During real-time streaming for demonstration, the ultrasound image screenshots were saved and loaded as numpy arrays to be used as inputs to the deep learning models for estimation. The ultrasound data were acquired at an average frame rate of $\sim$10 Hz. A total of 2400 frames and 4800 frames were acquired for the 4-hand gesture and 7-hand gesture classification studies respectively.

3) Data processing: The ultrasound data and corresponding labels were split into testing and training sets without shuffling with a test-train split of 20%. For the 4-hand gesture classification study, this led to 1920 training images and 480 testing images. For the 7-hand gesture classification study, this led to 3840 training images and 960 testing images. A CNN previously proven to work well on hand movement estimation [12] was used to train a model using Tensorflow (`https://www.tensorflow.org/`). Adam optimizer and sparse categorical cross-entropy loss were utilized. A batch size of 5 and 25 epochs were chosen for evaluation.

4) Evaluation metrics: For multiclass classification, accuracy is the fraction of correct classifications over the total number of classificatiton. Accuracy percentage was used as the metric to evaluate the performance of the trained model for both offline and online estimation. The equation for accuracy percentage (Acc) is described in equation 1.

$$Acc = \frac{CC}{TC} * 100 \tag{1}$$

where, CC is the number of correct classifications and TC is the number of total classifications.

B. VR interfacing

The experimental setup incorporated a virtual reality (VR) component alongside the forearm ultrasound data acquisition and deep learning gesture estimation. A Meta Quest 2 headset

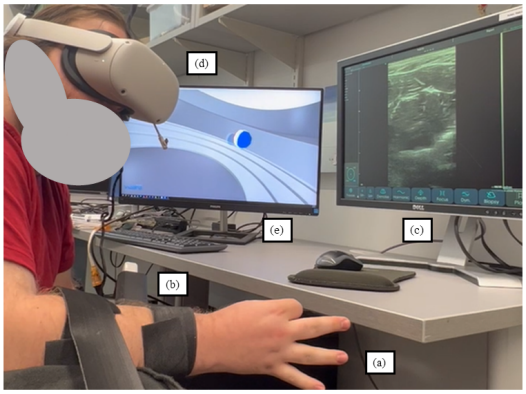


Fig. 2. The system setup: (a) Hand gesture, (b) Ultrasound Probe, (c) Desktop with Ultrasound Image, (d) VR Headset, (e) VR game visualization.

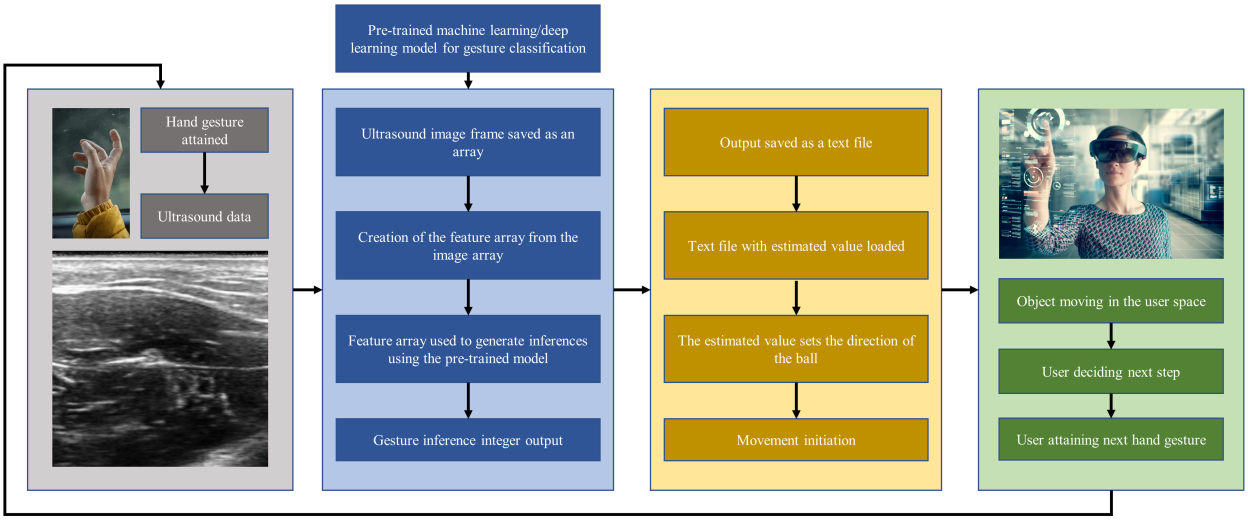


Fig. 3. Pipeline for virtual object manipulation using ultrasound data from the forearm.

was used for interacting with the VR environment designed using Vizard (`https://www.worldviz.com/`). It was connected to the Windows machine via a USB-C cable.

1) Custom designed VR game: To showcase the efficacy of the developed system, an interactive game was designed. In this game, the player maneuvers a ball in all six cardinal directions to align it with a translucent target ball. Each hand gesture for both the 4-hand gesture and the 7-hand gesture classification studies corresponded to a specific direction for ball movement. For the 4-hand gesture study (gesture, movement): index-pinch, right; hand horns, forward; thumbs-up, upwards. For the 7-hand gesture study: index-pinch, right; middle-pinch, left; ring-pinch, forward; pinky-pinch, backward; hand-horns, upwards; thumbs-up: downwards. These movements are customizable and are not tied to these specific hand gestures. Upon successfully aligning the ball with the target, the target would relocate to a random position, facilitating continuous gameplay.

2) Environment interfacing with ultrasound estimation: For streaming the CNN output to the VR space, two Python scripts were run simultaneously. One of the scripts generated the ultrasound image's CNN-predicted output and mapped the predicted hand positions in the VR space. This was done by attaching a specific letter to each prediction, which was then saved in a text file. The second script ran in the Vizard environment and was responsible for running the game. This script read the same text file and adjusted the ball's movement accordingly. The full pipeline is shown in Figure 3.

III. Results

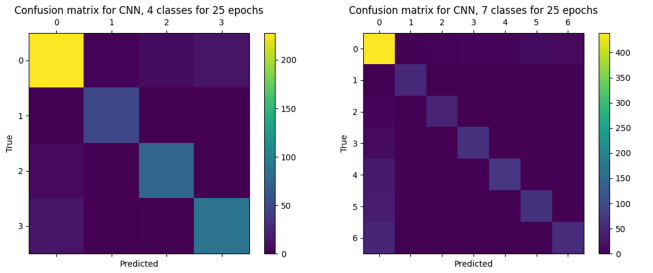
Classification accuracy results for offline and online cases for both 4-hand gesture and 7-hand gesture studies were acquired. For the offline analysis, classification accuracy is obtained on both the training and testing sets. For the online evaluation, the subject was made to attain a hand gesture based on a random gesture selection implementation. 5 frames of ultrasound data and the CNN's corresponding estimation were then acquired based on the attained gesture. The mode of the results based on each 5-frame window was then used to record the final prediction for that hand gesture. This was done 100 times. The results are supplemented with a video presentation containing several videos with a demonstration of the 4-hand gesture and 7-hand gesture classification and corresponding virtual object manipulation.

A. Classification results on the subject (offline results)

The confusion matrix for the 4-hand gesture classification for the CNN trained on 1920 training images and evaluated on 480 testing images is shown in Figure 4.(a). The accuracy on the training images was 92.1%. The accuracy on the test images was 91.3%. The confusion matrix for the 7-hand gesture classification for the CNN trained on 3840 training images and evaluated on 960 testing images is also shown in Figure 4.(b). The accuracy on the training images was 96.0%. The accuracy on the test images was 81.5%.

B. Classification results on the subject (online)

For the 4-hand gesture classification's online evaluation, the estimations were accurate 92% of the time. These results are close to from what was observed with the offline analysis. Forearm ultrasound-based 4-hand gesture classification for a real-time system control has previously reported a similar accuracy percentage for controlling a soft robotic gripper [10]. It must be noted that the average time per estimation using the trained CNN for generating estimations was 6.9 milliseconds. For the 7-hand gesture classification's online evaluation, the estimations were accurate 50% of the time. The average time per estimation using the trained CNN for generating estimations was 7.3 milliseconds. Future work will focus on improving the results for 7-hand gesture classification



(a) (b)

Ultrasound image screenshot	Image gray-scaling	Ultrasound image reshaping	CNN predict-ion	Extracting the gesture	Writing to the text file	VR goal validat-ion	Sending data to VR game	Clearing the text file
47.47	2.32	0.01	7.84	0.06	0.67	0.08	0.36	0.67

(c)

Fig. 4. Results: (a) Confusion matrix for 4-hand gesture classification (Evaluation on 960 testing samples using a CNN trained on 3840 samples), (b) Confusion matrix for 7-hand gesture classification (Evaluation on 480 testing samples using a CNN trained on 1920 samples), and (c) Latency in different components of the pipeline in milliseconds, totaling 59.48 milliseconds.

by hyperparameter optimization and analyzing the system performance for a higher number of hand gestures.

C. Demonstration and latency analysis

Here is the link to the presentation video: (`https://youtu.be/8Cx___jnLJM`). It contains several demonstrations of the pipeline for the 4-hand gesture and the 7-hand gesture classification. While the 4-hand gesture classification performs very well consistent with the online evaluation performance, even the 7-hand gesture classification performs well compared to the evaluation metrics that were obtained during the online performance analysis. The video shows two good cases and one case in which the subject struggled to reach the goal for each study.

Throughout multiple sessions of interacting with the VR game, the CNN estimations and the VR integration exhibited prompt responsiveness, with no noticeable delay between hand position changes and the corresponding adjustment in the ball's trajectory. The accuracy of the algorithm, yielded minimal discrepancy between the desired and actual directions of the ball. The most frequent errors occurred for hand configurations that closely resembled one another, indicating the intricacies of distinguishing subtle differences. The latency results averaged over 10000 samples are described in Figure 4.(c). The overall system latency from the ultrasound image screenshots to the VR manipulation and text file clearing was obtained to be 59.48 milliseconds.

IV. Conclusions

This paper presented a data-driven approach using forearm ultrasound data and deep learning for hand gesture recognition for manipulating objects in virtual reality (VR) interfaces. For 4-hand gesture classification, the methodology achieved high accuracy percentages of 91% for offline evaluation and 92% for online evaluation. 7-hand gesture classification was also analyzed. The study demonstrated the effectiveness of utilizing forearm ultrasound and deep learning based neural networks to recognize hand gestures during both training and real-time inference. The pipeline demonstrated an impressive latency of 59.48 milliseconds, encompassing the entire process from acquiring screenshots of ultrasound data to sending commands for manipulating the VR object. This research provides valuable insights into ultrasound-based hand gesture recognition techniques, enhancing user interaction and immersion in VR applications. By leveraging deep learning and real-time data acquisition, this study contributes to the advancement of body sensor networks in VR environments.

References

[1] R. Gravin and G. Fortino, "Wearable body sensor networks: state-of-the-art and research directions," IEEE Sensors Journal 21.11 (2020): 12511-12522.

[2] C. Ahmadizadeh, M. Khoshnam, and C. Menon, "Human machine interfaces in upper-limb prosthesis control: A survey of techniques for preprocessing and processing of biosignals," IEEE Signal Processing Magazine (2021), 38(4), 12-22.

[3] L. Dipietro, A. M. Sabatini, and P. Dario, "A survey of glove-based systems and their applications," IEEE transactions on systems, man, and cybernetics, part c (applications and reviews) 38.4 (2008): 461-482.

[4] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," IET Computer Vision 12, no. 1 (2018): 3-15.

[5] K. Bimbraw, and M. Zheng, "Towards The Development of a Low-Latency, Biosignal-Controlled Human-Machine Interaction System," In 2023 IEEE/SICE International Symposium on System Integration (SII), pp. 1-7. IEEE, 2023.

[6] C. L. Toledo-Peral, G. Vega-Martínez, J. A. Mercado-Gutiérrez, G. Rodríguez-Reyes, A. Vera-Hernández, L. Leija-Salas, and J. Gutiérrez-Martínez, "Virtual/Augmented Reality for Rehabilitation Applications Using Electromyography as Control/Biofeedback: Systematic Literature Review," Electronics 11, no. 14 (2022): 2271.

[7] M. C. Tosin, J. C. Machado, and A. Balbinot, "semg-based upper limb movement classifier: Current scenario and upcoming challenges," Journal of Artificial Intelligence Research 75 (2022): 83-127.

[8] S. Patwardhan, J. Schofield, W. M. Joiner, and S. Sikdar, "Sonomyography shows feasibility as a tool to quantify joint movement at the muscle level," In 2022 International Conference on Rehabilitation Robotics (ICORR), pp. 1-5. IEEE, 2022.

[9] H. Wang, S. Zuo, M. Cerezo-Sánchez, N. G. Arekhloo, K. Nazarpour, and H. Heidari, "Wearable super-resolution muscle-machine interfacing," Frontiers in Neuroscience (2022).

[10] K. Bimbraw, E. Fox, G. Weinberg, and F. L. Hammond, Towards sonomyography-based real-time control of powered prosthesis grasp synergies," In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 4753-4757. IEEE, 2020.

[11] K. Bimbraw, C. J. Nycz, M. J. Schueler, Z. Zhang, and H. K. Zhang, "Prediction of Metacarpophalangeal joint angles and Classification of Hand configurations based on Ultrasound Imaging of the Forearm," In 2022 International Conference on Robotics and Automation (ICRA), pp. 91-97. IEEE, 2022.

[12] K. Bimbraw, C. J. Nycz, M. J. Schueler, Z. Zhang, and H. K. Zhang, "Simultaneous Estimation of Hand Configurations and Finger Joint Angles Using Forearm Ultrasound," IEEE Transactions on Medical Robotics and Bionics 5, no. 1 (2023): 120-132.

[13] J. He, H. Luo, J. Jia, J. T. Yeow, and N. Jiang, "Wrist and finger gesture recognition with single-element ultrasound signals: A comparison with single-channel surface electromyography," IEEE Transactions on Biomedical Engineering (2018), 66(5), pp.1277-1284.

[14] R. Tsumura, J. W. Hardin, K. Bimbraw, A. V. Grossestreuer, O. S. Odusanya, Y. Zheng, J. C. Hill, B. Hoffmann, W. Soboyejo, and H. K. Zhang, "Tele-operative low-cost robotic lung ultrasound scanning platform for triage of COVID-19 patients," IEEE robotics and automation letters 6, no. 3 (2021): 4664-4671.