Mirror-based ultrasound system for hand gesture classification through convolutional neural network and vision transformer

Keshav Bimbraw*[a], Haichong K. Zhang[a]
[a]Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA, USA 01605

ABSTRACT

This research presents an innovative mirror-based ultrasound system designed for hand gesture classification using Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures. Hand gesture recognition using ultrasound has garnered significant interest due to its potential applications in various fields such as prosthetic control and human-machine interfacing. Traditionally, ultrasound probes are placed perpendicular to the forearm causing discomfort and interference with natural arm movements due to the center of mass of the wearable ultrasound system being distanced from the body. To address this challenge, a novel approach utilizing the advantages of acoustic reflection is proposed. A convex ultrasound probe is strategically aligned with the forearm, and ultrasound waves are transmitted to the forearm, and received back using a mirror placed at 45 degrees to the imaging region and the forearm. By aligning the probe parallel to the arm, the center of mass is brought closer to the body, ensuring enhanced stability and reduced strain on the user's arm during data collection. A dataset comprising 5 hand gestures was collected to train and evaluate the performance with Support Vector Machines with linear kernel, CNN, and ViT based approaches. It was observed that the performance of the mirror-based ultrasound system is comparable to the traditional perpendicular approach for hand gesture classification. The experimental results demonstrate the potential of the system in assisting with data acquisition and device development for hand gesture recognition using ultrasound in the field of human-machine interfacing, prosthetic control, human-computer interaction, and beyond.

Keywords: Hand gesture estimation, Mirror based ultrasound, Forearm ultrasound, Vision Transformers, Deep Learning, Machine Learning, Sonomyography, Hardware Design

1. INTRODUCTION

Hand gesture classification is a crucial component to facilitate effective human-machine interfacing. Various modalities such as camera, wearable hand gloves and distance sensors have been explored for this purpose. Biosignal based methods can be used to capture high-quality biological data, enabling the inference of hand movements through subtle changes in the human body. Although surface electromyography (EMG) stands out as a widely researched modality for this purpose, ultrasound emerges as a promising alternative, providing comprehensive visualization of forearm musculature to infer hand configurations [1-2]. With the advances in machine learning, ultrasound based human-machine interfaces have been used to control robots and AR/VR interfaces [3-4]. Ultrasound has not just been used for hand gesture classification but also for estimating finger angles and finger forces [5]. While ultrasound can provide a wide array of benefits for reconstruction of hand behavior through changes in the forearm musculature, there is a need to make the ultrasound data acquisition more comfortable for the user. There has been a lot of research on miniaturizing ultrasound data acquisition systems for effective data acquisition and interfacing with the human body [6-7]. However, the previous research does not focus on directly using the widely available B-mode ultrasound probes specifically for forearm data acquisition.

To that end, we propose a reflector based ultrasound system for forearm ultrasound data acquisition and hand gesture classification. The innovative system uses a mirror at 45º angle to both the imaging surface and the forearm. Doing so achieves a dual purpose: The spatial footprint required for probe data acquisition is minimized and the stability of the wearable for foreram ultrasound data acquisition is enhanced. The latter is achieved because the center of the mass of the probe is closer to the body compared to the traditional perpendicular configuration. We use a convolutional neural network (CNN) based on [1-2], in addition to training a vision transformer (ViT) based on [8] to train models to estimate 5 hand gestures from ultrasound images obtained using both traditional perpendicular configuration as well as our proposed reflector based configuration. Section II describes the methods and the experimental design, with the results discussed in Section III. The paper concludes with a discussion of our results.

*kbimbraw@wpi.edu; bimbrawkeshav@gmail.com; phone 1 678 436-9426; https://bimbraw.github.io

Medical Imaging 2024: Ultrasonic Imaging and Tomography, edited by Christian Boehm, Nick Bottenus,
Proc. of SPIE Vol. 12932, 129320Z · © 2024 SPIE · 1605-7422
doi: 10.1117/12.3005247

2. METHODS

A system was designed to evaluate the performance of hand gesture classification using forearm ultrasound for mirror based parallel and perpendicular probe configurations as shown in figure 1(a). A support vector classifier (SVC) with different kernels, convolutional neural network (CNN) and vision transformer (ViT) were used for data analysis.

Data was acquired using a convex ultrasound probe, with the data continuously streamed to a computer. These streamed images were then saved locally for the two configurations. Figure 1 shows the workflow for data acquisition and analysis for both the proposed (figure 1(b)) and the traditional perpendicular (figure 1(c)) configurations. The corresponding ultrasound data is also shown in the figures.

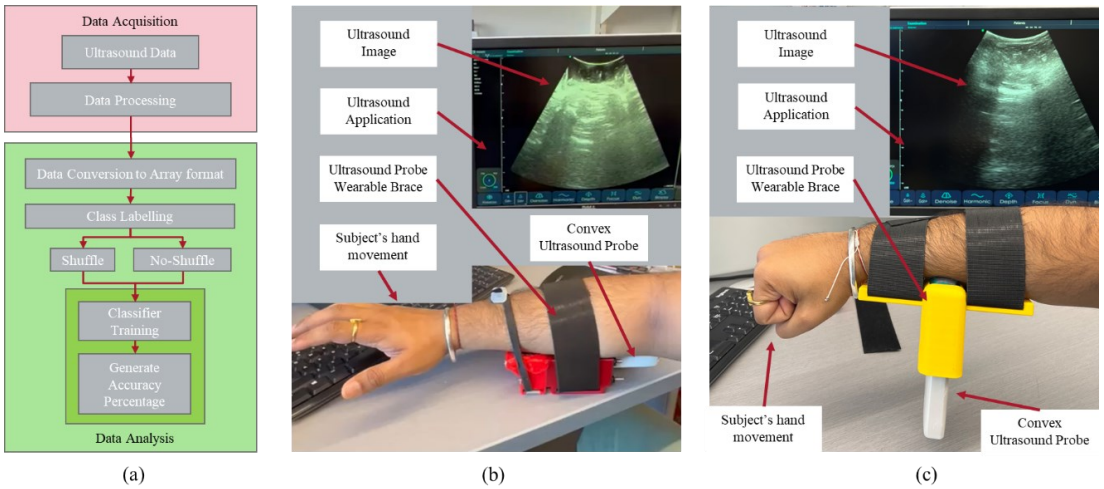

Figure 1. Data acquisition for mirror based parallel and perpendicular configurations. (a) Workflow for data acquisition and analysis, (b) System description for the mirror based parallel configuration, and (c) the perpendicular configuration.

The probe attachment for both the configurations was designed in SolidWorks and then 3-D printed. While the perpendicular configuration comprised of only the 3D printed attachment, the mirror configuration comprised of three elements: mirror, mirror base and probe attachment. The mirror was attached to the mirror base to maintain a 45-degree angle to both the imaging elements as well as the forearm, and the base and probe attachment were assembled using screws as fastening elements. Figure 2 shows the 3D models and various views for the mirror-based configuration.

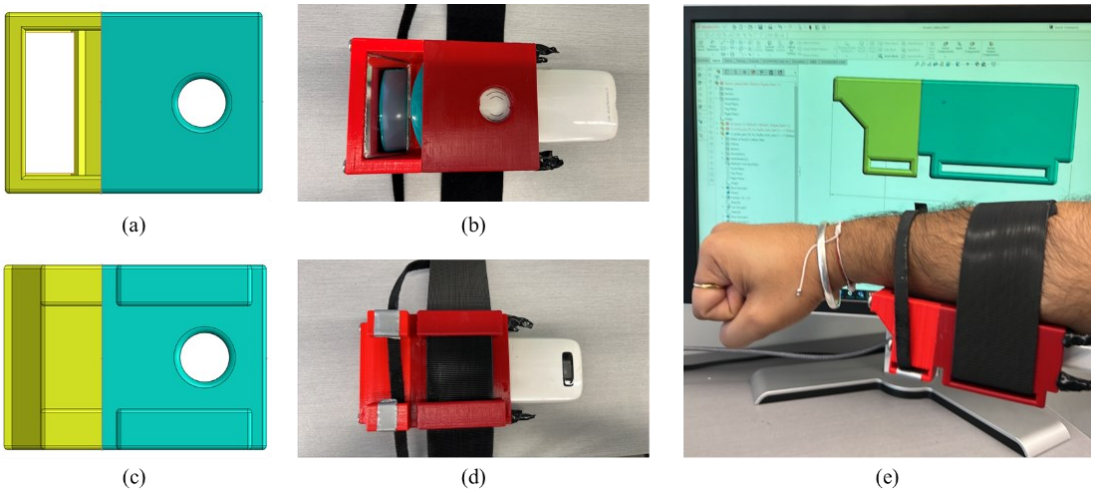

Figure 2. The mirror based wearable ultrasound brace for forearm ultrasound data acquisition. (a) 3D model base view, (b) Base view of the ultrasound probe with mirror attachment, (c) 3D model top view, (d) Top view of the ultrasound probe with the mirror attachment, and (e) Probe and attachment worn by the user with a 3D model in the background.

5 hand gestures were considered for the analysis. These included open hand, thumb flexion, index flexion, middle flexion, and ring flexion. 100 frames were acquired for each class of hand gestures. This was repeated 6 times, and this yielded a total of 3000 images for analysis. Each image was 640 x 640 pixels. Figure 3 shows the representative images for the 5 classes for parallel and perpendicular configurations. The data was subjected to a 5/6 train-test split. Analysis was done on shuffled (time-dependent) and non-shuffled (time independent) data.

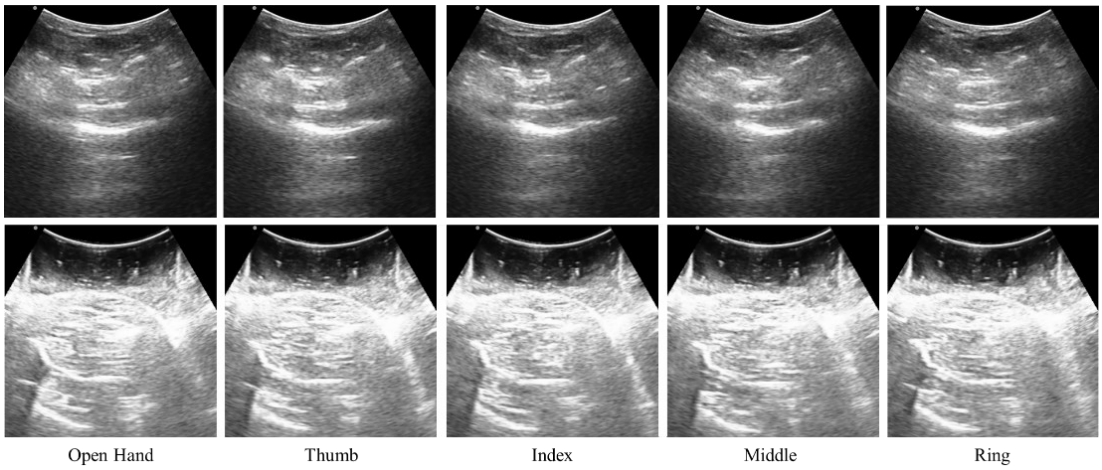

Figure 3. Representative ultrasound images acquired for each class for the perpendicular configuration (top row) and the mirror based parallel configuration (bottom row).

SVC and CNN have been shown to work well for hand gesture classification [1-3]. CNNs excel at extracting hierarchical features from image data, while ViT transforms image patches into meaningful embeddings by leveraging self-attention mechanisms, allowing for a comprehensive understanding of long-range dependencies and interdependencies among different regions within the image. The CNN from [1-2] was used for analysis. For the CNN, Adam optimizer with a learning rate of 0.001 was used. The loss was based on sparse categorical cross entropy. The training was done for a batch size of 16 and 20 epochs. Additionally, use ViT for hand gesture classification, with the architecture based on [8]. The system on which the training was done had NVIDIA GeForce RTX 2070 SUPER GPU, and AMD Ryzen 7 2700X Eight-Core Processor with 32 GB available RAM. The code was executed in Python 3.7 and TensorFlow library was used for implementing the architectures and running the training.

Accuracy percentage is used as the metric for evaluating classification accuracy. Accuracy percentage is defined 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.

3. RESULTS AND FUTURE WORK

Table 1 and figure 4 show the accuracy percentages for the mirror based and perpendicular configurations for different training approaches. It was observed that the perpendicular configuration performed better than the mirror based parallel configuration on an average. This can be attributed to various reasons, including the presence of air bubbles in the gel interface between the skin and the imaging surface for the mirror based configuration. Another observation was that the models performed better with the shuffled data than for non-shuffled data because of the introduction of temporal dependencies. 4 different SVC kernels were used: Sigmoid Linear, Polynomial and Radial Basis Function (RBF). SVC-Linear emerged as the top performer for shuffled data (consistent with results obtained in [3]), while SVC-RBF, closely trailed by ViT, exhibited superior performance for non-shuffled data across both configurations.

Our conclusive insights lean towards non-shuffled data, since it is closer to real-world scenarios where temporal dependencies are less prevalent during both training and testing phases. Using ViT, we achieved 93% classification accuracy for both mirror-based and perpendicular configurations indicating consistency of the proposed reflector-based

approach compared to the perpendicular approach. This also demonstrates the efficacy of using transformer based models for ultrasound image classification due to ViT's ability to capture long-range dependencies within data. It was interesting to note that while CNN outperformed ViT for the perpendicular configuration, ViT performed better than CNN for the mirror configuration. This could be due to ViT's strength in capturing global relationships, where understanding the overall forearm musculature and its interactions was crucial for effective hand gesture classification, giving it an edge over CNN, which focuses more on spatial features.

While the current results are encouraging, more work is needed towards improving the system and the study. It is important to address the issues caused by commercially available ultrasound gel used in the mirror based forearm ultrasound data acquisition system, such as air bubbles and low viscosity leading to leakages which affect the ultrasound image quality. Subsequent efforts will concentrate on developing custom gel-based interfaces to ensure improved ultrasound image quality and greater stability which would lead to less gel leakage. In addition to that, several systems have been proposed for biosignal based real time effective human-machine interfaces [4, 9]. Investigating real-time deployment of the reflector-based ultrasound system in human-machine interactions is of interest since it will offer insights into its feasibility for ultrasound based hand gesture recognition in dynamic, real-world scenarios.

Expanding the scope of the study involves acquiring data for a more extensive range of hand gestures. Expanding the gesture set will not only test the system's adaptability but also increase its potential applications in diverse human-machine interaction scenarios. Additionally, the current results are derived from data obtained from a single subject. To ensure the generalizability and applicability of the system across diverse individuals, future work will involve acquiring data from multiple subjects. This broader subject pool will enable an evaluation of the system's performance under varying anatomical characteristics, enhancing its versatility.

Table 1. Accuracy percentages for different configurations.

Model	***Acc*** **(Mirror, Shuffled)**	***Acc*** **(Mirror, Non-shuffled)**	***Acc*** **(Perp., Shuffled)**	***Acc*** **(Perp., Non-shuffled)**
SVC-Sigmoid	18	38	17	93
SVC-Linear	**98**	88	**98**	95
SVC-Polynomial	96	89	96	95
SVC-RBF	92	**93**	93	**94**
CNN	95	88	97	94
ViT	92	**93**	94	**93**

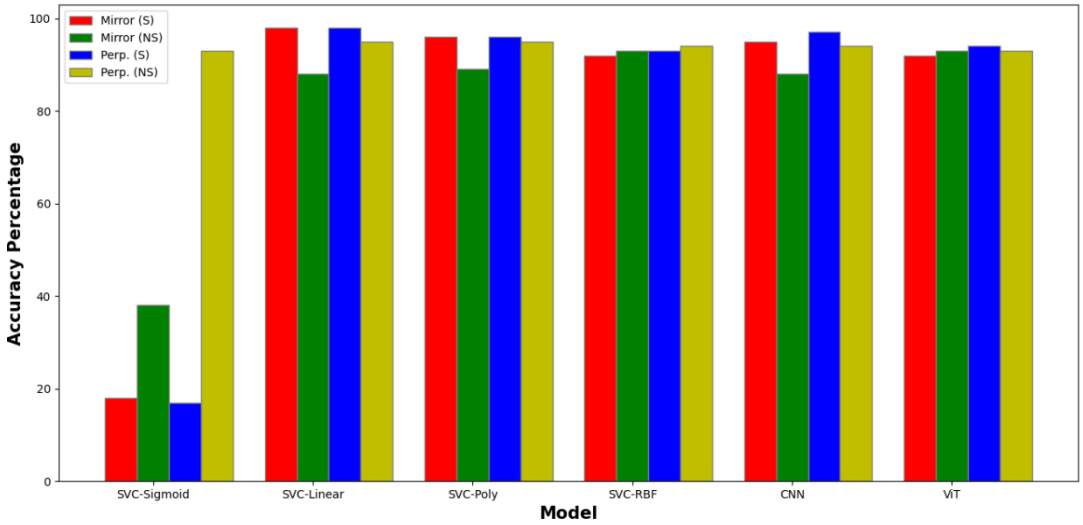

Figure 4. Results for accuracy percentage obtained for SVC with different kernels, CNN and ViT.

4. CONCLUSIONS

In this work, an innovative mirror-based forearm ultrasound data acquisition system is described with a mirror placed at 45 degrees to the imaging elements and the forearm for improved data acquisition. The shift from the traditional perpendicular to a parallel probe-arm configuration leads to the center of mass of the device being closer to the arm, improving overall stability and user experience. Ultrasound data for 5 hand gestures for both the configurations was acquired for one subject. Convolutional neural network and vision transformer were used to train classifiers. Using vision transformers led to a similar classification performance for both configurations. This sets a foundation for future work in wearable mirror/reflector-ultrasound based hand gesture recognition for better wearability and user mobility.

ACKNOWLEDGEMENTS

The authors are grateful to Amazon Robotics Greater Boston Tech Initiative, Worcester Polytechnic Institute Internal Fund and National Institutes of Health funding (Grant Number: DP5 OD028162).

REFERENCES

[1] K. Bimbraw, C. J. Nycz, M. Schueler, Z. Zhang and H. K. Zhang, "Simultaneous Estimation of Hand Configurations and Finger Joint Angles Using Forearm Ultrasound," in *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 1, pp. 120-132, Feb. 2023, doi: 10.1109/TMRB.2023.3237774.

[2] 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," *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 91-97, doi: 10.1109/ICRA46639.2022.9812287.

[3] K. Bimbraw, E. Fox, G. Weinberg and F. L. Hammond, "Towards Sonomyography-Based Real-Time Control of Powered Prosthesis Grasp Synergies," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 4753-4757, doi: 10.1109/EMBC44109.2020.9176483.

[4] K. Bimbraw, J. Rothenberg and H. Zhang, "Leveraging Ultrasound Sensing for Virtual Object Manipulation in Immersive Environments," *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, Boston, MA, USA, 2023, pp. 1-4, doi: 10.1109/BSN58485.2023.10331075.

[5] K. Bimbraw and H. K. Zhang, "Estimating Force Exerted by the Fingers Based on Forearm Ultrasound," *2023 IEEE International Ultrasonics Symposium (IUS)*, Montreal, QC, Canada, 2023, pp. 1-4, doi: 10.1109/IUS51837.2023.10306652.

[6] H. Huang, R. S. Wu, M. Lin and S. Xu, "Emerging Wearable Ultrasound Technology," in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, doi: 10.1109/TUFFC.2023.3327143.

[7] S. Frey, S. Vostrikov, L. Benini and A. Cossettini, "WULPUS: a Wearable Ultra Low-Power Ultrasound probe for multi-day monitoring of carotid artery and muscle activity," *2022 IEEE International Ultrasonics Symposium (IUS)*, Venice, Italy, 2022, pp. 1-4, doi: 10.1109/IUS54386.2022.9958156.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[9] K. Bimbraw and M. Zheng, "Towards The Development of a Low-Latency, Biosignal-Controlled Human-Machine Interaction System," *2023 IEEE/SICE International Symposium on System Integration (SII)*, Atlanta, GA, USA, 2023, pp. 1-7, doi: 10.1109/SII55687.2023.10039467.