Enhanced Human-Robot Teaming Through Attention Multi Convolutional Neural Network-Based Multi-Modal Sensor Fusion for Hand Gesture Recognition and Orientation Control

Alf Stian Sundo Gonsholt
University of Agder,
Grimstad, Norway
asgonsholt@uia.no

Eivind Enea Greca
University of Agder,
Grimstad, Norway
eivindeg@uia.no

Mustapha Haddad
University of Agder,
Grimstad, Norway
mustaphah@uia.no

Muhammad Hamza Zafar
University of Agder,
Grimstad, Norway
muhammad.h.zafar@uia.no

Filippo Sanfilippo
University of Agder,
Grimstad, Norway
filippo.sanfilippo@uia.no
Corresponding author

Abstract

Our study aims at enhancing Human-Robot Interaction, Collaboration, and Teaming (HRI/C/T) in industrial automation by developing a novel framework for real-time gesture control of a robotic hand. We use an Inertial Measurement Unit (IMU) sensor for precise orientation control of the end effector, and surface Electromyography (sEMG) sensors to detect muscle movements. The sEMG signals are processed by an Attention-based Multi Convolutional Neural Network (A-MCNN) for accurate gesture detection, enabling the robotic hand to mimic these gestures in real-time. Our method achieves notable results for gesture recognition, with the A-MCNN model attaining an accuracy of 97.89%, a precision of 97.49%, a recall of 97.71%, and an F1 score of 97.65%. This integration of IMU and sEMG technologies with advanced neural networks creates a responsive and intuitive control mechanism, improving safety, usability, and interaction of collaborative robots in shared workspaces. Our approach aims to transition towards Human-Robot Teaming (HRT), significantly advancing the seamless and safe integration of robots in industrial environments, enhancing productivity and collaboration.

Keywords: Multi-Modal Sensor Fusion, Inertial Measurement Units (IMUs), Electromyography (EMG) Sensors, Hand Gesture Recognition, Deep Neural Network, Gesture Recognition

1. Introduction

In the early days of industrial automation, robots were confined to cages or enclosed areas to ensure the safety of human workers. These caged robots were designed to perform tasks that were repetitive, hazardous, or required high precision without any direct human involvement. The physical separation was crucial because these machines operated at high speeds and with significant force, posing serious risks to humans in the event of errors or malfunctions. By placing robots in cages, manufacturers could maximise efficiency and productivity while protecting workers from potential harm.

With advancements in technology, there has been a significant shift towards more integrated Human-Robot Interaction, Collaboration, and Teaming (HRI/C/T). This new approach focuses on creating environments where robots and humans work more and more closely together. The development of more sophisticated and safer robotic systems has made this possible. These modern robots, often called collaborative robots or "cobots", are equipped with advanced sensors and control algorithms that enable them to understand and respond to human actions. Human involvement in HRI/C/T is increasingly crucial because it leverages the strengths of both humans and robots. Robots excel at performing precise and repetitive tasks, while humans bring creativity, problem-solving skills, and adaptability to the table. This synergy aims to enhance productivity and efficiency by allowing robots to assist humans rather than replace them.

However, integrating humans into HRI/C/T presents several challenges. One of the primary concerns is ensuring safety in a shared workspace, requiring sophisticated technology to ensure that robots can operate safely alongside humans. Additionally, developing intuitive interfaces and control systems is crucial for effective coordination and cooperation between humans and robots. These systems need to be user-friendly so that workers can easily interact with and

URI: https://hdl.handle.net/10125/108907
978-0-9981331-8-8

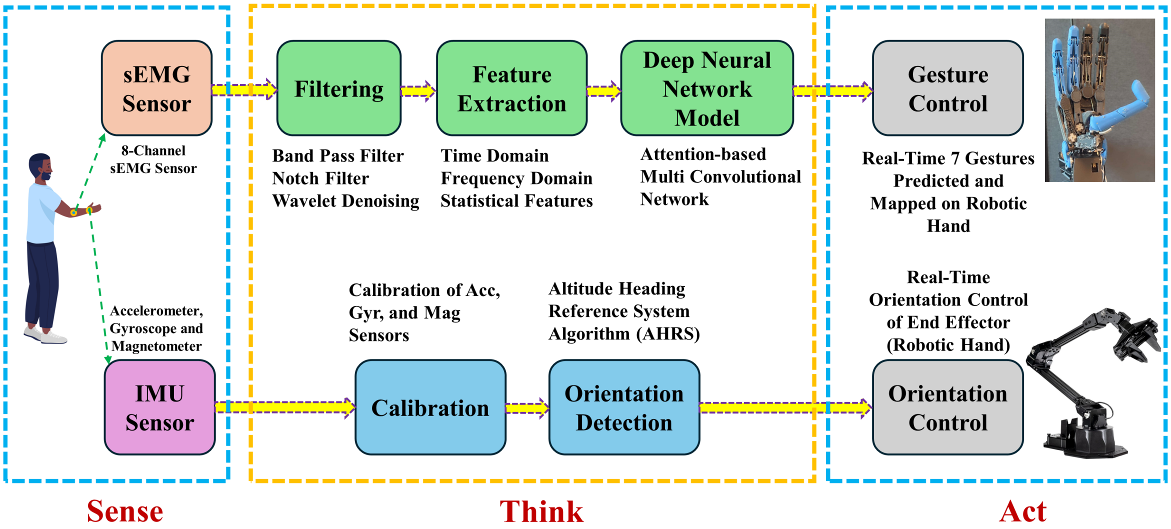
sEMG Sensor
8-Channel sEMG Sensor
Accelerometer, Gyroscope and Magnetometer
IMU Sensor
Filtering
Band Pass Filter
Notch Filter
Wavelet Denoising
Feature Extraction
Time Domain
Frequency Domain
Statistical Features
Deep Neural Network Model
Attention-based Multi Convolutional Network
Calibration of Acc, Gyr, and Mag Sensors
Calibration
Altitude Heading Reference System Algorithm (AHRS)
Orientation Detection
Gesture Control
Real-Time 7 Gestures Predicted and Mapped on Robotic Hand
Real-Time Orientation Control of End Effector (Robotic Hand)
Orientation Control
Sense
Think
Act

Figure 1: Detailed description of multi-modal fusion of IMU and sEMG sensors with deep neural network based gesture recognition and orientation control.

control the robots.

Despite the advancements in HRI/C/T, there remains a significant research gap in the seamless integration of human physiological signals for intuitive robot control. Current systems often rely on complex and less intuitive control interfaces, which can hinder the natural flow of HRI/C/T. There is a pressing need for innovative solutions that utilise human physiological signals, such as muscle activity, to predict and control robotic movements more effectively. Addressing this gap is essential for improving the responsiveness and accuracy of cobots, ultimately enhancing their usability and safety in industrial settings.

Addressing these challenges necessitates a multidisciplinary approach that integrates expertise from various fields such as robotics, human factors engineering, computer science, and artificial intelligence (AI). The underlying idea is shown in Figure 1. This involves creating advanced control algorithms, incorporating state-of-the-art sensor technologies, and developing intuitive user interfaces that facilitate seamless human-robot coordination. Successfully managing these complexities is crucial for the effective integration of humans and robots in the workplace, ultimately leading to enhanced productivity and safety.

Exploring these research trends, our method leverages the adoption of an Attention-based Multi Convolutional Neural Network (A-MCNN) to fuse two types of sensors: Inertial Measurement Units (IMUs) and Electromyography (EMG) sensors. The IMUs help control the direction and movement of a robot's hand, making sure it moves accurately and steadily. EMG sensors detect muscle signals from the human operator, which are used to predict hand gestures and allow the robot's hand to move accordingly. By combining these two technologies, we aim to create an intuitive and responsive way for humans to control robots. This approach ensures that robots can move safely and accurately, facilitating natural hand movements for human operators. By enabling robots to interpret and respond to human gestures, the system fosters more intuitive and effective interactions, advancing collaborative robotics by bridging human intent with robotic execution.

1.1. Contributions and Paper Organization

The main contributions of this work are listed in the following:

- Multi-Modal Sensing Integration: We introduce a novel approach based on an Attention-based Multi Convolutional Neural Network (A-MCNN) that combines IMUs and EMG sensors to enhance the precision and responsiveness of cobots in HRI/C/T.
- Enhanced Human-Robot Coordination: By leveraging IMUs for steady and accurate robot movement control and EMG sensors for detecting human muscle signals, we develop a system that allows natural and intuitive control of robots through hand gestures.
- Safety and Usability Improvements: Our method

addresses critical challenges in ensuring safety and usability in shared workspaces, contributing to the advancement of HRC towards HRT.

The rest of the paper is organised as follows: a review of related works, setting the context for the proposed framework, is presented in Section 2. Section 3 details the proposed methodology, describing the integration of IMUs and EMG sensors. In Section 4, we present the results and discussion, where we analyse the performance of our system, and discuss its implications for HRI/C/T. Finally, Section 5 provides the conclusion and future Work, summarising our key findings, outlining the limitations of our current approach, and proposing directions for further research.

2. Literature Review

HRI has been a fundamental area of research in robotics (Obaigbena et al., 2024), focusing on understanding and improving the ways humans and robots interact with each other. Initially, HRI mainly involved robots performing predefined tasks with minimal interaction with humans. Transitioning from HRI, there has been a notable shift towards HRC (Semeraro et al., 2023) in recent years. HRC emphasises closer cooperation between humans and robots, where robots are designed to work alongside humans, assisting them in various tasks. HRC has gained significant attention in recent years due to its potential to enhance various applications such as garbage sorting, environment exploration, military operations, and rescue missions (Duan et al., 2022). To enable effective collaboration, researchers have been exploring the integration of multimodal sensors and machine learning (ML) algorithms. These advancements aim to improve the interaction between humans and robots by fusing data from sensors like inertial measurement units (IMUs) and electromyography (EMG) sensors (Anvaripour et al., 2020; Colli Alfaro & Trejos, 2022).

Furthermore, the evolution continues towards HRT (Natarajan et al., 2023), marking a significant advancement in robotics research. HRT focuses on developing systems where humans and robots form cohesive teams, working together synergistically towards shared goals. Unlike traditional HRI or HRC, HRT emphasises not just cooperation but also mutual understanding, trust, and adaptability between humans and robots. This transition opens up new possibilities for applications where humans and robots collaborate seamlessly, complementing each other's capabilities to achieve tasks more effectively and safely. In the context of HRT, the fusion of data from IMUs and EMG sensors plays a crucial role in enhancing the understanding of human intentions and movements. By employing ML algorithms, biosignals acquired from users can be interpreted to plan appropriate robot reactions during shared tasks (Anvaripour et al., 2020). This fusion of sensor data enables robots to predict human intentions accurately, contributing to safer and more efficient collaboration between humans and robots (Rodrigues et al., 2022).

Moreover, the use of ML techniques in increasingly close the gap between robots and humans has been a subject of extensive research. Studies have highlighted the importance of cobots and the integration of human skills with robotic capabilities in various working environments. ML algorithms have been instrumental in enabling natural language interactions between humans and robots, enhancing the overall collaboration experience (Eloff & Engelbrecht, 2021).

Furthermore, the fusion of force-torque sensors, IMUs, and proprioception data has been widely adopted in robotics for tasks such as odometry, localisation, stabilisation, and control (Benallegue & Lamiraux, 2015). This integration of multiple sensor modalities allows for a more comprehensive understanding of the environment, facilitating safe space-sharing between humans and robots (Pedrocchi et al., 2013).

Despite these advancements, there remains a research gap in seamlessly integrating human physiological signals for intuitive robot control in HRC and HRT scenarios. Existing methods often lack robustness and real-time responsiveness required for dynamic human-robot coordination. This work aims to address this gap by proposing a novel method that leverages deep learning-based sensor fusion for enhanced HRI/C/T.

3. Proposed Methodology

3.1. Experimental Setup

The experimental setup uses a ViperX-300 produced by Trossen Robotics (Robotics, 2024). This robot has 5 degrees of freedom, a reach of 750mm, and a max payload of 750g (Zhou et al., 2022). This robot's end effector is replaced with a humanoid robotic hand, the Robot Nano Hand (Hand, 2024). The Robot operating system(ROS) is implemented to control the robot itself, and used to connect a wearable IMU sensor to the control system. The Robot Nano Hand is controlled via an sEMG sensor, through the use of a ML algorithm to predict human hand gestures. The diagram for this setup is shown in Figure 1, where orientation control algorithm is the ROS based IMU control. The gesture

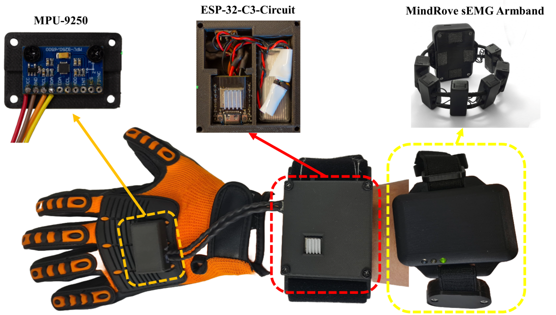

Figure 2: Detailed description of multi-modal setup.

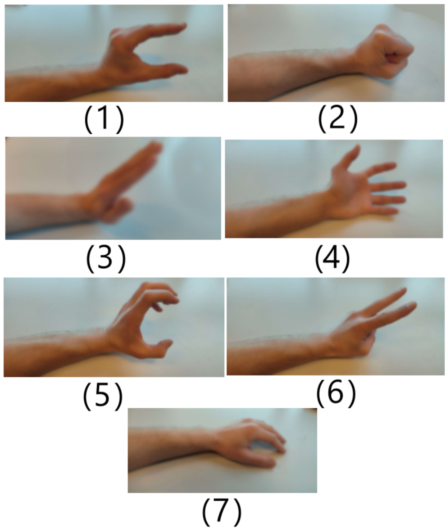

Figure 3: Description of the gestures used in this work.

control algorithm is the sEMG and ML based gesture control.

3.2. IMU based Orientation Control

For the orientation control of a robotic hand, the MPU9250 (InvenSense, 2024) is utilised, a low-cost IMU. The choice of the MPU9250 is driven by its affordability and sufficient accuracy for our application. The sensor is strategically placed on top of the robotic hand to mimic the role of an end effector on a robotic arm. The primary interaction with the system is through a terminal interface, emphasising the core mechatronics aspect without delving into advanced user interface (UI) enhancements. This terminal-based interface allows for straightforward, real-time communication with the IMU, facilitating various control and monitoring tasks.

The use of low-cost IMUs like the MPU9250 is crucial in making sophisticated robotic control systems accessible and affordable. Implementing these sensors in a way that they can be easily calibrated and utilised by users at any time is essential. This approach democratises advanced robotics technology, allowing hobbyists, researchers, and small-scale developers to experiment and innovate without significant financial barriers. At the heart of our orientation control system is the Attitude and Heading Reference System (AHRS) Justa et al., 2020, an algorithm that processes signals from the IMU. With AHRS, it is possible to implement precise navigation and tracking of the end effector's heading, providing critical information about the device's orientation. This functionality ensures that we can accurately determine and control the direction in which the robotic hand is pointing, enhancing its operational efficiency and effectiveness.

3.3. sEMG Sensor Based Data Acquisition

To detect the myoelectrical signals of the forearm, a MindRove armband surface EMG device (MindRove, 2024) with a sampling frequency of 500 Hz is utilised. This device is equipped with eight electrodes, which are equidistantly placed around the armband. For standardisation purposes, a bias channel and a reference channel is positioned on the brachioradialis muscle. To facilitate the connection to the MindRove device and initiate the reception of raw signals, the PyCharm Integrated Development Environment (IDE) (s.r.o., 2024) is employed in conjunction with the MindRove library and other built-in libraries. The experimental protocol involves testing the gestures depicted in Figure 3, specifically gestures (1) through (6), with the rest position shown as gesture (7). Eight human subjects are involved in the experiment. Each of the eight subjects perform five repetitions of each gesture, alternating between the rest phase and the active phase, with each phase lasting between three to four seconds.

3.3.1. Filtering The raw sEMG signal is known to be noise contaminated. Therefore, filtering is needed to get informative signals. During our tests, the EMG signals has a frequency range between 3 and 60 Hz. A band-pass filter is used to target this frequency range. A potential noise source is the radiation emitted from electrical sources, having 50/60 Hz frequency. A band-stop filter is employed to eliminate the effect of this noise source. Similarly, a notch filter is used to assure eliminating this noise source. In addition, a wavelet denoising filter with 6 vanishing moments (db6) and 3 decomposition levels is implemented. This is done to eliminate high frequency noise merged with the raw signal in the targeted range.

3.4. Feature Extraction

In this study, features are extracted from an 8-channel EMG sensor with a 250ms sample size and a 50% overlap. We use a combination of time-domain, frequency-domain, and statistical features to capture important characteristics of the EMG signals.

For the time-domain features, we calculate the Mean Absolute Value (MAV), Root Mean Square (RMS), Zero Crossing (ZC), and Slope Sign Changes (SSC). The MAV is the average of the absolute values of the EMG signal, providing a measure of the overall signal amplitude, and is computed as:

$$\text{MAV} = \frac{1}{N}\sum_{i=1}^{N}|x_i| \tag{1}$$

The RMS is the square root of the average power of the EMG signal, reflecting the energy content of the signal, and is given by:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2} \tag{2}$$

Zero Crossing (ZC) counts the number of times the signal crosses zero, indicating the frequency content of the signal. It is calculated by counting the sign changes in the signal. Slope Sign Changes (SSC) counts the number of times the slope of the signal changes sign, capturing changes in the frequency characteristics of the signal.

In the frequency-domain, we compute the Mean Frequency (MNF) and Median Frequency (MDF). MNF is the average frequency of the EMG signal, calculated from the Fourier transform, and provides an indication of the central frequency of the signal:

$$\text{MNF} = \frac{\sum_{k=1}^{M} f_k P(f_k)}{\sum_{k=1}^{M} P(f_k)} \tag{3}$$

where f_k is the frequency and $P(f_k)$ is the power at frequency f_k. MDF is the frequency that divides the power spectrum into two equal halves, calculated by finding the frequency at which the cumulative power spectrum is half of the total power.

Lastly, we extract statistical features such as Variance, Skewness, and Kurtosis. Variance measures the variability of the EMG signal, indicating how much the signal deviates from the mean, and is computed as:

$$\text{Variance} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{4}$$

Skewness measures the asymmetry of the signal distribution, showing whether the signal has a tendency to lean towards higher or lower values, and is calculated as:

$$\text{Skewness} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{3/2}} \tag{5}$$

Kurtosis measures the ”tailedness” of the signal distribution, indicating the presence of outliers or extreme values in the signal, and is given by:

$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{2}} \tag{6}$$

These features collectively provide a comprehensive representation of the EMG signals, capturing their amplitude, frequency, and statistical properties, which are essential for further analysis and classification tasks.

3.5. Attention based Multi Convolutional Neural Network

The model architecture for hand gesture recognition using sEMG sensors is based on a multi 1D convolutional neural network (CNN) enhanced with an attention mechanism. The input to the model is a sequence of sEMG signals represented as $X \in R^{N \times T}$, where N denotes the number of channels (sensors) and T represents the time steps of the signal.

The first layer consists of multiple 1D convolutional layers designed to extract temporal features from the sEMG signals. Each convolutional layer applies a set of filters $W^{(l)} \in R^{K \times F}$, where l indicates the layer number, K is the kernel size, and F is the number of filters. The convolution operation for the l-th layer can be expressed as:

$$H^{(l)} = \text{ReLU}(W^{(l)} * X + b^{(l)}) \tag{7}$$

where $H^{(l)}$ is the output feature map, $*$ denotes the convolution operation, and $b^{(l)}$ represents the bias term. The ReLU function introduces non-linearity into the model.

Following the convolutional layers, an attention mechanism is incorporated to focus on the most relevant parts of the feature maps. The attention mechanism computes attention scores α_t for each time step t, which helps the model to weigh the importance of different time steps in the sequence. The attention scores are computed as:

$$e_t = \tanh(W_a H_t + b_a) \tag{8}$$

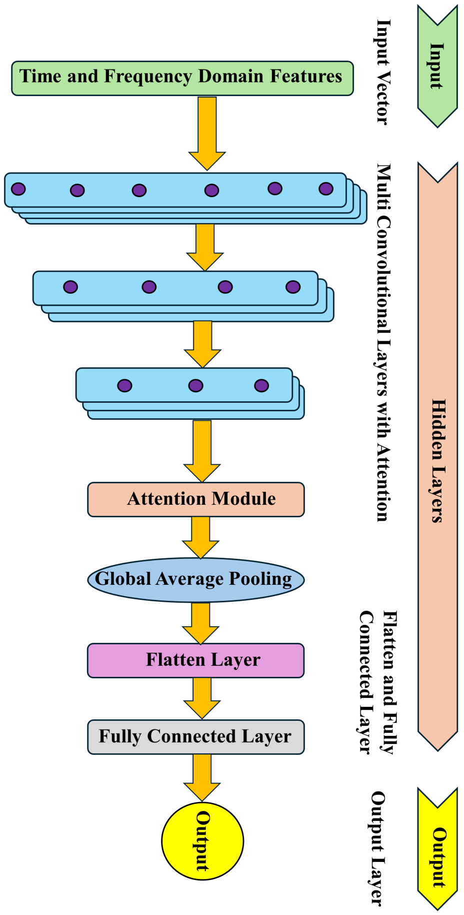

Figure 4: Detailed architecture of attention-based multi convolutional neural network for gesture recognition using sEMG sensors.

$$\alpha_t = \frac{\exp(e_t)}{\sum_{t=1}^{T} \exp(e_t)} \tag{9}$$

where W_a and b_a are learnable parameters of the attention mechanism. The context vector c is then obtained by taking a weighted sum of the feature maps:

$$c = \sum_{t=1}^{T} \alpha_t H_t \tag{10}$$

The context vector c is then passed through a fully connected layer to perform the final classification of hand gestures. The output layer uses a softmax function to produce probability distributions over the gesture classes:

$$y = \text{softmax}(W_c c + b_c) \tag{11}$$

where W_c and b_c are the weights and bias of the fully connected layer, respectively.

4. Results and Discussion

4.1. Orientation Control

In our study, we implement an IMU based orientation control system for the robot's end effector. Given the robot's limitations in degrees of freedom, our orientation control is restricted to managing pitch and roll angles. The core of our system relied on Madgwick's algorithm for the Attitude and Heading Reference System (AHRS), which provided a reliable means to estimate angles accurately despite some inherent limitations of the sensor. The accuracy of our IMU-based orientation control is heavily reliant on the proper calibration of the IMU sensor. During the initial stages, we observed that without meticulous calibration, the system suffered from noticeable drift and offsets. These inaccuracies could lead to significant deviations in the robot's movements, thereby affecting the overall performance and precision of the end effector's control. To address this, a comprehensive calibration routine is developed and implemented , ensuring that the sensor's measurements are both precise and reliable. Post-calibration, the drift and offsets is significantly minimised, resulting in stable and accurate angle estimations.

The wearable IMU sensor is strategically placed on the operator's left hand. In contrast, the robot is configured to operate using as if it was the operator's right hand. This mirrored setup is designed to test the system's ability to translate the operator's movements accurately despite the lateral inversion. The robot's end effector is thus tasked with mirroring the operator's hand rotations in real-time. Our results show that the robot could actively and accurately replicate the operator's hand movements, including both pitch and roll rotations. The implementation of Madgwick's algorithm in the AHRS system is instrumental in achieving this level of precision. The robot demonstrated a high degree of responsiveness and fidelity in mimicking the operator's hand movements. This is evident from several test scenarios where the robot's end effector is capable to maintain consistent and precise orientation control, closely following the operator's gestures.

We conducted a series of tests to quantify the system's performance, including latency measurement, accuracy tests, and stability analysis. In our evaluation, we tested the IMU orientation control with 500 samples, resulting in an average response time of 13 ms, demonstrating real-time control capability. The inference time for the EMG-based gesture recognition was measured at 15 ms. The angular accuracy of the robot's movements was measured against predefined

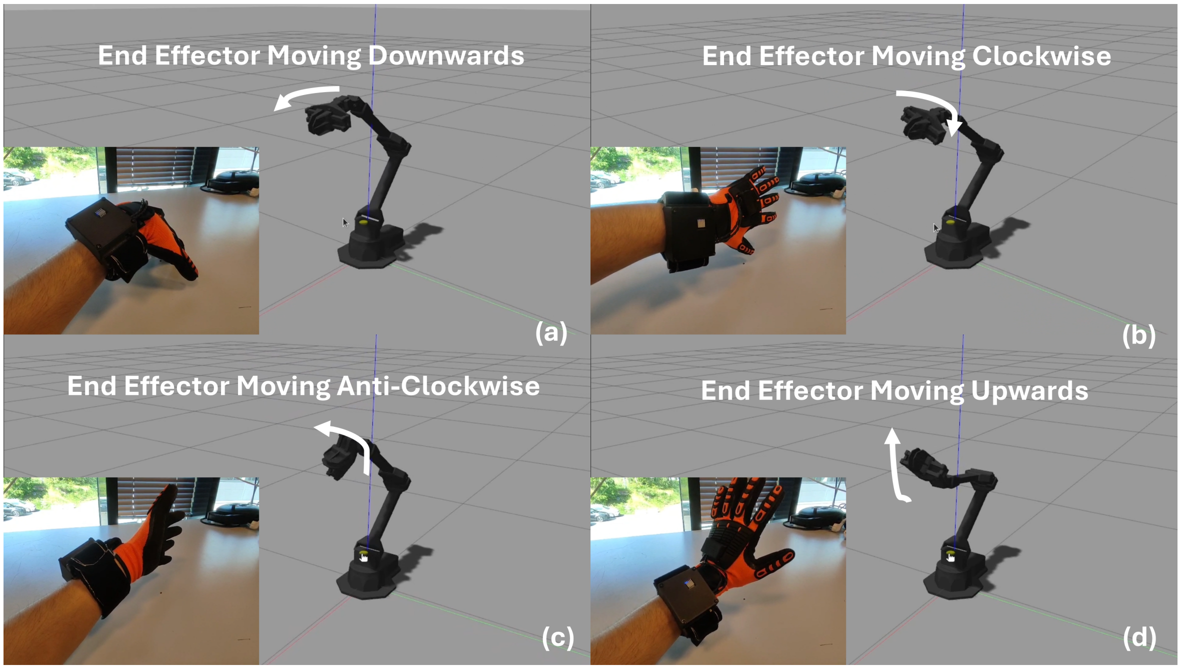
End Effector Moving Downwards
(a)
End Effector Moving Clockwise
(b)
End Effector Moving Anti-Clockwise
(c)
End Effector Moving Upwards
(d)

Figure 5: Orientation control of end effector at different times in real time testing. (a) End Effector moving downwards as the hand is moving downwards (b) End Effector moving clockwise as the hand is moving clockwise (c) End Effector moving anti-clockwise as the hand is moving anti-clockwise (d) End Effector moving upwards as the hand is moving updwards.

Table 1: Hyperparameters of the Convolutional Neural Network Model with Attention Mechanism

Hyperparameter	**Value**
Convolutional Layers	3 layers: Conv1D: 128 filters, kernel size 7, activation ReLU Conv1D: 64 filters, kernel size 5, activation ReLU Conv1D: 64 filters, kernel size 3, activation ReLU
Attention Mechanism	Yes
Global Average Pooling	Yes
Dense Layer	300 units, activation ReLU
Dropout	Rate 0.5
Output Layer	Softmax activation
Optimizer	Adam
Loss Function	Sparse Categorical Crossentropy

targets, with deviations found to be within acceptable limits. Over prolonged periods of operation, the system maintained its accuracy, with negligible drift, showcasing the robustness of the calibration process and the effectiveness of Madgwick's algorithm. The successful implementation of the IMU-based orientation control using Madgwick's algorithm has proven to be effective for the robot's end effector. The system's ability to accurately follow the operator's hand movements, even in a mirrored setup, highlights its potential for applications requiring precise and responsive control. Our findings underscore the importance of sensor calibration and the robustness of Madgwick's algorithm in achieving reliable orientation control. This setup provides a foundation for future enhancements and applications in various fields,

Table 2: Evaluation Metrics for Gesture Prediction Models

Model	Acc.	Prec.	Recall	F1 Score
A-MCNN	97.89	97.49	97.71	97.65
MCNN	95.23	94.57	94.24	94.35
DNN	84.87	84.11	83.52	83.75
SVM	62.45	61.08	61.36	61.25
RF	60.34	69.55	68.66	68.75

including teleoperation and assistive robotics. The detailed orientation control of the robotic end effector is shown in Figure 5.

4.2. sEMG based Gesture Recognition

The performance metrics for various gesture classification models are summarised in Table 2. These metrics include Accuracy, Precision, Recall, and F1 Score, providing a comprehensive evaluation of each model's effectiveness. The comparative analysis of proposed attention based multi convolutional neural network (A-MCNN) is made with other state of the art techniques like Multi convolutional neural network (MCNN), Feedforward deep neural network (DNN), support vector machine (SVM) and random forest (RF).

The Attention-based Multi CNN Network (A-MCNN) demonstrated the highest performance across all metrics, with an Accuracy of 97.89%, Precision of 97.49%, Recall of 97.71%, and an F1 Score of 97.65%. These results indicate that the A-MCNN model is highly effective at correctly identifying gestures with minimal errors, likely due to the enhanced feature extraction capabilities provided by the attention mechanism, which helps the model focus on the most relevant parts of the input data.

The Multi CNN Network (MCNN) also performed well but was outperformed by the A-MCNN. The MCNN achieved an Accuracy of 95.23%, Precision of 94.57%, Recall of 94.24%, and an F1 Score of 94.35%. While these results are strong, the slightly lower scores compared to the A-MCNN suggest that the attention mechanism in the A-MCNN provides a significant performance boost by better capturing the intricacies of the gesture data.

The Deep Neural Network (DNN) showed moderate performance with an Accuracy of 84.87%, Precision of 84.11%, Recall of 83.52%, and an F1 Score of 83.75%. The DNN's lower performance metrics compared to the CNN-based models indicate that while it is capable of learning and generalising from the gesture data, it may not capture spatial hierarchies and local dependencies as effectively as convolutional networks.

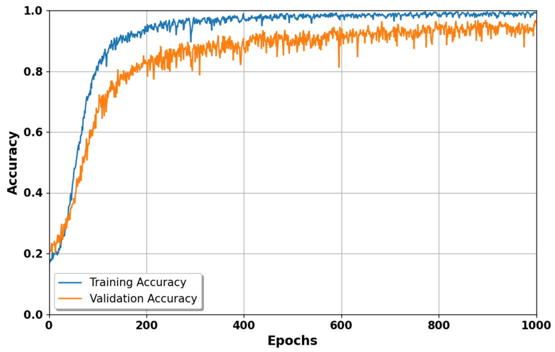

Figure 6: Comparison of training and validation accuracy vs epochs.

The Support Vector Machine (SVM) model performed relatively poorly with an Accuracy of 62.45%, Precision of 61.08%, Recall of 61.36%, and an F1 Score of 61.25%. These results suggest that the SVM, while useful in many classification tasks, struggles with the complexity and high dimensionality of the gesture data, leading to suboptimal performance.

The Random Forest (RF) model also showed limited effectiveness, with an Accuracy of 60.34%, but interestingly, it had a relatively high Precision of 69.55%, Recall of 68.66%, and an F1 Score of 68.75%. The higher Precision and Recall values compared to its Accuracy suggest that the RF model is somewhat better at identifying positive instances of gestures but still misses a significant number of correct classifications, leading to lower overall accuracy.

4.3. Discussion

The performance metrics for various gesture classification models reveal that the Attention-based Multi CNN Network (A-MCNN) outperformed all other models, achieving the highest scores in Accuracy, Precision, Recall, and F1 Score. This superior performance can be attributed to the enhanced feature extraction capabilities provided by the attention mechanism, which enables the model to focus on the most relevant parts of the input data. The Multi CNN Network (MCNN), while also performing well, showed slightly lower metrics, indicating that the attention mechanism indeed provides a significant boost in capturing the intricacies of gesture data. The Deep Neural Network (DNN) demonstrated moderate effectiveness, but its lower performance compared to CNN-based models suggests it may not capture spatial hierarchies and local dependencies as effectively. The Support Vector Machine (SVM) and Random Forest

(RF) models showed limited effectiveness, with the SVM struggling with the complexity of the gesture data and the RF model, despite higher Precision and Recall values, failing to achieve high overall accuracy due to significant misclassifications. These results underscore the importance of advanced deep learning techniques, such as attention mechanisms, in enhancing the accuracy and reliability of gesture recognition systems.

Additionally, the IMU-based orientation control system for the robot's end effector, employing Madgwick's algorithm for the Attitude and Heading Reference System (AHRS), demonstrated high accuracy in managing pitch and roll angles. The successful implementation of this system, despite initial drift and offsets that were mitigated through meticulous calibration, highlights its effectiveness in accurately mirroring the operator's hand movements. The robot's ability to maintain consistent and precise orientation control in a mirrored setup further showcases the robustness of the calibration process and the effectiveness of Madgwick's algorithm. Together, these findings emphasise the critical role of precise sensor calibration and advanced algorithms in achieving reliable and responsive control in both gesture recognition and orientation control systems.

4.4. Limitations

One limitation of this work is its limited generalizability. The study focuses on a specific type of robotic hand and sensor setup, which may not be easily applicable to other types of robotic systems or industrial settings. Significant modifications might be required for the framework to work in different contexts, such as with robots that have varying degrees of freedom or operate under different environmental conditions. Environmental sensitivity is also a concern. The study likely assumes controlled conditions during testing, but real-world industrial environments can vary greatly. Factors like temperature changes, humidity, electromagnetic interference, and physical obstructions could affect the system's performance, and these variables were not thoroughly evaluated. User variability poses another challenge. The system's effectiveness may differ depending on individual user characteristics, such as muscle signals influenced by factors like fatigue or body mass. This variability could impact the consistency and accuracy of the system across different users, which might not have been fully accounted for in the study.

Another important area for future development is the creation of edge AI models for EMG-based gesture recognition. The current system relies on high computational power to achieve accurate and fast gesture recognition, which may not be feasible in all industrial settings, particularly those with limited resources or where minimal latency is crucial. Developing edge AI models would enable the system to operate efficiently on devices with lower computational power while maintaining quick response times, making the technology more accessible and scalable for a wider range of applications. This advancement would significantly enhance the system's practicality and usability in real-world industrial environments.

5. Conclusion and Future Work

Our study demonstrates a significant advancement in industrial automation by developing a sophisticated system for real-time gesture control of a robotic hand, enhancing Human-Robot Interaction, Collaboration, and Teaming (HRI/C/T). By integrating Inertial Measurement Units (IMUs) for precise orientation control and surface Electromyography (sEMG) sensors for detecting muscle movements, processed through an Attention-based Multi Convolutional Neural Network (A-MCNN), we achieve highly accurate gesture recognition. The A-MCNN model achieves an accuracy of 97.89%, a precision of 97.49%, a recall of 97.71%, and an F1 score of 97.65%. This integration creates a responsive and intuitive control mechanism, significantly improving the safety, usability, and interaction of collaborative robots in shared workspaces. Our approach not only enhances efficiency and effectiveness but also aims to transition towards Human-Robot Teaming (HRT), promising a seamless and safe integration of robots in industrial environments, thus enhancing productivity and collaboration.

Future research could focus on further refining the deep learning models for gesture recognition and exploring additional modalities for sensor fusion to enhance the capabilities of HRT. The possibility of integrating haptic feedback may also be considered (Moosavi et al., 2022; Sanfilippo & Pacchierotti, 2020; Sanfilippo et al., 2021). Additionally, real-world deployment and testing of such systems will be crucial to validate their effectiveness and address practical challenges.

Acknowledgement

This research is supported by the Artificial Intelligence, Biomechatronics, and Collaborative Robotics research group at the Top Research Center

Mechatronics (TRCM), University of Agder (UiA), Norway.

References

Anvaripour, M., Khoshnam, M., Menon, C., & Saif, M. (2020). Fmg-and rnn-based estimation of motor intention of upper-limb motion in human-robot collaboration. *Frontiers in Robotics and AI*, *7*, 573096.

Benallegue, M., & Lamiraux, F. (2015). Estimation and stabilization of humanoid flexibility deformation using only inertial measurement units and contact information. *International Journal of Humanoid Robotics*, *12*(03), 1550025.

Colli Alfaro, J. G., & Trejos, A. L. (2022). User-independent hand gesture recognition classification models using sensor fusion. *Sensors*, *22*(4), 1321.

Duan, H., Wang, P., Li, Y., Li, D., & Wei, W. (2022). Learning human-to-robot dexterous handovers for anthropomorphic hand. *IEEE Transactions on Cognitive and Developmental Systems*.

Eloff, K. M., & Engelbrecht, H. A. (2021). Toward collaborative reinforcement learning agents that communicate through text-based natural language. *Proc. of the Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 1–6.

Hand, R. N. (2024). *Robot nano hand*. https://robotnanohand.com (accessed: 12.06.2024).

InvenSense. (2024). *Mpu-9250 datasheet*. https://invensense.tdk.com/download-pdf/mpu-9250-datasheet/ (accessed: 12.06.2024).

Justa, J., Šmídl, V., & Hamáček, A. (2020). Fast ahrs filter for accelerometer, magnetometer, and gyroscope combination with separated sensor corrections. *Sensors*, *20*(14), 3824.

MindRove. (2024). *Mindrove armband surface emg (electromyography) device*. https://mindrove.com/armband/ (accessed: 12.06.2024).

Moosavi, S. K. R., Zafar, M. H., & Sanfilippo, F. (2022). A review of the state-of-the-art of sensing and actuation technology for robotic grasping and haptic rendering. *Proc. of the 5th International Conference on Information and Computer Technologies (ICICT)*, 182–190.

Natarajan, M., Seraj, E., Altundas, B., Paleja, R., Ye, S., Chen, L., Jensen, R., Chang, K. C., & Gombolay, M. (2023). Human-robot teaming: Grand challenges. *Current Robotics Reports*, *4*(3), 81–100.

Obaigbena, A., Lottu, O. A., Ugwuanyi, E. D., Jacks, B. S., Sodiya, E. O., & Daraojimba, O. D. (2024). Ai and human-robot interaction: A review of recent advances and challenges. *GSC Advanced Research and Reviews*, *18*(2), 321–330.

Pedrocchi, A., Ferrante, S., Ambrosini, E., Gandolla, M., Casellato, C., Schauer, T., Klauer, C., Pascual, J., Vidaurre, C., Gföhler, M., et al. (2013). Mundus project: Multimodal neuroprosthesis for daily upper limb support. *Journal of neuroengineering and rehabilitation*, *10*, 1–20.

Robotics, T. (2024). *Viperx 300 s*. https://www.trossenrobotics.com/viperx-300 (accessed: 12.06.2024).

Rodrigues, I. R., Barbosa, G., Oliveira Filho, A., Cani, C., Sadok, D. H., Kelner, J., Souza, R., Marquezini, M. V., & Lins, S. (2022). A new mechanism for collision detection in human–robot collaboration using deep learning techniques. *Journal of Control, Automation and Electrical Systems*, 1–13.

Sanfilippo, F., Blažauskas, T., Girdžiūna, M., Janonis, A., Kiudys, E., & Salvietti, G. (2021). A multi-modal auditory-visual-tactile e-learning framework. *International Conference on Intelligent Technologies and Applications*, 119–131.

Sanfilippo, F., & Pacchierotti, C. (2020). A low-cost multi-modal auditory-visual-tactile framework for remote touch. *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, 213–218.

Semeraro, F., Griffiths, A., & Cangelosi, A. (2023). Human–robot collaboration and machine learning: A systematic review of recent research. *Robotics and Computer-Integrated Manufacturing*, *79*, 102432.

s.r.o., J. (2024). *Pycharm integrated development environment (ide)*. https://www.jetbrains.com/pycharm/ (accessed: 12.06.2024).

Zhou, C., Peers, C., Wan, Y., Richardson, R., & Kanoulas, D. (2022). Teleman: Teleoperation for legged robot loco-manipulation using wearable imu-based motion capture. *arXiv preprint arXiv:2209.10314*.