Interpretable Multimodal Gesture Recognition for Drone and Mobile Robot Teleoperation via Log-Likelihood Ratio Fusion

Seungyeol Baek[1], Jaspreet Singh[2], Lala Ray[2,3], Hymalai Bello[3], Paul Lukowicz[2,3], and Sungho Suh[1]

***Abstract*— Human operators are still frequently exposed to hazardous environments such as disaster zones and industrial facilities, where intuitive and reliable teleoperation of mobile robots and Unmanned Aerial Vehicles (UAVs) is essential. In this context, hands-free teleoperation enhances operator mobility and situational awareness, thereby improving safety in hazardous environments. While vision-based gesture recognition has been explored as one method for hands-free teleoperation, its performance often deteriorates under occlusions, lighting variations, and cluttered backgrounds, limiting its applicability in real-world operations. To overcome these limitations, we propose a multimodal gesture recognition framework that integrates inertial data (accelerometer, gyroscope, and orientation) from Apple Watches on both wrists with capacitive sensing signals from custom gloves. We design a late fusion strategy based on the log-likelihood ratio (LLR), which not only enhances recognition performance but also provides interpretability by quantifying modality-specific contributions. To support this research, we introduce a new dataset of 20 distinct gestures inspired by aircraft marshalling signals, comprising synchronized RGB video, IMU, and capacitive sensor data. Experimental results demonstrate that our framework achieves performance comparable to a state-of-the-art vision-based baseline while significantly reducing computational cost, model size, and training time, making it well suited for real-time robot control. We therefore underscore the potential of sensor-based multimodal fusion as a robust and interpretable solution for gesture-driven mobile robot and drone teleoperation.**

***Index Terms*— Multimodal Gesture Recognition, Wearable Sensors, Log-Likelihood Ratio Fusion, Drone and Mobile Robot Teleoperation**

I. Introduction

Despite rapid advances in automation, many hazardous tasks in domains such as industrial manufacturing, construction, and emergency response are still performed directly by human workers. Firefighters, rescue teams, and industrial operators remain regularly exposed to extreme heat, toxic gases, unstable structures, and other life-threatening conditions. To reduce human risk, mobile robots and Unmanned Aerial Vehicles (UAVs) are increasingly deployed as substitutes or assistants in such environments, supporting tasks such as surveillance in confined industrial facilities [1], [2], industrial automation [3], delivery of supplies in disaster zones [4], and emergency rescue operations [5], [6]. Although these systems are expected to operate where humans cannot remain safely or efficiently, they often still rely on effective human teleoperation to ensure reliable task execution.

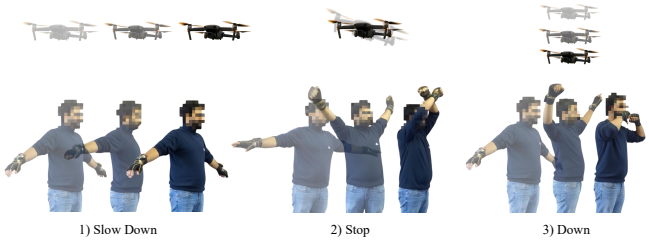


Fig. 1: An example of drone control using the proposed sensor-based gesture recognition.

Teleoperation methods provide operators with a means to control robots remotely, bridging the gap between human decision-making and robotic actuation. Conventional approaches rely on physical controllers such as joysticks and consoles [7], [8], which remain the standard in many industrial and defense applications. However, these devices restrict operator mobility and demand continuous manual engagement, limiting their suitability in highly dynamic or dangerous scenarios. As an alternative, gesture-based control has gained attention for its intuitive, hands-free, and natural interaction paradigm [9], [10]. By leveraging human body movements, operators can issue commands while maintaining situational awareness and physical flexibility—critical requirements in real-world deployments.

A large body of work has explored vision-based gesture recognition, driven by the maturity of computer vision methods and the availability of visual sensors. Yet vision-based systems remain highly sensitive to environmental conditions: occlusions, poor lighting, camera placement, and background clutter often degrade recognition performance [11]. In real operational settings such as smoke-filled disaster zones, dark industrial environments, or outdoor missions with variable lighting, these limitations hinder reliability. Consequently, vision-only systems struggle to meet the robustness required for mission-critical human–robot interaction.

Sensor-based gesture recognition using wearable devices offers a promising solution [12]. By capturing motion and physiological signals directly from the operator's body, wearable sensors provide robustness against environmental disturbances, enabling reliable operation where vision-based methods fail. However, single-modality systems often fall short of capturing the full richness of human gestures. Fur-

[1]S. Baek and S. Suh are with the Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea {mbaek01, sungho_suh}@korea.ac.kr

[2] J. Singh, P. Lukowicz are with the Department of Computer Science, RPTU Kaisersalutern-Landau, Kaiserslautern, Germany

[3] L.S.S. Ray, H. Bello, and P. Lukowicz are with the Embedded Intelligence, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany {lala_shakti_swarup.ray, hymalai.bello, paul.lukowicz}@dfki.de

Code and additional materials: https://github.com/mbaek01/Gesture-Recognition-for-Drone-Control.

thermore, many multimodal fusion techniques treat modality integration as a black box, offering little transparency regarding how different sensors contribute to recognition, a serious concern for safety-critical robotic control.

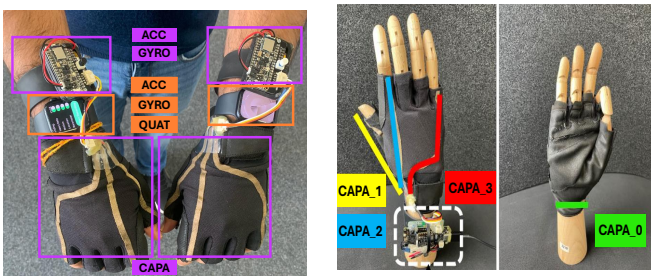
To address these challenges, we propose a multimodal gesture recognition framework tailored for mobile robots and drones. Our framework integrates accelerometer, gyroscope, and orientation data from Apple Watches worn on both wrists, together with capacitive sensing from our custom gloves [13]. These heterogeneous signals are combined through a log-likelihood ratio (LLR)–based late fusion strategy, which not only improves recognition accuracy but also provides interpretability by quantifying modality-specific contributions. This property is essential in robotic applications where operators and system designers must understand which signals drive the model's decisions.

In support of this framework, we introduce a novel dataset of 20 distinct gestures inspired by aircraft marshalling signals [14], a standardized communication protocol originally developed for ground–aircraft coordination. The dataset captures synchronized RGB video, wearable inertial data, and capacitive signals across both hands, making it a unique resource for evaluating multimodal gesture recognition models. Beyond drone control, the dataset and framework generalize to other mobile robot interaction scenarios, supporting broader adoption of gesture-based control systems.

Finally, we evaluate multiple fusion strategies, including LLR and self-attention, and conduct interpretability analysis and ablation studies to assess the importance of each modality. Comparisons with the state-of-the-art (SoTA) vision-based method PoseConv3D [15] show that our sensor-based multimodal approach outperforms the vision-based method while reducing computational cost, model size, and training time, making it suitable for real-time robotic deployment.

Our contributions are summarized as follows:

- We propose a deep learning framework that integrates heterogeneous wearable sensor modalities through a log-likelihood ratio (LLR)–based late fusion strategy, combining accuracy with interpretability.
- We provide interpretability analyses of fusion methods, leveraging LLR values and attention weights to quantify modality-specific contributions to recognition.
- We introduce a novel multimodal dataset for mobile robot and drone control, comprising 20 distinct gestures inspired by aircraft marshalling signals, with synchronized RGB video, inertial, and capacitive sensor data.
- We conduct ablation studies to evaluate robustness under missing-modality conditions and to highlight the role of different sensors.
- We show that sensor-based multimodal approaches achieve performance comparable to SoTA vision-based methods, while reducing computational cost, model size, and training time for practical robotic deployment.



(a) Textile sensing gloves and Apple Watch devices

(b) Capacitive sensor channels on the gloves [13]

Fig. 2: Hardware setup with textile sensing gloves and Apple Watch devices on both hands is shown in Fig. 2a. Purple outlines mark capacitive sensors (bottom) and IMU sensors (top), while orange outlines mark Apple Watch devices on the wrist used for IMU sensing and quaternion orientation.

II. Related Works

A. Applications of Mobile Robotics

Mobile robots and UAVs are increasingly deployed in safety-critical domains such as disaster response, industrial inspection, and logistics [16], [17]. For example, mobile robots are used for monitoring and surveying hazardous industrial facilities, detecting safety risks at construction sites, and enabling smart manufacturing through robot-assisted production. UAVs and drones are also being explored for contactless delivery of medical supplies, disaster detection, and emergency response [18].

Despite advances in autonomy, many of these tasks still require reliable human oversight and control to ensure safe and effective operation [19]. This makes intuitive and robust teleoperation methods critical for real-world deployment.

B. Control Methods for Mobile Robotics

Traditional robot teleoperation relies on physical interfaces such as joysticks, keyboards, and specialized control consoles, which remain standard in industrial applications. While these devices provide precision, they limit operator mobility and require continuous manual engagement, reducing effectiveness in dynamic and hazardous environments [20], [21]. Alternative approaches have explored mixed reality interfaces for more flexible robot control [22], [23]. However, these systems often demand cumbersome infrastructure and remain difficult to scale in the field. Overall, existing control methods highlight a trade-off between precision and operator flexibility, motivating the search for more natural, hands-free interaction paradigms.

C. Gesture Recognition as a Control Method

Gesture recognition has been widely studied as a natural interface for human–computer interaction, enabling intuitive control in domains such as augmented and virtual reality [24], gaming [25], and smart environments [26]. Vision-based approaches have shown strong performance in controlled laboratory settings, but their reliability degrades under occlusion, variable lighting, and cluttered backgrounds [27].

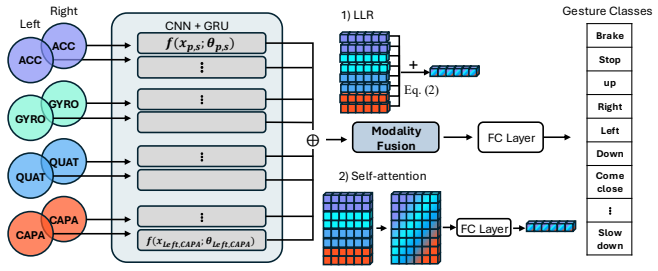


Fig. 3: Framework architecture for sensor-based gesture recognition on our dataset. Each sensor modality is processed individually through convolutional and temporal feature extraction layers, and the resulting modality-specific features are fused using either log-likelihood ratio (LLR) fusion Eq. (2) or self-attention fusion for classification.

To address these limitations, wearable sensor-based methods have gained attention, leveraging accelerometers, gyroscopes, and capacitive sensing to directly capture body movements [13]. These approaches have demonstrated robustness in human activity recognition tasks, where environmental variability poses significant challenges. More recent work has focused on multimodal fusion to exploit the complementary strengths of different sensors, though many fusion strategies still operate as black boxes, offering little interpretability of modality contributions [28].

While these techniques were originally developed for broader HCI and wearable computing applications, their advantages naturally extend to human–robot interaction. Gesture-based teleoperation of UAVs and multi-robot systems illustrates the potential of adapting these recognition methods to safety-critical robotic contexts, where robustness, efficiency, and interpretability are essential.

III. Methodology

A. Hardware Setting

The dataset was collected using a multimodal wearable setup consisting of textile sensing gloves, Apple Watch devices, and an RGB camera (Fig. 2). The gloves, adapted from a previously developed design [13], integrate stretchable textile electrodes and an inertial measurement unit (IMU) into lightweight sports gloves. Four capacitive channels were distributed across the fingers and wrist, connected to an FDC2214 capacitance-to-digital converter, with signals sampled at approximately 50 Hz. Each glove included a wrist-mounted IMU to capture acceleration and angular velocity.

To complement the glove-based signals, Apple Watch devices (Series 7) were worn on both wrists, providing accelerometer, gyroscope, and orientation data at approximately 100 Hz. A ZED Mini stereo camera recorded synchronized RGB videos at 30 fps, ensuring alignment between visual and wearable modalities. This combined setup—capacitive glove electrodes, inertial wrist sensors, Apple Watches, and RGB video—enables comprehensive multimodal capture of hand and arm gestures for drone and robot teleoperation.

B. Network Architecture

The framework for multi-sensor gesture recognition processes each sensor modality individually to obtain its feature representation. Fig. 3 provides an overview of the entire network architecture, where we define the set of modality pairs as $\mathcal{M} = \{(p,s) \mid p \in \{\text{Left}, \text{Right}\}, s \in \{\text{ACC}, \text{GYRO}, \text{QUAT}, \text{CAPA}\}\}$, with ACC, GYRO, QUAT, and CAPA denoting the accelerometer, gyroscope, quaternion orientation, and capacitive sensor, respectively. The input sensor data are represented as $X_{p,s} \in \mathbb{R}^{T_{p,s} \times C_{p,s}}$, where $T_{p,s}$ denotes the temporal sliding window size and $C_{p,s}$ the number of channels for modality (p,s). Each input sequence is processed independently by a feature extraction pipeline comprising convolutional and temporal layers to produce latent representations. These extracted features are subsequently fused using one of two approaches: log-likelihood ratio (LLR) fusion or self-attention fusion.

1) Feature Extraction per Modality: Feature extraction for each modality is carried out in two stages: an initial convolutional subnet for local feature extraction, followed by a temporal subnet based on gated recurrent units (GRUs) [29]. Inspired by human activity recognition (HAR) models such as DeepConvLSTM [30] and TinyHAR [31], we employ four convolutional subnets for initial feature extraction, one per modality. Each subnet consists of a 1D convolution layer applied along the temporal axis, followed by ReLU activations and batch normalization. The output shape of the convolutional subnet $\mathbb{R}^{T^*_{p,s} \times D}$, where $T^*_{p,s}$ indicates the reduced temporal dimension and D the output channel dimension.

The extracted features from each modality are then passed independently through an attention-based GRU subnet to capture temporal dependencies. The subnet consists of a two-layer GRU that outputs hidden states at every timestep. Following prior work [32], [31], rather than relying solely on the final hidden state as the temporal representation, we apply a self-attention mechanism to compute a weighted sum of all hidden states as a global temporal context representation. This global temporal representation is multiplied by a trainable gating parameter, allowing the model to learn whether to emphasize or disregard the global representation, and is then added to the hidden state of the last timestep [31]. Therefore, the output of this temporal subnet has shape $\mathbb{R}^D$.

2) Late Fusion Approaches (LLR and Self-Attention): We then apply either LLR or self-attention fusion to combine the per-modality features. In the LLR approach, the likelihood that a feature from a given modality belongs to a specific class is first computed. Denoting the per-modality feature as $h_{p,s} \in \mathbb{R}^D$, the predicted softmax output for the modality (p,s) is $\hat{y}_{p,s}$, defined over a set of N classes $\{C_i, ..., C_n\}$. The estimated probability that $\hat{y}_{p,s}$ belongs to class C_i, denoted $\hat{p}(\hat{y}_{p,s} \in C_i)$, is used to compute the LLR for that class as:

$$LLR_{p,s}(\hat{y}_{p,s} \in C_i) = \ln\left(\frac{\hat{p}(\hat{y}_{p,s} \in C_i)}{\hat{p}(\hat{y}_{p,s} \notin C_i)}\right) = \ln\left(\frac{\hat{p}(\hat{y}_{p,s} \in C_i)}{\sum_{\substack{j=1 \\ j\neq i}}^{N} \hat{p}(\hat{y}_{p,s} \in C_j)}\right) \quad (1)$$

The computed LLRs from all modality pairs form a tensor of shape $\mathbb{R}^{|\mathcal{M}|\times D}$, where $|\mathcal{M}| = 8$ is the number of modality pairs. Summing over all modality pairs $(p, s) \in \mathcal{M}$ yields the fused representation of shape $\mathbb{R}^D$:

$$LLR_{\text{fused}} = \sum_{(p,s)\in\mathcal{M}} LLR_{p,s}. \quad (2)$$

The self-attention fusion module operates on the concatenated features from all modality pairs, represented as a matrix of shape $\mathbb{R}^{|\mathcal{M}|\times D}$. Each row corresponds to the GRU-encoded representation of one modality pair, with no sequence dimension. Inspired by [31], [33], [34], we apply a scaled dot-product self-attention mechanism across the modality dimension to model inter-modality dependencies. The self-attention mechanism assigns relative importance to each modality by computing its similarity with all other modalities; these similarity scores are normalized into attention weights that determine how strongly each modality aggregates information from the others. Using these weights, each modality feature integrates information from the other modalities to produce a fused representation of shape $\mathbb{R}^{|\mathcal{M}|\times D}$. Finally, a fully connected layer reduces the modality dimension, mapping the fused representation from $\mathbb{R}^{|\mathcal{M}|\times D}$ to a single vector in $\mathbb{R}^D$.

The outputs of the LLR and self-attention fusion modules therefore share the same shape, $\mathbb{R}^D$, which is further projected to $\mathbb{R}^N$ by an additional fully connected layer for final classification. To preserve a direct and interpretable relationship between the learned features and the final class predictions, no nonlinear activation is applied to the fusion outputs [35].

C. Model Interpretability

We focus on post-hoc interpretability at the fusion stage, examining how LLR values and attention weights reflect modality contributions. Each fusion method provides interpretable signals of modality influence. Pre-fusion LLR values indicate each modality's relative contribution to the class prediction. Similarly, attention weights are examined to identify modality interactions indicated by high weights. Although attention weights do not directly quantify contributions, they provide insight into how strongly a query modality attends to others when forming a prediction [34], [36]. We visualize these attention weights as heatmaps for inter-modality dependencies, alongside the per-modality LLR contributions. Additionally, we conduct an ablation study—evaluating performance by excluding sensors and testing different combinations—to further assess modality contributions.

TABLE I: Gesture Label Mapping

Gesture Label	ID
Brake	0
Brake Fire Left	1
Brake Fire Right	2
Come Close	3
Cut Engine Left	4
Cut Engine Right	5
Down	6
Engine Start Left	7
Engine Start Right	8
Follow	9
Left	10
Move Away	11
Negative	12
Release Brake	13
Right	14
Slow Down	15
Stop	16
Straight	17
Take Photo	18
Up	19
Null Class	20

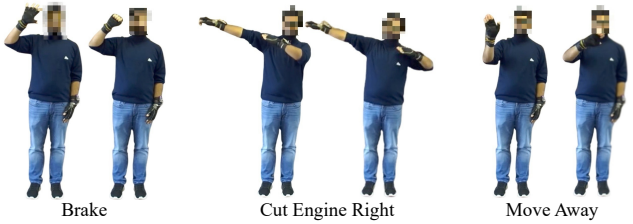


Fig. 4: Example gesture classes are illustrated, including 'brake,' 'brake fire right,' and 'move away.' Each gesture is represented by two sequential image frames, displaying a non-mirrored, first-person perspective.

IV. Experimental Setting

A. Gesture Definition

We define 20 distinct gesture classes, inspired by aircraft marshalling signals [14]. The gesture set, along with the corresponding label map ID used for classification, is summarized in I. Gestures were selected for their distinctive hand and arm motions to leverage the gloves' textile capacitive sensors and the wrist-mounted IMU sensors. They were also designed for visibility and clarity across diverse environmental conditions. Example gestures are shown in Fig. 4. While the gestures are well-suited and intuitive for aerial vehicles such as unmanned aerial vehicles (UAVs) due to their similarity in movement patterns to aircraft marshalling signals [14], their general control patterns are applicable to a broader range of mobile robotic platforms. A comprehensive video of the defined gestures will be included as the accompanying video.

B. Data Collection

The dataset was collected from eleven participants, comprising seven males and four females. They received instructions and demonstrations of the defined gestures, with natural variations permitted. Each participant completed five sessions. In each session, they performed 80 gesture instances, consisting of four randomized repetitions of each

gesture. The recorded sensor channels from each device are summarized in Table II and Table III. To synchronize modalities, five claps at the beginning and end of each session were used as reference points between the video and sensor streams. A temporal scaling factor—defined as the ratio between the video sequence length and the sensor sequence length—was applied to compress or stretch sensor data to match video duration. All gesture segments were labeled using LabelStudio [37].

Several preprocessing steps were applied to the raw sensor data before training the gesture recognition model. Sensor channels from the accelerometer, gyroscope, quaternion orientations, and capacitive sensors were used for model training. IMU data were obtained from the Apple Watch to ensure alignment with the quaternion orientations. Capacitive sensors from the gloves were included as additional inputs, while the glove IMU data served only as a complement to the Apple Watch IMU. Standardization was applied to each sensor device's raw data to mitigate inter-device variability and measurement noise. The sensor data were then segmented using a sliding-window approach, where each window corresponded to a fixed-length segment of sensor readings. A window size of 3.0 seconds was used to capture sufficient temporal context for gesture recognition, with a step size of 1.0 seconds to maintain continuity between windows. Each window was labeled by a threshold-voting mechanism, which determined its label based on the proportion of overlap between the window and the fully annotated gesture sequence. A threshold of 0.75 was employed, and windows that did not meet the threshold for any class were labeled as `Null`. Windows labeled as `Null` were excluded, ensuring that the model was trained exclusively on valid gestures. All windows containing claps for synchronization were also removed before training.

C. Training Details

Training was conducted using both leave-one-session-out(LOSO) and leave-one-participant-out(LOPO) strategies across five sessions and six participants. The dataset splits were generated from the remaining six participants after excluding all sessions identified as corrupted and unsuitable for training. Models were trained using cross-entropy loss for multi-class gesture classification. Training was performed with the Adam optimizer [38], using a learning rate of 1×10^{-4} and beta values of $(0.9, 0.999)$. Models were trained for a maximum of 150 epochs with a batch size of 32 on a single NVIDIA GeForce RTX 4090 24GB GPU. Early stopping was applied based on the F1-score of the validation set. Model performance on gesture recognition was evaluated using precision, recall, and macro F1-score. The proposed multimodal dataset and associated code for preprocessing and model implementation will be released publicly as supplementary materials upon acceptance.

TABLE II: Glove Sensor Channels

Sensor Type	Feature Names
Capacitive Sensors (CAPA)	CAPA_0, CAPA_1, CAPA_2, CAPA_3
Accelerometer (ACC)	ACC_X, ACC_Y, ACC_Z
Gyroscope (GYRO)	GYRO_X, GYRO_Y, GYRO_Z

TABLE III: Apple Watch Sensor Channels

Sensor Type	Feature Names
Accelerometer (ACC)	ACC_X, ACC_Y, ACC_Z
Gyroscope (GYRO)	GYRO_X, GYRO_ GYRO_Z
Quaternion Orientation (QUAT)	QUAT_W, QUAT_X, QUAT_Y, QUAT_Z

TABLE IV: Model Performance Comparison

Model	F1 Score (%)	Precision (%)	Recall (%)
(LOSO Split)			
LLR Fusion (sensor)	**95.40 ± 5.91**	**95.99 ± 4.84**	**95.52 ± 5.85**
Self-Attention Fusion (sensor)	94.42 ± 5.20	95.05 ± 4.44	94.45 ± 5.09
PoseConv3D (video)	94.70 ± 2.33	95.03 ± 1.93	94.73 ± 2.35
(LOPO Split)			
LLR Fusion (sensor)	**93.59 ± 5.20**	94.45 ± 4.70	93.59 ± 5.31
Self-Attention Fusion (sensor)	92.97 ± 6.11	94.40 ± 4.73	93.10 ± 5.98
PoseConv3D (video)	93.39 ± 5.44	**95.09 ± 3.40**	**93.75 ± 4.91**

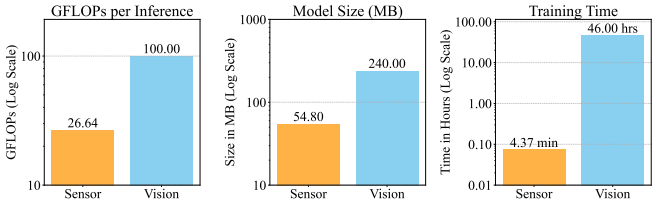


Fig. 5: Computational resource comparison between sensor-based (orange) and vision-based (blue) models. From left to right, the plots compare GFLOPs per inference, model size, and training time. Training time is reported as the average across all dataset split variations.

V. Results

A. Quantitative Evaluation with Comparisons

We evaluate two fusion methods in our sensor-based model and a single vision-based approach (PoseConv3D) using F1-score, precision, and recall, with results shown in Table IV. Under the LOSO split, the sensor-based LLR fusion model achieved the highest scores across all metrics. Under the LOPO split, it outperformed PoseConv3D in F1-score, while showing comparable precision and recall with only minor differences. Overall, metrics in the LOPO split were lower than in the LOSO split, suggesting that LOPO is a more challenging task. In addition, the sensor-based model was computationally more efficient, requiring fewer GFLOPs per inference, a smaller model size, and shorter training time,

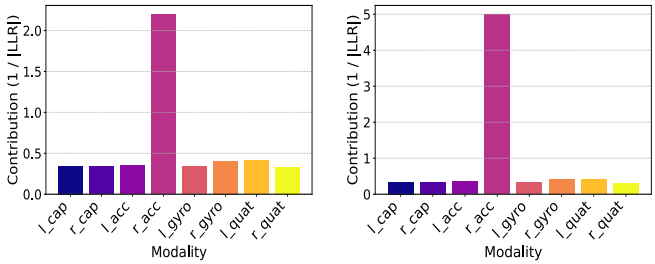


Fig. 6: Randomly sampled LLR contribution per modality, when correctly predicted `Move Away` gesture

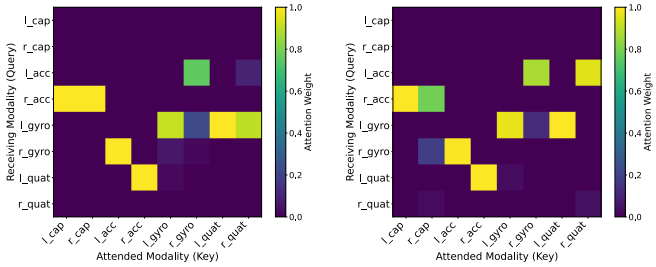


Fig. 7: Randomly sampled attention weights for a case when correctly predicted `Move Away` gesture

as shown in Fig. 5. For resource usage comparison, the LLR fusion was used to represent the sensor-based approach, while PoseConv3D represented the vision-based approach.

B. Qualitative Evaluation for Interpretability

Model interpretability was assessed using models trained on the LOPO split. From each fusion method, LLR values per modality and attention weights were randomly sampled during inference. Fig. 6 and Fig. 7 illustrate these values for correct predictions of the class `Move Away`. LLR values are transformed to the inverse of their magnitude so that contribution magnitudes are represented as positive rather than negative values. As shown in Fig. 4, this gesture involves waving the right hand horizontally with the palm facing forward. The predominantly linear motion of this gesture led to the hypothesis that accelerometer features would be especially relevant for correct classification. Fig. 6 shows that the right-hand accelerometer feature representation exhibited the highest LLR magnitude, indicating its strong contribution to correctly classifying the gesture as `Move Away`. All other modalities showed visibly lower magnitudes than that of the right-hand accelerometer. The self-attention fusion mechanism provides a more granular insight that complements the LLR findings. As shown in Fig. 7, the attention patterns also highlighted the right-hand accelerometer as a key component. The highest attention weights were associated with the right-hand accelerometer, whose representation attended most strongly to the capacitive sensor features, while also being the main target of attention from the left-quaternion orientation feature. While these patterns suggest that the model is learning relevant inter-feature relationships, we treat them as diagnostic, as high attention does not guarantee a causal role in the final prediction [36].

TABLE V: Performance Comparison in Modality Combinations in Terms of F1 Score (LOSO vs. LOPO)

Sensor Modalities	**F1 Score (%)**	
	LOSO	**LOPO**
(All 4 sensors)		
CAP + ACC + GYRO + QUAT	95.40 ± 5.91	93.59 ± 5.20
(3 sensors)		
CAP + ACC + GYRO	95.80 ± 5.61	91.78 ± 4.11
CAP + ACC + QUAT	95.24 ± 5.40	92.11 ± 4.52
ACC + GYRO + QUAT	95.75 ± 5.39	**93.75 ± 4.38**
(2 sensors)		
ACC + QUAT	95.30 ± 5.54	92.36 ± 4.92
ACC + GYRO	**95.81 ± 5.05**	92.00 ± 4.25
CAP + QUAT	93.88 ± 5.51	87.96 ± 7.23
CAP + ACC	94.66 ± 5.91	89.82 ± 5.01
(1 sensor)		
CAP	49.22 ± 8.99	7.98 ± 4.27
ACC	94.51 ± 5.89	90.05 ± 5.96
GYRO	95.21 ± 5.42	90.05 ± 6.70

C. Ablation Study

We conducted an ablation study to evaluate the impact of individual sensor modalities on model performance, quantified by the F1 score. This analysis employed our LLR fusion-based sensor approach and was evaluated under both LOSO and LOPO cross-validation splits.

In the LOSO splits, we observed minimal F1 score degradation across many sensor combinations, even when reducing the sensor count from four to two. This suggests that the gesture recognition task within the LOSO setting may be sufficiently simple, allowing models with fewer sensors to achieve performance comparable to those using the full sensor suite.

Conversely, the performance disparities between different sensor combinations were significantly more pronounced in the LOPO splits. This indicates that the LOPO evaluation presents a more challenging generalization task, where a reduced sensor count is often insufficient to match the performance of a richer sensor set. Notably, in the LOPO split, the top-performing combination resulted from excluding the capacitive sensor. In contrast, the same combination of capacitive sensors, accelerometer, and gyroscope ranked third in the LOSO split, but its F1 score was only marginally different from the other top-five combinations. These results suggest the LOSO split lacks the sensitivity to robustly differentiate the efficacy of various sensor combinations.

Further analysis of the capacitive sensor modality reinforces this finding. The model relying solely on capacitive sensors yielded significantly lower F1 scores than any other single-sensor or multi-sensor combination in both validation splits, with a particularly substantial performance degradation in the LOPO split. We hypothesize that this is due to our defined gesture set, which predominantly involves large-

scale arm and hand motions rather than the fine-grained, individuated finger movements that capacitive sensors are designed to detect.

This hypothesis is substantiated by the observation that the capacitive-only model collapsed its predictions to the `Take Photo` gesture. This specific gesture requires participants to "raise both hands in front of the face and form a rectangular frame by extending the index fingers vertically and the thumbs horizontally," a motion that uniquely engages the fingers in a static, articulated pose. The model's comprehensive failure to recognize other classes, which lack this intensive finger engagement, underscores the capacitive sensor's limited utility for our target gesture vocabulary.

D. Discussion

Our LLR-based multimodal framework addresses key limitations of gesture-driven teleoperation in hazardous, real-world deployments. By integrating heterogeneous wearable modalities into an interpretable and efficient model, the system enables reliable control in safety-critical scenarios such as disaster response and industrial maintenance, where understanding sensor-level contributions is essential for diagnosing failures and building operator trust. Unlike vision-based systems that degrade under smoke, low light, or occlusion, inertial and capacitive sensing remain robust in such environments. The gesture vocabulary, inspired by aircraft marshalling [14], further supports intuitive adoption in emergency operations without extensive retraining. Compared with vision-based baselines, the sensor-based framework reduces computational cost and model size, enabling real-time deployment on resource-constrained edge devices without reliance on cloud infrastructure, and extending battery life by lowering computational load and minimizing operational interruptions [39].

Despite strong performance, several limitations should be noted. Attention weights were used for model explainability, but they can be misleading [36]. To mitigate this, LLR values were incorporated to capture direct modality-level contributions, preserving a clear relationship between features and final predictions [35]. Additionally, several recording sessions were lost due to hardware issues, and all data were collected indoors, leaving robustness in uncontrolled environments uncertain. Future work should expand data collection to more diverse operational conditions, including outdoor, dynamic, and cluttered environments. The current implementation processes modality features sequentially, introducing avoidable latency during training and inference. Future work will adopt a parallel architecture in which unimodal encoders operate simultaneously, reducing inference delay and improving scalability for real-time robotic deployment.

Another limitation concerns the scope of the gesture set. While its foundation in aircraft marshalling protocols [14] is a key strength for standardization, this choice also means it requires further operation-specific adjustment to achieve universal applicability. This protocol is highly optimized for vehicle navigation (e.g., `Stop`, `Left`, `Right`), but it lacks the necessary vocabulary for the complex manipulation or instrumentation commands required in other domains. For broader adoption, the mapping between gestures and control commands must be adapted or extended. Future work should incorporate new, domain-specific gestures for tasks such as industrial inspection or emergency rescue.

Relatedly, the physical properties of our current gesture set introduced a sensor-specific limitation. As shown in our ablation study, the capacitive sensor's contribution was minimal, which we attribute to the vocabulary's emphasis on large-scale arm motions rather than fine finger articulation. Although they did not yield significant accuracy gains overall, the glove sensors provided complementary information in cases where wrist-based inertial signals were ambiguous. This limited improvement reflects the aircraft marshalling–inspired design, which prioritizes gross limb movements. Nonetheless, their inclusion highlights the framework's potential for future gesture dictionaries that involve finer finger-based interactions, where capacitive sensing is expected to play a more significant role. Expanding the gesture vocabulary to include actions requiring precise finger control—such as pinches, taps, or other subtle manipulations—could therefore unlock a substantial performance boost, with capacitive sensing providing unique and critical signals that enhance the model's discriminative power. Releasing this multimodal dataset further enables future studies to explore such extensions and develop sensor fusion strategies that more effectively combine wrist and glove modalities.

VI. Conclusions

In this work, we presented a multimodal gesture recognition framework for mobile robots and drones that integrates inertial signals from Apple Watches with capacitive sensing from textile gloves. Using a log-likelihood ratio (LLR)–based late fusion strategy, our method enhances recognition performance while providing interpretability through modality-specific contributions. We also introduced a dataset of 20 aircraft marshalling–inspired gestures with synchronized RGB, IMU, and capacitive data. Experiments showed that our approach outperforms vision-based baselines in F1 score while requiring lower computational cost and a smaller model size.

Future work will expand data collection to larger and more diverse settings, including continuous gesture streams, multi-robot coordination, and operational conditions beyond controlled indoor settings to further assess robustness. Beyond drone teleoperation, the proposed approach may extend to broader human–robot interaction scenarios where robustness, efficiency, and interpretability are critical for safe operation. This work advances practical, interpretable, and sensor-based gesture recognition for intuitive and reliable robot control in real-world environments.

ACKNOWLEDGMENT

This work was supported by the IITP grant funded by the MSIT (RS-2019-II190079, AI Graduate School Pro-

gram (Korea University) and IITP-2026-RS-2024-00436857, IITP-ITRC), and by the Technology Innovation Program (RS-2025-25453819) funded by the MOTIR. This work was also supported by the BMBF in the project CrossAct (01IW25001).

REFERENCES

[1] J. Azeta, C. Bolu, D. Hinvi, A. Abioye, H. Boyo, P. Anakhu, and P. Onwordi, "An android based mobile robot for monitoring and surveillance," *Procedia Manufacturing*, vol. 35, pp. 1129–1134, 2019.

[2] M. Z. Shanti, C.-S. Cho, B. G. de Soto, Y.-J. Byon, C. Y. Yeun, and T. Y. Kim, "Real-time monitoring of work-at-height safety hazards in construction sites using drones and deep learning," *Journal of safety research*, vol. 83, pp. 364–370, 2022.

[3] S. Wang, L. Jiang, J. Meng, Y. Xie, and H. Ding, "Training for smart manufacturing using a mobile robot-based production line," *Frontiers of Mechanical Engineering*, vol. 16, no. 2, pp. 249–270, 2021.

[4] J. Euchi, "Do drones have a realistic place in a pandemic fight for delivering medical supplies in healthcare systems problems?" pp. 182–190, 2021.

[5] F. Rubio, F. Valero, and C. Llopis-Albert, "A review of mobile robots: Concepts, methods, theoretical framework, and applications," *International Journal of Advanced Robotic Systems*, vol. 16, no. 2, p. 1729881419839596, 2019.

[6] A. Khan, S. Gupta, and S. K. Gupta, "Emerging uav technology for disaster detection, mitigation, response, and preparedness," *Journal of Field Robotics*, vol. 39, no. 6, pp. 905–955, 2022.

[7] T. Fong and C. Thorpe, "Vehicle teleoperation interfaces," *Autonomous robots*, vol. 11, no. 1, pp. 9–18, 2001.

[8] C. Cruz Ulloa, D. Domínguez, J. Del Cerro, and A. Barrientos, "A mixed-reality tele-operation method for high-level control of a legged-manipulator robot," *Sensors*, vol. 22, no. 21, p. 8146, 2022.

[9] K. B. de Carvalho, D. K. D. Villa, M. Sarcinelli-Filho, and A. S. Brandao, "Gestures-teleoperation of a heterogeneous multi-robot system," *The International Journal of Advanced Manufacturing Technology*, vol. 118, no. 5, pp. 1999–2015, 2022.

[10] Y. Ma, Y. Liu, R. Jin, X. Yuan, R. Sekha, S. Wilson, and R. Vaidyanathan, "Hand gesture recognition with convolutional neural networks for the multimodal uav control," in *2017 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS)*. IEEE, 2017, pp. 198–203.

[11] Y. Zhu, Z. Yang, and B. Yuan, "Vision based hand gesture recognition," in *2013 international conference on service sciences (ICSS)*. IEEE, 2013, pp. 260–265.

[12] S. Suh, V. F. Rey, S. Bian, Y.-C. Huang, J. M. Rožanec, H. T. Ghinani, B. Zhou, and P. Lukowicz, "Worker activity recognition in manufacturing line using near-body electric field," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 11 554–11 565, 2023.

[13] H. Bello, S. Suh, D. Geißler, L. S. S. Ray, B. Zhou, and P. Lukowicz, "Captainglove: Capacitive and inertial fusion-based glove for real-time on edge hand gesture recognition for drone control," in *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, 2023, pp. 165–169.

[14] C. Choi, J.-H. Ahn, and H. Byun, "Visual recognition of aircraft marshalling signals using gesture phase analysis," in *2008 IEEE Intelligent Vehicles Symposium*. IEEE, 2008, pp. 853–858.

[15] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[16] M. Moniruzzaman, A. Rassau, D. Chai, and S. M. S. Islam, "Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey," *Robotics and Autonomous Systems*, vol. 150, p. 103973, 2022.

[17] S. Opiyo, J. Zhou, E. Mwangi, W. Kai, and I. Sunusi, "A review on teleoperation of mobile ground robots: Architecture and situation awareness," *International Journal of Control, Automation and Systems*, vol. 19, no. 3, pp. 1384–1407, 2021.

[18] M. Chiou, N. Hawes, and R. Stolkin, "Mixed-initiative variable autonomy for remotely operated mobile robots," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–34, 2021.

[19] C. Zhou, C. Peers, Y. Wan, R. Richardson, and D. Kanoulas, "Teleman: Teleoperation for legged robot loco-manipulation using wearable imu-based motion capture," *arXiv preprint arXiv:2209.10314*, 2022.

[20] H. Esaki and K. Sekiyama, "Immersive robot teleoperation based on user gestures in mixed reality space," *Sensors*, vol. 24, no. 15, p. 5073, 2024.

[21] J. Lee, T. Lim, and W. Kim, "Investigating the usability of collaborative robot control through hands-free operation using eye gaze and augmented reality," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4101–4106.

[22] C. Cruz Ulloa, D. Domínguez, J. del Cerro, and A. Barrientos, "Analysis of mr–vr tele-operation methods for legged-manipulator robots," *Virtual Reality*, vol. 28, no. 3, p. 131, 2024.

[23] O. S. Adekoya, A. Sgorbissa, and C. T. Recchiuto, "Horus: A mixed reality interface for managing teams of mobile robots," *arXiv preprint arXiv:2506.02622*, 2025.

[24] T. Di Qi, F. L. Cibrian, M. Raswan, T. Kay, H. M. Camarillo-Abad, and Y. Wen, "Toward intuitive 3d interactions in virtual reality: A deep learning-based dual-hand gesture recognition approach," *IEEE Access*, vol. 12, pp. 67 438–67 452, 2024.

[25] A. Sharma, L. Verma, H. Kaur, A. Modgil, A. Soni *et al.*, "Hand gesture recognition gaming control system: Harnessing hand gestures and voice commands for immersive gameplay," in *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*. IEEE, 2024, pp. 101–107.

[26] C. Panagiotou, E. Faliagka, C. P. Antonopoulos, and N. Voros, "Multidisciplinary ml techniques on gesture recognition for people with disabilities in a smart home environment," *AI*, vol. 6, no. 1, p. 17, 2025.

[27] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: a review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581–1606, 2024.

[28] L. S. S. Ray, D. Geißler, M. Liu, B. Zhou, S. Suh, and P. Lukowicz, "Als-har: Harnessing wearable ambient light sensors to enhance imu-based human activity recognition," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 133–147.

[29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[30] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[31] Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, and M. Beigl, "Tinyhar: A lightweight deep learning model designed for human activity recognition," in *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 89–93.

[32] H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu, "Attnsense: Multi-level attention mechanism for multimodal human activity recognition." in *IJCAI*, 2019, pp. 3109–3115.

[33] A. Abedin, M. Ehsanpour, Q. Shi, H. Rezatofighi, and D. C. Ranasinghe, "Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–22, 2021.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[36] S. Serrano and N. A. Smith, "Is attention interpretable?" *arXiv preprint arXiv:1906.03731*, 2019.

[37] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2025, open source software available from https://github.com/HumanSignal/label-studio. [Online]. Available: https://github.com/HumanSignal/label-studio

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] A. Romero, C. Delgado, L. Zanzi, R. Suárez, and X. Costa-Pérez, "Cellular-enabled collaborative robots planning and operations for search-and-rescue scenarios," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5942–5948.