Gesture-Controlled Aerial Robot Formation for Human-Swarm Interaction in Safety Monitoring Applications

Vít Krátký[1⋆], Giuseppe Silano[1], Matouš Vrba[1], Christos Papaioannidis[2], Ioannis Mademlis[2], Robert Pěnička[1], Ioannis Pitas[2], and Martin Saska[1]

***Abstract*— This paper presents a formation control approach for contactless gesture-based Human-Swarm Interaction (HSI) between a team of multi-rotor Unmanned Aerial Vehicles (UAVs) and a human worker. The approach is intended for monitoring the safety of human workers, especially those working at heights. In the proposed dynamic formation scheme, one UAV acts as the leader of the formation and is equipped with sensors for human worker detection and gesture recognition. The follower UAVs maintain a predetermined formation relative to the worker's position, thereby providing additional perspectives of the monitored scene. Hand gestures allow the human worker to specify movements and action commands for the UAV team and initiate other mission-related commands without the need for an additional communication channel or specific markers. Together with a novel unified human detection and tracking algorithm, human pose estimation approach and gesture detection pipeline, the proposed approach forms a first instance of an HSI system incorporating all these modules onboard real-world UAVs. Simulations and field experiments with three UAVs and a human worker in a mock-up scenario showcase the effectiveness and responsiveness of the proposed approach.**

***Index Terms*— Aerial Systems: Applications, Human-Swarm Interaction, Multi-Robot Systems.**

Supplementary Material

Video: `https://mrs.felk.cvut.cz/gestures2024`

I. Introduction

The multi-rotor Unmanned Aerial Vehicles (UAVs) applied in challenging-to-access real-world work environments such as wind turbines [1], large construction sites [2], and power transmission lines [3], prove to be exceptionally beneficial. The introduction of UAVs as *robotic co-workers* [4] in these settings offers numerous benefits, including the ability to access locations that are challenging for humans to reach, assist in tool handling, monitor workers' safety, and reduce the physical and cognitive workload imposed on the human workers [5], [6]. Within the context of the European AERIAL-CORE project[1], the application for safety monitoring is driven by the observation that violations of safety protocols are a primary cause of fatal injuries during maintenance tasks on electric power infrastructures. To tackle this issue, the concept of *Aerial Co-Workers* (ACWs) has been developed [7], encompassing three roles: the inspection-ACW [3], the safety-ACW [8], and the physical-ACW [9]. These roles constitute key components of future human-robot missions aimed at maintaining electric power transmission infrastructures and, more broadly, the entire energy system.

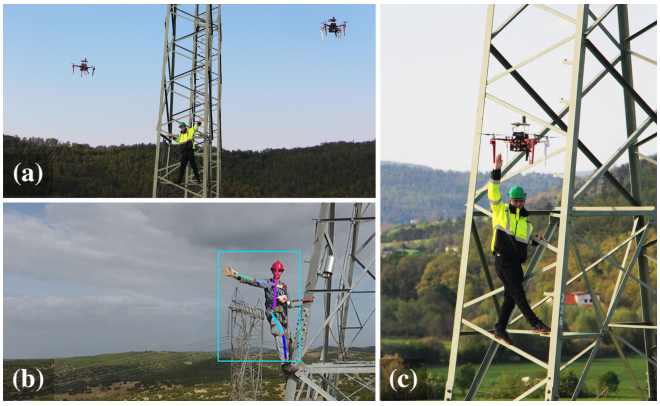

Fig. 1: The gesture-based interaction between a human worker and a team of UAVs using the proposed system (a, c) with an example output of the developed gesture recognition pipeline overlaid on the corresponding input video frame (b).

Safety monitoring applications require the human operator, tasked with situation assessment, to be provided with a comprehensive view of the scene. The operator greatly benefits from the capability to adapt this view interactively, with their situation awareness significantly enhancing as the number of simultaneous scene perspectives increases. This fact underpins the utility of deploying multiple UAVs for this purpose. However, when monitoring human workers, it is vital to balance the operator's preferences with the monitored individuals' safety and comfort, ensuring their performance is not adversely affected by the UAVs' proximity.

This study introduces an approach for UAV formation control in contactless Human-Swarm Interaction (HSI), focusing on multi-rotor UAV teams. Leveraging gesture-based controls, our approach aims to improve situational awareness and facilitate precise command execution in real-world scenarios, such as maintenance operations on electric power transmission infrastructures. The framework allows a remote operator to dynamically adjust the UAV formation to optimize observation angles, while a monitored worker can use gestures to request assistance or modify the UAVs' proximity for safety reasons. This dual-control mechanism ensures online system adaptation to operator needs and task context, empowering the monitored individual to influence

[1]Authors are with the Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic.

[2]Authors are with the Department of Informatics, Faculty of Sciences, Aristotle University of Thessaloniki, Greece.

⋆Corresponding author: `vit.kratky@fel.cvut.cz`

This work was partially funded by the EU's H2020 AERIAL-CORE grant no. 871479, by the CTU grant no. SGS23/177/OHK3/3T/13, and by the Czech Science Foundation (GAČR) grant no. 23-07517S.

[1]`https://aerial-core.eu`

TABLE I: Comparison of addressed features in related papers and our proposed approach: *included* (✓) and *not included* (✗).

Ref.	Features				
	Human Gestures	**Formation Control**	**Onboard Computation**	**Adaptive Parameters**	**Human Tracking**
[8]	✗	✓	✓	✗	✗
[14]	✓	✓	✗	✓	✗
[15]	✓	✗	✗	✗	✗
[16]	✓	✓	✗	✓	✗
[18]	✓	✓	✗	✗	✗
[20]	✗	✓	✓	✗	✗
[21]	✗	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓

UAVs behavior without additional equipment (e.g., wearable sensors). This capability is crucial for ensuring the worker's safety during unforeseen events, as it leverages their superior awareness of UAV proximity, nearby obstacles, and prevailing weather conditions compared to the remote operator.

The presented work bridges the gap in multi-UAV interactions with humans, introducing an innovative system that integrates advanced HSI features, combining vision and control strategies directly on UAV platforms for seamless and responsive collaboration. Validated through rigorous testing in both simulated and real-world outdoor conditions, as depicted in Figures 1 and 4, our system shows significant advancements in practical deployment, robustness with respect to changes in the environment, and potential to augment safety and efficiency in critical infrastructure maintenance and inspection tasks.

A. Related Work

While extensive research has been dedicated to collaborative and safe interactions involving human and ground robots, the methods for UAVs in this context are less developed [10]. Particularly, the dynamics of human interaction with multi-robot UAV teams present a significant research gap. Although there has been considerable advancement in computer vision and autonomous systems to facilitate human-UAV interaction, these efforts primarily concentrate on specific applications [11], [12]. Studies in computer vision have focused on recognizing human features, such as faces [13], hand gestures [14]–[16], hand motion [17], and body postures [18], with some examining gaze detection for robot selection [19]. Concurrently, research on UAV autonomy has tackled perception-aware control [20], formation control to enhance visibility [8], and optimization-based obstacle avoidance [21], aiming to improve UAVs' independent navigation and safety around humans and other UAVs. These contributions are vital in shaping UAVs' interaction capabilities. Yet, the integration of these technologies into cohesive systems for human-UAV teams in complex, real-world environments remains an area ripe for further exploration [22].

The presented studies tend to primarily focus either on the vision component, sometimes neglecting or oversimplifying vehicle dynamics [13]–[16], or on the control aspect, often abstracting the use of generic onboard sensors [8], [20], [21]. Failing to consider both vision and control aspects in the design of an HSI framework can result in severe failures. For instance, shortcomings in estimating human position (attributed to, e.g., unbalanced camera vibration or motion blur) can compromise system stability, leading to crashes and potentially endangering the operator. Furthermore, none of the aforementioned studies [8], [13]–[21] have addressed the challenge of integrating onboard gesture recognition modules within the UAV formation control scheme. Additionally, the majority of these methods [13]–[18], [20] are evaluated indoors without accounting for external elements, such as lighting conditions and wind gusts, and often rely on off-board computation that could introduce notable delays in real-world scenarios.

This work builds upon our previous studies [23]–[26] and advances beyond existing methodologies by designing algorithms that jointly address the human detection, pose estimation, gesture recognition, and UAV formation control that can be executed onboard light-weight UAVs as a unified system. Thus, it eliminates off-board processing latency and enables real-world deployment independent on external infrastructure. The HSI framework also incorporates safety features like obstacle avoidance, collision prevention with other UAVs, and compliance with distance regulations to ensure human comfort. For a comprehensive comparison of the features addressed in the related papers and in the proposed approach, we refer the reader to Table I.

B. Contributions

To the best of our knowledge, this work represents the first instance of a contactless HSI system involving a human and a team of multi-rotor UAVs that incorporates onboard human state estimation and gesture recognition, enabling dynamic and intuitive interaction between humans and UAVs in real-world applications. The presented work addresses the challenges hindering the application of existing UAV-based HSI studies [13]–[18], [20] in real-world applications through the following contributions. First, we propose a novel dynamic formation control strategy supporting online, on-demand adaptation of the shape of the UAV formation in complex environments. This enables rapid response to changes in the environment and provides a convenient way to control the relative positions of multiple robots to the human worker through operator's commands. Second, we design an onboard approach for recognizing 2D human body poses and hand gestures using a Deep Neural Network configuration that minimizes latency and eliminates the need for off-board computation, thereby enhancing the system's responsiveness. Further, we introduce a multi-modal approach to human pose estimation tailored for dynamic UAV systems, which operates independently of external infrastructure and can optionally function without additional equipment carried by the human. Such an approach ensures the system's applicability across diverse scenarios and environments. Lastly, we demonstrate that a system with advanced HSI capabilities directly combining vision and control strategies for immediate responsiveness can be implemented onboard lightweight UAV platforms.

II. Gesture-controlled Aerial Formation

The system architecture, as depicted in Fig. 2, comprises four layers: *Detection*, *Localization*, *Planning*, and *UAV Plant*. The *Detection* block interfaces directly with the human worker, translating hand gestures into commands for the UAV formation. An RGB-D camera captures images, enabling human detection, tracking, and gesture recognition (Section II-A). The *Localization* block combines sensor data from the UAV plant, including the vehicle's relative distance from the worker, with information from the *Detection* block and an Ultra Wide Bandwidth (UWB) module. This fusion provides inputs for a Kalman filter that estimates the human's 3D position and velocity for the formation controller (Section II-B). The *Planning* block generates feasible trajectories for the individual vehicles based on the status of the UAV formation leader, the requests from a remote human operator, the output of the gesture classifier, the human worker's state, and the status of other UAV team members obtained through a wireless network (Section II-C). Lastly, the *UAV Plant* receives and executes the trajectories, ensuring precise flight maneuvers [27].

A. Human detection and gesture recognition

RGB images from the onboard camera are processed during flight to detect and track human worker, leveraging the authors' prior work on Convolutional Neural Networks (CNNs) [23]. A fast deep neural object detector based on Single-Shot multibox Detector (SSD) [28] is employed along with a custom LDES-ODDA visual tracker [24]. The two components are combined in a novel unified detection-and-tracking configuration where detection and tracking are performed alternately, exploiting the advantages of both worlds towards achieving both increased accuracy and fast inference.

The output of this pipeline is a predicted bounding box for the tracked human in each input image where the human is visible, as shown in Fig. 1(b). These bounding boxes are then used for gesture recognition and human state estimation. To maximize accuracy, both the detector and the tracker were pretrained on a manually annotated dataset[2] and then fine-tuned using videos of a human operator wearing safety equipment. These videos were captured in diverse outdoor environments and lighting conditions.

Given a sequence of images captured by the RGB-D camera of the leader UAV and the corresponding bounding boxes of the tracked human, the developed gesture recognition module predicts the type of the gesture from a predefined set (e.g., extending one arm to the side) [29], [30]. The gesture recognition proceeds as a sequential pipeline. First, each video frame is cropped using the corresponding bounding box of the tracked human (see Fig. 1(b)). 2D skeletons, i.e., visible human body joints in pixel coordinates, are subsequently extracted from each cropped image via an enhanced version of our multi-branch CNN from [25], which has been crucially improved here by replacing the simple interbranch skip connections with more powerful cross-attention synapses [31]. The last N outputs of the skeleton extractor, covering N successive video frames, are stored in a FIFO buffer. This buffer is subsequently processed by our gesture classifier [26], a lightweight Long Short-Term Memory (LSTM) neural architecture that determines the type of the performed gesture based on a temporal sliding window of size N. The pipeline was trained on a large, manually annotated dataset of gestures[3], and fine-tuned to perform effectively on aerial images. The parameter N was empirically tuned to $N = 9$, based on the camera's update rate and the pipeline's performance when running onboard the UAVs.

The gesture recognition pipeline's output undergoes a post-processing step to improve HSI reliability by mitigating undesired shape adaptation from false positive detections. In each iteration, the K most recent valid measurements are considered, with older data filtered out beyond a time threshold, $t_c \in \mathbb{R}_{>0}$, to maintain relevance. The dominant gesture's ratio, $f_d \in [0,1]$, is computed. If it exceeds a predefined threshold $\Pi_d \in [0,1]$, the corresponding formation parameter is adjusted (Section II-C). However, a new adjustment can only occur after a time delay, $t_d \in \mathbb{R}_{>0}$, preventing repeated updates based on the same set of measurements. This enhances the worker's control and prevents unwanted shape adaptation. The values of t_c, Π_d, and t_d were determined through real-world experiments.

B. Human 3D position estimation

The estimated human's 3D position, denoted as ${}^H\mathbf{p} = [{}^Hp_x, {}^Hp_y, {}^Hp_z]^\top \in \mathbb{R}^3$, is derived from detections and available onboard sensors, and subsequently refined through a Kalman filter. We employ a constant velocity model within the Kalman filter, formulated as

$$\begin{bmatrix} {}^H\mathbf{p} \\ {}^H\mathbf{v} \end{bmatrix}_{[k+1]} = \begin{bmatrix} \mathbf{I}_3 & \Delta t\mathbf{I}_3 \\ \mathbf{0}_3 & \mathbf{I}_3 \end{bmatrix} \begin{bmatrix} {}^H\mathbf{p} \\ {}^H\mathbf{v} \end{bmatrix}_{[k]} + \boldsymbol{\varepsilon}_{[k]}, \tag{1}$$

$$\mathbf{z}_{[k]} = {}^H\mathbf{p}_{[k]} + \boldsymbol{\zeta}_{[k]}, \tag{2}$$

$$\boldsymbol{\varepsilon}_{[k]} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \qquad \boldsymbol{\zeta}_{[k]} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_{[k]}\right), \tag{3}$$

where the subscript $\bullet_{[k]}$ indicates the time step, ${}^H\mathbf{v} = [{}^Hv_x, {}^Hv_y, {}^Hv_z]^\top \in \mathbb{R}^3$ represents the human's velocity, $\mathbf{I}_3 \in \mathbb{R}^{3\times3}$ and $\mathbf{0}_3 \in \mathbb{R}^{3\times3}$ are the identity and zero matrices, respectively, Δt signifies the time step duration, and $\mathbf{z}$ is the measurement. The variables $\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}$ denote the process noise and the measurement noise, respectively, both assumed to follow a normal distribution with zero mean. The covariance matrices for these distributions are represented by $\mathbf{Q}$ for the process noise and $\boldsymbol{\Sigma}$ for the measurement noise. We define the matrix $\mathbf{Q}$ as

$$\mathbf{Q} = \mathrm{diag}\left(\sigma_{p_x}^2, \sigma_{p_y}^2, \sigma_{p_z}^2, \sigma_{v_x}^2, \sigma_{v_y}^2, \sigma_{v_z}^2\right), \tag{4}$$

where $\sigma_{p_\bullet}$ and $\sigma_{v_\bullet}$ denote empirically derived parameters for the human positional and velocity uncertainties, respectively.

[2] https://aiia.csd.auth.gr/open-multidrone-datasets

[3] https://aiia.csd.auth.gr/auth-uav-gesture-dataset

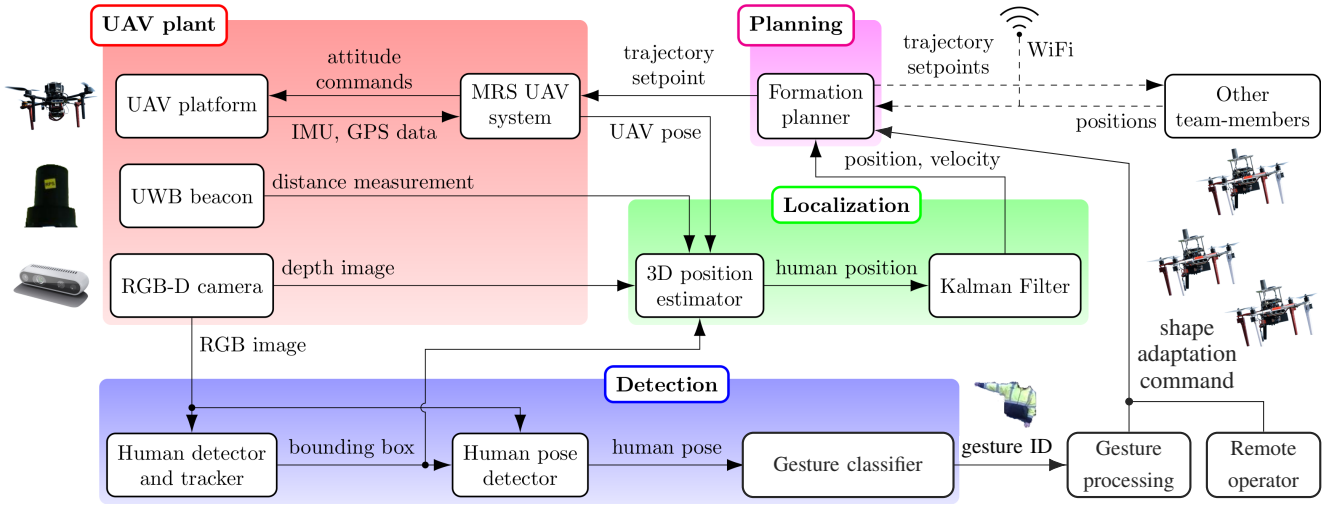

Fig. 2: A system architecture overview showing data exchange between blocks using arrows and highlighted layers.

To simplify our notation, we will henceforth not explicitly mention the time $\bullet_{[k]}$.

The measurement vector $\mathbf{z}$ is obtained using a unit vector $\vec{\mathbf{d}}$, which indicates the direction from the camera to the human, and a distance estimate ${}^{C}d$. This direction vector $\vec{\mathbf{d}}$ is determined by projecting the center of the bounding box using a calibrated camera projection model. The distance ${}^{C}d$ is computed by aggregating estimates from three sources:

1) *apparent distance* d_{apparent} calculated based on the apparent size in the image and the known physical height of the human, employing techniques from existing literature [32],
2) *stereo camera distance* d_{stereo} derived from the median of distance measurements within the bounding box captured by the stereo camera,
3) *UWB system distance* d_{UWB} obtained from an UWB system[4] mounted on the UAV and worn by the human.

Given the manufacturers' specifications, we assumed d_{UWB} to be generally more accurate than d_{stereo}, which, in turn, is deemed more accurate than d_{apparent}. However, the availability of UWB and stereo measurements may be inconsistent due to factors like limited range or absence of the UWB beacon and stereo camera, radio interference, and camera blur. To account for these limitations and utilize the best available data, we select the most reliable distance measurement for use, adjusting the covariance matrix $\mathbf{\Sigma}$ accordingly

$$\mathbf{z} = {}^{C}\mathbf{R}\left({}^{C}d\vec{\mathbf{d}}\right) + {}^{C}\mathbf{p}, \tag{5}$$

$$\left\{{}^{C}d,\ \mathbf{\Sigma}\right\} = \begin{cases} \left\{d_{\text{UWB}},\ {}^{C}\mathbf{R}\mathbf{\Sigma}_{\text{UWB}}{}^{C}\mathbf{R}^{\top}\right\}, & \text{if } d_{\text{UWB}} \text{ available,} \\ \left\{d_{\text{stereo}},\ {}^{C}\mathbf{R}\mathbf{\Sigma}_{\text{stereo}}{}^{C}\mathbf{R}^{\top}\right\}, & \text{if } d_{\text{stereo}} \text{ available,} \\ \left\{d_{\text{apparent}},\ {}^{C}\mathbf{R}\mathbf{\Sigma}_{\text{apparent}}{}^{C}\mathbf{R}^{\top}\right\}, & \text{otherwise,} \end{cases} \tag{6}$$

where ${}^{C}\mathbf{R} \in \mathbb{R}^{3\times3}, {}^{C}\mathbf{p} = [{}^{C}p_x, {}^{C}p_y, {}^{C}p_z]^{\top} \in \mathbb{R}^3$ represent the camera's rotation matrix and position, respectively, describing the camera's pose in the world frame. The covariance matrices for each distance measurement type are defined as follows

$$\mathbf{\Sigma}_{\text{UWB}} = \text{diag}\left(\sigma_{xy}^2, \sigma_{xy}^2, \sigma_{z,\text{UWB}}^2\right), \tag{7}$$

$$\mathbf{\Sigma}_{\text{stereo}} = \text{diag}\left(\sigma_{xy}^2, \sigma_{xy}^2, \sigma_{z,\text{stereo}}^2\right), \tag{8}$$

$$\mathbf{\Sigma}_{\text{apparent}} = \text{diag}\left(\sigma_{xy}^2, \sigma_{xy}^2, \sigma_{z,\text{apparent}}^2\right), \tag{9}$$

where σ_{xy} indicates the uncertainty in determining the bounding box's center, and $\sigma_{z,\text{UWB}}$, $\sigma_{z,\text{stereo}}$, $\sigma_{z,\text{apparent}}$ reflect the uncertainties associated with the respective distance estimation methods. These uncertainties are either empirically determined or based on the known characteristics of the sensors used. It is worth noting that we assume the camera's optical axis aligns with z-axis in the camera frame.

C. Formation control

The proposed formation control approach incorporates a leader-follower formation scheme coupled with a receding horizon control. In this scheme, one UAV takes on the role of the formation leader, equipped with onboard sensors and modules responsible for detecting the human worker and recognizing their gestures. The obtained information is then shared with other UAVs within the team, as depicted in Fig. 2. In contrast, follower UAVs, equipped with supplementary cameras, use the information provided by the leader to maintain a predefined formation relative to the worker's position. Concurrently, they capture additional perspectives to enhance safety monitoring. All UAVs within the formation maintain their respective cameras oriented toward the worker. A visual depiction of this scenario is provided in Fig. 3.

The state of the i-th UAV in the formation, denoted as ${}^{i}\mathbf{x} = [{}^{i}\mathbf{p}, {}^{i}\varphi, {}^{i}\xi]^{\top} \in \mathbb{R}^5$, consists of the UAV's position coordinates ${}^{i}\mathbf{p} = [{}^{i}p_x, {}^{i}p_y, {}^{i}p_z]^{\top} \in \mathbb{R}^3$ and the orientation of its camera, represented by the heading ${}^{i}\varphi$ and pitch ${}^{i}\xi$. The label i in the upper left indicates a specific UAV within the team, where $i = L$ refers to the leader UAV, and $i \in \mathbb{N}_{>0}$ pertains to the

[4] https://www.terabee.com/shop/mobile-robotics/terabee-robot-positioning-system

follower UAVs. Similarly, the state of the human worker is denoted as ${}^{H}\mathbf{x} = [{}^{H}\mathbf{p}, {}^{H}\varphi, 0]^{\top}$. We introduce the concept of adaptive parameters for specifying desired observation angles (${}^{i}\beta$ and ${}^{i}\gamma$) and distances (${}^{i}d$), as depicted in Fig. 3, to incorporate dynamic inputs from operators and monitored workers. These parameters evolve based on gestures made by the human worker and requests communicated by the remote human operator. Such flexibility enables mid-flight adaptation of the view on the scene during the continuous tracking of human workers and their interactions with the formation. Adaptation of formation parameters in response to gestures is executed incrementally, enhancing the worker's situational awareness by observing the behavior of the UAVs. The trajectory generation process is designed to accommodate such step changes, resulting in smooth and feasible trajectories.

Given the desired observation angles in the horizontal (${}^{L}\beta$) and vertical (${}^{L}\gamma$) planes and the required distance to the human (${}^{L}d$), the desired state of the leader is given by

$$ {}^{L}\mathbf{x} = [{}^{H}\mathbf{p}^{\top}, \mathbf{0}]^{\top} - \begin{bmatrix} {}^{L}d\cos({}^{H}\varphi - {}^{L}\beta)\cos({}^{L}\gamma) \\ {}^{L}d\sin({}^{H}\varphi - {}^{L}\beta)\cos({}^{L}\gamma) \\ {}^{L}d\sin(-{}^{L}\gamma) \\ {}^{L}\beta - {}^{H}\varphi \\ {}^{L}\gamma \end{bmatrix}. \quad (10) $$

Similarly, the desired state of the follower UAVs, with the required observation distance ${}^{i}d$ and observation angles ${}^{i}\beta$ and ${}^{i}\gamma$ defined with respect to the leader UAV's observation angles, is computed as

$$ {}^{i}\mathbf{x} = [{}^{H}\mathbf{p}^{\top}, \mathbf{0}]^{\top} - \begin{bmatrix} {}^{i}d\cos({}^{L}\varphi - {}^{i}\beta)\cos({}^{i}\gamma - {}^{L}\xi) \\ {}^{i}d\sin({}^{L}\varphi - {}^{i}\beta)\cos({}^{i}\gamma - {}^{L}\xi) \\ {}^{i}d\sin({}^{L}\xi - {}^{i}\gamma) \\ {}^{i}\beta - {}^{L}\varphi \\ {}^{i}\gamma - {}^{L}\xi \end{bmatrix}. \quad (11) $$

Note that ${}^{L}\beta = 0$ represents an observation angle aligned with the heading of the worker ${}^{H}\varphi$, and that ${}^{H}\varphi$ does not necessarily need to match the orientation of the worker's body, but can coincide with the estimated motion direction or be set to a constant value.

The formation controller first applies (10) and (11) to every pose on the prediction horizon using the worker's predicted trajectory and the leader's planned trajectory. This step generates reference trajectories for the UAVs, initially excluding the collision avoidance constraints to alleviate complexity. Subsequently, collision-free paths along the reference trajectories are established for each UAV using a map of the environment. Following this, safe corridors are computed along these paths using a convex decomposition of free space [33]. As a final step, trajectory optimization is executed within these safe corridors to obtain dynamically feasible, collision-free trajectories. To prevent inter-UAV collisions, the projected planned trajectories of the team members, inflated by a safety distance Γ_{dis}, are incorporated as obstacles in the map of the environment for other UAVs. This three-stage trajectory generation process operates onboard each UAV, following a receding horizon strategy

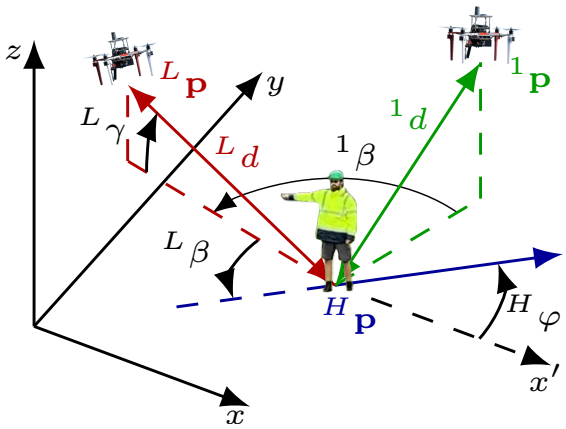

Fig. 3: Illustration of the proposed formation scheme for tracking the human worker, while providing a diverse view of the scene from multiple angles (given by ${}^{i}\beta$ and ${}^{i}\gamma$) and distances (${}^{i}d$).

that enables online response to dynamic changes in the environment and requests for view adaptation. For detailed information on UAV coordination method, refer to [8].

III. Results

The effectiveness of the proposed HSI approach was evaluated through Gazebo simulations and field experiments in a mock-up scenario. The simulations were performed using the MRS software stack [27], on a computer with an i7-10510U processor and 16GB of RAM. Videos of the experiments can be found at `https://mrs.felk.cvut.cz/gestures2024`, with snapshots in Fig. 9.

A. Simulation

The validation simulation scenarios mirror real-world applications for the proposed methodology. In one scenario, a formation of three UAVs is tasked with monitoring a human worker performing maintenance operations at two power transmission towers (see Fig. 4). Throughout the mission, the formation receives 25 requests to adjust the views provided by the UAVs, altering both the observation angles and the distance of individual UAVs from the worker's estimated position. Most of these requests are initiated by an operator monitoring safety compliance, while the remainder are triggered by the human worker's commands, who requests an increase in the UAVs' relative distance to ensure comfort and safety given the wind conditions and proximity to obstacles.

The mission showcases the system's ability to navigate close to obstacles, including maneuvering through narrow gaps formed by electrical power lines while maintaining tracking of the human subject. As illustrated in Fig. 5, the UAVs successfully maintain the required distance from the target and obstacles, adhering to the desired observation angles throughout the mission. The simulation also demonstrates that, beyond providing multiple perspectives of the monitored scene, the UAV formation effectively tracks the human even without explicit implementation of direct visibility constraints. Temporary loss of the worker from the Field of View (FoV) or occlusions occurred only during a circular flight around a transmission tower. Nevertheless,

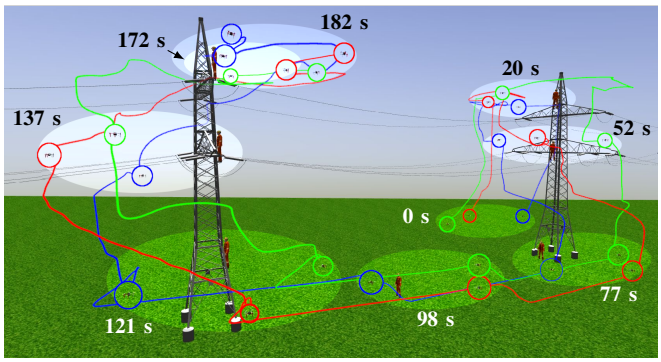

Fig. 4: Simulation illustrating a safety monitoring scenario with three UAVs responding to numerous view adaptation requests. Ellipses highlight the UAV formation at specific time instances, while the trajectories are denoted by colored lines: the leader UAV in red and the follower UAVs in blue and green.

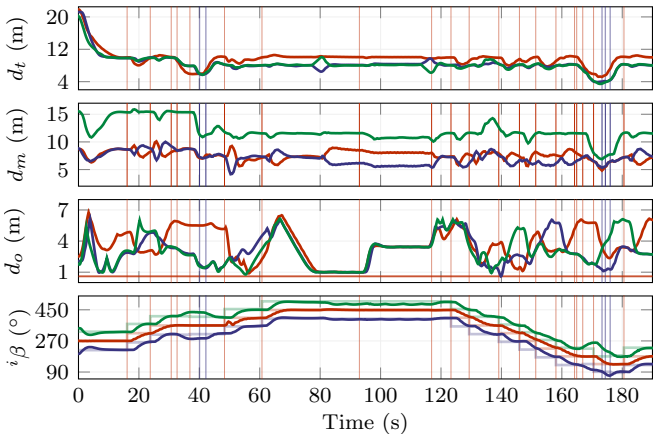

Fig. 5: Simulation timeline depicting the evolution of the key metrics over time: the distance between UAVs and human worker (d_t), the mutual distance between UAVs (d_m), distance to the nearest obstacle (d_o), and the observation angle (${}^i\beta$). These metrics are individually represented for each UAV, with the leader's data shown in red and the followers' in green and blue. In the bottom graph, the opaque lines indicate the reference observation angles. Vertical lines mark the instances when commands from the operator (in red) and the human worker (in blue) were received.

these brief information gaps from the camera sensors were mitigated by the proposed human pose estimation module, highlighting the system's robustness and adaptability.

B. Experiments

The integration of introduced modules into a single system running onboard UAVs is demonstrated through field experiments involving three UAVs collaborating with a human worker wearing a reflective safety vest. The worker's gestures were mapped to changes in formation parameters for the purpose of real-world responsiveness demonstration. The mapping was as follows: crossing arms (gesture ID = 1) decreased ${}^L\beta$, extending an arm to the side (ID = 2) increased ${}^L\beta$, palms put together (ID = 3) decreased ${}^L\gamma$, and raising an arm upwards (ID = 4) increased ${}^L\gamma$. The increments and decrements of ${}^L\beta$ and ${}^L\gamma$ were set to 30° and 5°, respectively. The heading of the human worker, ${}^H\varphi$, was assumed constant, and gestures were filtered using the twenty most recent measurements. Additional parameters used for the experiments are listed in Table II.

TABLE II: Values of parameters applied in the experiments. Part of the parameters is influenced by requirements of safety monitoring procedures of the industrial partners in the AERIAL-CORE project.

Parameter	Symbol	Value	Parameter	Symbol	Value
CNN sliding window	N	9 [–]	^{L}UAV obs. heading	${}^L\beta$	90°
Data filtering thr.	t_c	20 s	^{L}UAV obs. pitch	${}^L\gamma$	11°
Ratio threshold	Π_d	0.8 [–]	^{L}UAV des. distance	${}^Ld$	10.00 m
Command threshold	t_d	5 s	^{1}UAV obs. heading	${}^1\beta$	60°
Pos. process noise	$\sigma_{p_\bullet}$	0.1 m	^{1}UAV obs. pitch	${}^1\gamma$	0°
Vel. process noise	$\sigma_{v_\bullet}$	0.1 m s^{-1}	^{1}UAV des. distance	${}^1d$	8.00 m
Direction meas. noise	σ_{xy}	0.05 m	^{2}UAV obs. heading	${}^2\beta$	−60°
UWB meas. noise	$\sigma_{z,\text{UWB}}$	0.1 m	^{2}UAV obs. pitch	${}^2\gamma$	0°
Stereo meas. noise	$\sigma_{z,\text{stereo}}$	0.3 m	^{2}UAV des. distance	${}^2d$	8.00 m
Apparent size meas. noise	$\sigma_{z,\text{apparent}}$	0.6 m	Mutual distance thr.	Γ_{dis}	2.50 m

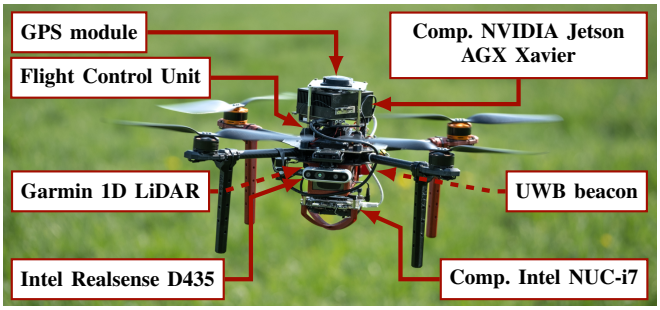

Fig. 6: The aerial platform used as primary UAV for the experiments.

Two types of multi-rotor UAVs were used in the experimental validation. The primary UAV utilizes a Tarot 650 frame and is equipped with a Pixhawk Flight Control Unit (FCU) with sensors for UAV state estimation, gesture recognition, and human detection. Refer to Fig. 6 for detailed visuals. Onboard computation is facilitated by an NVIDIA Jetson AGX Xavier computer, which manages the human detection and gesture recognition pipeline, while an Intel NUC-i7 handles the core functions of state estimation, control, and planning (see Fig. 2). The computers are interconnected via an Ethernet interface, ensuring reliable data transfer. While running the entire pipeline on a single AGX Xavier computer is feasible, leveraging additional computational resources allows for faster image processing and separation of the computationally intensive image processing pipeline from the safety-critical modules essential for autonomous UAV flight. The secondary UAVs are constructed using F450 platforms, with a payload limited to a single onboard computer Intel NUC-i7, Pixhawk FCU, and the necessary sensors for state estimation and scene capture. A detailed description of the hardware platforms is provided in [34], [35].

The final evaluation of the system followed a series of experiments involving varying numbers of UAVs and diverse environments, which helped to fine-tune the performance of individual modules and their interconnections. The presented evaluation of the system is based on three autonomous flights conducted under a consistent setup and configuration of the modules. In these flights, the system achieved a success rate of 87% in propagating human gestures to scene view adaptation. In this metric, an invoked shape adaptation is considered successful if executed while the human performs the corre-

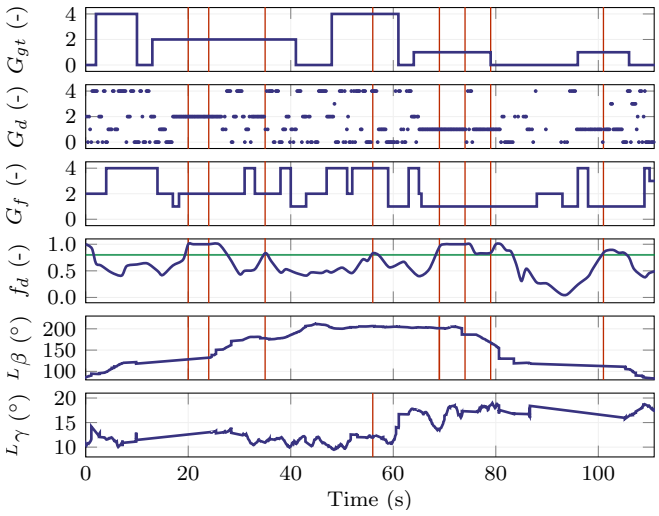

Fig. 7: A timeline illustrating the progression from human gestures to the adaptation of the relative view provided by the UAVs. The graphs depict the IDs of the gestures performed by the human worker (G_{gt}), gestures detected by the gesture recognition module (G_d), the dominant gesture (G_f), and its corresponding relative frequency (f_d). Additionally, the graphs include the observation angles of the leading UAV (${}^L\beta$, ${}^L\gamma$). Each gesture is represented by its associated ID, as detailed in Section III. The gesture ID = 0 corresponds to the detection of a human not performing any gesture. The green line represents the threshold Π_d on f_d, while the vertical red lines denote instances associated with confirmed requests for scene view adaptation.

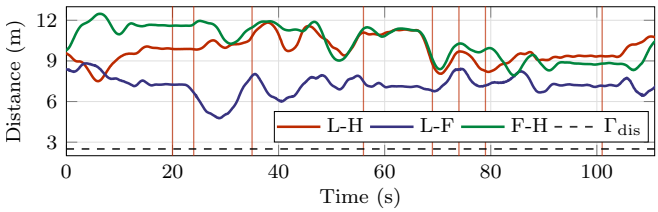

Fig. 8: Mutual distances between the leader UAV and the human (L-H), between the leader UAV and one of the followers (L-F), and between the follower UAV and the human (F-H) depicted during one of the real-world experiments. The red vertical lines indicate times when a request for scene view adaptation was sent. The follower's deviation from the required distance is caused mainly by the propagation of the imprecision in human pose estimation.

sponding gesture. A timeline illustrating the propagation of a human's gesture through the gesture recognition and filtering pipeline, up to the adaptation of observation angles during one of the experiments, is depicted in Fig. 7. The average time from the initiation of a gesture to the onset of shape adaptation during the final experiments was 7 s. This response time is influenced by a conservative parameter setup, which is necessary to prevent undesired view adaptation due to incorrect gesture classification or a temporary worker's pose resembling one of the predefined gestures. Throughout the experiments, the UAV team effectively maintained the safety distance Γ_{dis} between both the human and among the UAVs while keeping their cameras oriented toward the worker. This ensured safety and showcased the capability of the proposed approach in real-world scenarios, as depicted in Fig. 8 and Fig. 9.

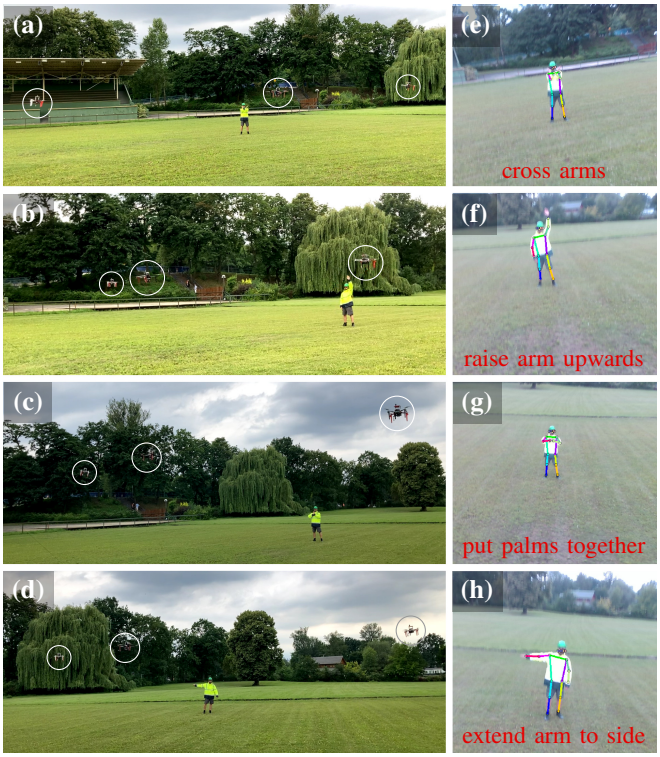

Fig. 9: Sequence of snapshots showing a team of UAVs following a human worker (a)-(d) and adapting the relative view based on the detected gestures (e)-(h). The experiment presents a full 3D deployment, which requires adapting the observation angles in both the horizontal and vertical directions.

IV. Discussion

The conducted experiments underscore the potential of employing hand gestures for the intuitive control and coordination of multi-robot aerial systems. This feature proves especially advantageous in scenarios such as safety monitoring and assisting human workers in challenging environments, as it does not impose an extra workload on the workers nor necessitate additional equipment for conventional wireless communication.

However, gesture-based control presents specific challenges distinct from other modes of interaction. Firstly, the admissible observation angles and distance range are limited by the gesture recognition module's performance and the safety requirements of the workers. The recognition range of the worker in the image is constrained, which must be considered in the scene view adaptation process. Similarly, adhering to safety regulations involves maintaining a minimum distance from the worker. Our approach addresses this challenge by imposing stringent limits on the parameters ${}^L\gamma$, ${}^i d$ and Γ_{dis}. In future work, we intend to enhance the system's robustness by implementing a worker detection pipeline on multiple UAVs coupled with distributed estimation of worker's position. Such an approach not only provides multiple perspectives but also prevents the loss of the tracked worker due to occlusion or obstacle avoidance maneuvers.

The second important aspect of employing gestures to

interact with teams of aerial robots is particularly relevant in contexts where workers involved in maintenance tasks may unintentionally assume positions that resemble predefined gestures. This scenario is amplified by the safety challenges inherent in such environments, characterized by constrained mobility and the necessity to maintain uncomfortable postures. Given the potential for misinterpretation, it becomes crucial to configure the gesture processing pipeline with care. This cautious approach is necessary to avoid the propagation of false positive detections to mission-related commands.

Lastly, our experimental campaigns have revealed that providing clear feedback from the formation to the human executing gestures is one of the most significant and not immediately evident aspects of HSI via gestures. To avoid the need for additional equipment for visual feedback, we have structured the behavior of UAVs such that the human can gauge the acceptance of their command based on the observable actions of the UAVs. In this regard, making incremental alterations in the formation parameters and avoiding continuous scene view adaptation commands have proven advantageous.

V. Conclusion

In this paper, we introduced a novel approach for contactless Human-Swarm Interaction using hand gestures to control a team of UAVs applicable in safety monitoring scenarios. The proposed approach enables safe and efficient interaction between remote human operators, human workers and autonomous aerial systems, offering benefits in real-world scenarios. Integrating hand gestures as a control modality allows human workers to command and adjust various formation parameters, such as relative distance to the worker, request immediate assistance, and initiate other mission-related commands. The proposed approach directly incorporates robust algorithms for human worker detection and gesture recognition, ensuring an accurate and prompt response. Simulations and field experiments validated the effectiveness of the approach, demonstrating successful navigation in complex environments while providing varying required perspectives controlled by both remote commands and based on the detected hand gestures.

References

[1] S. S. Mansouri *et al.*, "Cooperative Coverage Path Planning for Visual Inspection," *Contr. Eng. Pr.*, vol. 74, pp. 118–131, 2018.

[2] G. Loianno *et al.*, "Autonomous Flight and Cooperative Control for Reconstruction Using Aerial Robots Powered by Smartphones," *Int. J. Rob. Res.*, vol. 37, no. 11, pp. 1341–1358, 2018.

[3] A. Caballero *et al.*, "A Signal Temporal Logic Motion Planner for Bird Diverter Installation Tasks With Multi-Robot Aerial Systems," *IEEE Acc.*, vol. 11, pp. 81 361–81 377, 2023.

[4] S. Haddadin *et al.*, "Towards the Robotic Co-Worker," in *Robotics Research*, 2011, pp. 261–282.

[5] M. Tognon *et al.*, "Physical Human-Robot Interaction With a Tethered Aerial Vehicle: Application to a Force-Based Human Guiding Problem," *IEEE Trans. Rob.*, vol. 37, no. 3, pp. 723–734, 2021.

[6] F. Benzi *et al.*, "Adaptive Tank-based Control for Aerial Physical Interaction with Uncertain Dynamic Environments Using Energy-Task Estimation," *Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9129–9136, 2022.

[7] A. Ollero *et al.*, "AERIAL-CORE: AI-Powered Aerial Robots for Inspection and Maintenance of Electrical Power Infrastructures," *arXiv:2401.02343*, 2024.

[8] V. Kratky *et al.*, "Autonomous Aerial Filming with Distributed Lighting by a Team of Unmanned Aerial Vehicles," *Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7580–7587, 2021.

[9] A. Afifi *et al.*, "Toward Physical Human-Robot Interaction Control with Aerial Manipulators: Compliance, Redundancy Resolution, and Input Limits," in *Int. Conf. Rob. Aut.*, 2022, pp. 4855–4861.

[10] A. Ajoudani *et al.*, "Progress and Prospects of the Human-Robot Collaboration," *Aut. Rob.*, vol. 42, no. 5, pp. 957–975, 2018.

[11] A. Kolling *et al.*, "Human Interaction With Robot Swarms: A Survey," *IEEE Trans. on Hum.-Mach. Syst.*, vol. 46, no. 1, pp. 9–26, 2016.

[12] A. Dahiya *et al.*, "A Survey of Multi-Agent Human–Robot Interaction Systems," *Rob. and Aut. Sys.*, vol. 161, no. 104335, pp. 1–18, 2023.

[13] A. Couture-Beil *et al.*, "Selecting and Commanding Individual Robots in a Multi-Robot System," in *Can. Conf. on Comp. and Rob. Vis.*, 2010, pp. 159–166.

[14] J. Nagi *et al.*, "Human-Swarm Interaction Using Spatial Gestures," in *Int. Conf. Int. Rob. Syst.*, 2014, pp. 3834–3841.

[15] G. Abbate *et al.*, "Pointing at Moving Robots: Detecting Events from Wrist IMU Data," in *Int. Conf. Rob. Aut.*, 2021, pp. 3604–3611.

[16] S. S. Abdi *et al.*, "Safe Operations of an Aerial Swarm via a Cobot Human Swarm Interface," in *Int. Conf. Rob. Aut.*, 2023, pp. 1701–1707.

[17] M. Macchini *et al.*, "Personalized Human-Swarm Interaction Through Hand Motion," *Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8341–8348, 2021.

[18] V. M. Monajjemi *et al.*, "HRI in the Sky: Creating and Commanding Teams of UAVs with a Vision-mediated Gestural Interface," in *Int. Conf. Int. Rob. Syst.*, 2013, pp. 617–623.

[19] L. Zhang *et al.*, "Optimal robot selection by gaze direction in multi-human multi-robot interaction," in *Int. Conf. Int. Rob. Syst.*, 2016, pp. 5077–5083.

[20] M. Jacquet *et al.*, "Motor-Level N-MPC for Cooperative Active Perception With Multiple Heterogeneous UAVs," *Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2063–2070, 2022.

[21] A. Alcantara *et al.*, "Optimal Trajectory Planning for Cinematography With Multiple Unmanned Aerial Vehicles," *Rob. and Aut. Sys.*, vol. 140, no. 103778, 2021.

[22] R. Bonatti *et al.*, "Autonomous Aerial Cinematography in Unstructured Environments with Learned Artistic Decision-making," *J. of Field Rob.*, vol. 37, no. 4, pp. 606–641, 2020.

[23] C. Symeonidis *et al.*, "Neural Attention-Driven Non-Maximum Suppression for Person Detection," *IEEE Trans. on Im. Pro.*, vol. 32, pp. 2454–2467, 2023.

[24] I. Karakostas *et al.*, "Occlusion Detection and Drift-avoidance Framework for 2D Visual Object Tracking," *Sig. Proc.: Im. Com.*, vol. 90, p. 116011, 2021.

[25] C. Papaioannidis *et al.*, "Fast CNN-based Single-Person 2D Human Pose Estimation for Autonomous Systems," *IEEE Trans. on Cir. and Sys. for Video Tech.*, vol. 33, no. 3, pp. 1262–1275, 2023.

[26] ——, "Learning Fast and Robust Gesture Recognition," in *Eur. Conf. on Sig. Pro.*, 2021, pp. 761–765.

[27] T. Baca *et al.*, "The MRS UAV System: Pushing the Frontiers of Reproducible Research, Real-world Deployment, and Education with Autonomous Unmanned Aerial Vehicles," *J. of Int. & Rob. Sys.*, vol. 102, no. 26, pp. 1–28, 2021.

[28] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Eur. Conf. on Comp. Vis.*, 2016, pp. 21–37.

[29] F. Patrona *et al.*, "Self-Supervised Convolutional Neural Networks for Fast Gesture Recognition in Human-Robot Interaction," in *Int. Conf. on Inf. and Aut. for Sust.*, 2021, pp. 88–93.

[30] ——, "An Overview of Hand Gesture Languages for Autonomous UAV Handling," in *Aer. Rob. Sys. Ph. Int. with the Env.*, 2021, pp. 1–7.

[31] A. Vaswani *et al.*, "Attention is all you need," *arXiv:1706.03762*, 2017.

[32] M. Vrba *et al.*, "Marker-Less Micro Aerial Vehicle Detection and Localization Using Convolutional Neural Networks," *Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2459–2466, 2020.

[33] S. Liu *et al.*, "Planning Dynamically Feasible Trajectories for Quadrotors Using Safe Flight Corridors in 3-D Complex Environments," *Robot. Autom. Lett.*, vol. 2, no. 3, pp. 1688–1695, 2017.

[34] D. Hert *et al.*, "MRS Drone: A Modular Platform for Real-World Deployment of Aerial Multi-Robot Systems," *J. of Int. & Rob. Sys.*, vol. 108, no. 64, pp. 1–34, 2023.

[35] D. Hert *et al.*, "MRS Modular UAV Hardware Platforms for Supporting Research in Real-World Outdoor and Indoor Environments," in *Int. Conf. on Unm. Air. Sys.*, 2022, pp. 1264–1273.