TagTeam: Towards Wearable-Assisted, Implicit Guidance for Human–Drone Teams

Kasthuri Jayarajah† , Aryya Gangopadhyay† , Nicholas Waytowich‡
† University of Maryland, Baltimore County
‡US Army Research Lab
USA

ABSTRACT

The availability of sensor-rich smart wearables and tiny, yet capable, unmanned vehicles such as nano quadcopters, opens up opportunities for a novel class of *highly interactive*, *attention-shared* human–machine teams. Reliable, lightweight, yet passive exchange of intent, data and inferences within such human–machine teams make them suitable for scenarios such as search-and-rescue with significantly improved performance in terms of speed, accuracy and semantic awareness. In this paper, we articulate a vision for such human–drone teams and key technical capabilities such teams must encompass. We present *TagTeam*, an early prototype of such a team and share promising demonstration of a key capability (i.e., motion awareness).

ACM Reference format:
Kasthuri Jayarajah† , Aryya Gangopadhyay† , Nicholas Waytowich‡ . 2022. *TagTeam*: Towards Wearable-Assisted, Implicit Guidance for Human–Drone Teams. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 6 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The availability of sensor-rich smart wearables and tiny, yet capable, unmanned vehicles such as nano quadcopters[1], opens up opportunities for a novel class of *highly interactive*, *attention-shared* human–machine teams. Reliable, lightweight, yet passive exchange of intent, data and inferences within such human–machine teams make them suitable for scenarios such as search-and-rescue with significantly improved performance in terms of speed, accuracy and semantic awareness.

Our proposed paradigm of human–drone teams are motivated by two salient trends:

- ***Sensor-Rich, Pervasive Wearable Devices:*** Whilst smart watches, bands and rings have been widely studied for fine-grained, gesture-based control of smart environments and machines, more recent technologies are capable of continuously capturing more than just inertial motion. For instance, miniaturized sensors such as earables (i.e., wearables worn on the ear such as the Emotiv MN8[2]) and smart glasses (e.g., AttentiveU[3]) are equipped with a range of inertial and physiological sensors that can measure brain activity, eye movements, etc. that can passively measure the wearer's cognitive processes. More recently, devices such as the Microsoft HoloLens 2[4] embedded with a variety of vision, depth and time-of-arrival sensors coupled with IMU sensors, head and gaze tracking, open up interesting possibilities for capturing both point-of-view visual information and the individual's physiological and neurological states. Together with their low power requirements and connectivity capabilities, such devices can provide cues about individual's intents and states, at real-time, to their robotic teammates for implicit coordination in various scenarios including battlefields and Industry 4.0 settings.
- ***Highly Mobile, SWAP-Constrained Unmanned Vehicles:*** Robotic platforms that are size, weight, and power constrained are attractive for human-machine teams that operate side by side, occupying the same physical space. While high levels of mobility can pose serious safety concerns for the human agents, the physical configurations of SWAP-constrained platforms make them safer options without compromising on the sensing capabilities they can offer. However, such platforms can be extremely restrictive in terms of performing fully autonomous navigation, sensing and onboard computation for sense-making. We posit that such platforms can leverage guidance from their human partners for intelligently adapting between *autonomous* and *assisted* operations for longer operation windows without compromising on the sensing efficacy.

1.1 Motivating Scenarios

"Passive guidance" from human(s) in the team can help direct machines for maximizing the effectiveness of human-machine combined sensing objectives. Communication, the process of information exchange, between man and machine is key for successful team performance [28]. While advances in natural language processing or exchange of visuals aid in more direct, explicit communication between teammates,

[1] https://www.bitcraze.io/

[2] https://www.emotiv.com/setup/mn8/

[3] https://www.media.mit.edu/projects/attentivu/overview/

[4] https://docs.microsoft.com/en-us/hololens/hololens2-hardware

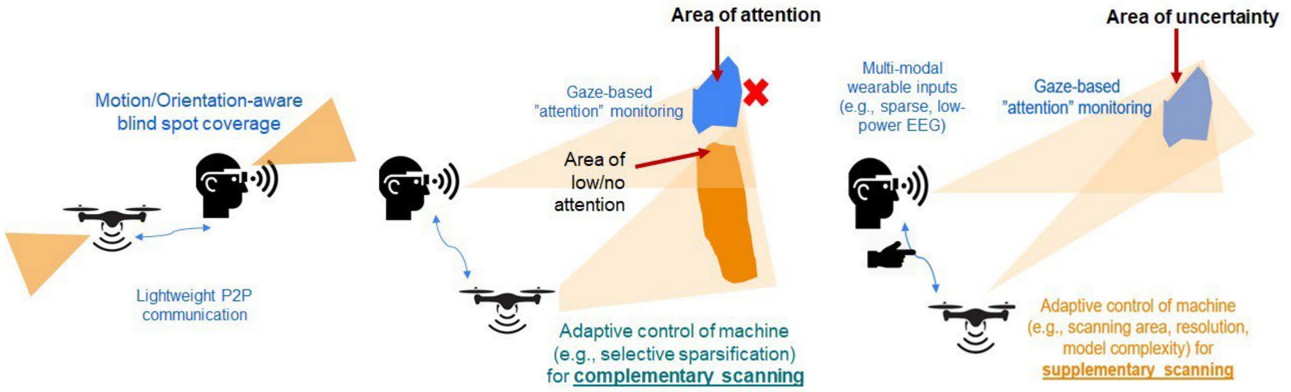
Motion/Orientation-aware blind spot coverage
Lightweight P2P communication
Gaze-based "attention" monitoring
Area of attention
Area of low/no attention
Adaptive control of machine (e.g., selective sparsification) for complementary scanning
Multi-modal wearable inputs (e.g., sparse, low-power EEG)
Gaze-based "attention" monitoring
Area of uncertainty
Adaptive control of machine (e.g., scanning area, resolution, model complexity) for supplementary scanning

Figure 1: Illustrative *TagTeam* Scenarios: Gaze-Assisted Cooperative Visual Scanning in Indoor Environments. See https://youtu.be/KYeo2Aichgs for a video demonstration of the blind spot coverage scenario.

"implicit coordination" where machines are able to synchronize with their human-teammates without explicit intervention has its advantages. Previous studies [18] show that implicit coordination is helpful under high workload situations due to a reduction in communication overheads and the resulting distractions. Here, we describe two scenarios where we envision tightly-coupled and responsive drones that adapt to human intent based on implicit guidance to be highly effective. We illustrate this in Figure 1.

Attention-Shared, coordinated visual scanning for reconnaissance and search-and-rescue missions: An exemplar scenario is where a dismount is teamed up with a robotic teammate, with the robotic teammate equipped with a variety of sensors such as RGB cameras and LIDAR. Coordinated scanning can include two goals: (1) complementary - achieve wider coverage where the machine is able to scan regions where the human is not paying attention to, and as a team, achieve faster and efficient scanning, and (2) collaborative supplementary – where the machine provides enhanced resolution scanning and inference when a human requires more accurate visibility and augmentation from the machine. In both cases, machines require to know where the human is paying attention to, and not just what is within the "visible" range. A collection of wearable sensors can help intelligently infer the type of assistance the human teammate requires (e.g., whether complementary or supplementary) as well as in passively guiding the drone to areas that need augmented attention and/or at varied perception configurations (e.g., resolution, coverage, etc.).

Non-verbal, interactive communication for continuous learning While natural language is a more direct interface for communication with machines, non-language behaviours (voice quality, body language, etc.) [5] and motor correlates of speech and verbal communication (e.g., gaze, facial expressions, gestures [19] play a crucial role in effective communication of the human expressor. The ability to sense such cues can be beneficial in many battlefield scenarios including (1) machine learning and adaptation with passive human reinforcement (e.g., via affirmative or negative thinking inferred from brain activity [25]) and (2) to interactively resolve comprehension ambiguities of human-to-robot instructions (e.g., gestures to zoom in/out for a vision sensing task, corrective or altered behaviours based on cues such as a frown or shaking of the head).

Key Contributions: Through this initial work, we articulate a paradigm of highly interactive, attention-shared human-drone teams and identify key technical capabilities such teams require to address. We also share details of an initial prototype we built and share early results from enabling accurate motion transfer from human to drone.

2 DESIGN GOALS

To support the *TagTeam* scenarios that we envision, we enumerate the following key design goals that human–drone teaming systems should achieve.

2.1 Wearable-based attentive state estimation:

Visual attention: In this work, we will develop techniques to gauge human visual attention using a combination of wearable technologies that allow for accurate tracking of eye movements (e.g., using noisy EOG signals) in the presence of motion artifacts (determined using on-body inertial and EEG signals) and investigate techniques for continuous, light-weight exchange of attention information. The machines then adapt their attentional focus and/or resolution, on-the-fly, to synchronize with their human-teammates' intent. Evaluations of such systems will require both recreations of dynamic environments in an augmented-reality based experimental setups and real-world studies to study the trade-offs between accuracy on vision tasks, energy efficiency and latency, baselining against attention-agnostic models for both the complementary and collaborative goals. We believe that these models of attention-responsive adaptation of machine intelligence will not only improve the overall inference and situational intelligence accuracy, but also provide a practical way to reduce energy and computational requirements, thereby enabling longer operational lifetimes and more ergonomic machine form factors.

Non-verbal intent: A key technical capability that this proposed work builds is how such nonverbal intent can be reliably measured with a combination of wearable sensors including EEG sensors, on-body inertial sensors, gaze trackers, etc. While the effectiveness of physiological sensing has been demonstrated with high-fidelity sensors, in lab and controlled settings, for tasks such as inferring emotion, the ability to infer nonverbal cues using energy-efficient, but sparse, physiological signals is relatively under-explored (e.g., a 2-channel Emotiv MN8 earable as opposed to a 64-channel Biosemi ActiveTwo EEG sensor). More recently, ear-worn inertial sensors have indeed been shown to be effective in detecting activities such as head and neck movements [10] and more finer-grained motion such as tapping and sliding of the teeth [23]. Furthermore, recent work [29] has demonstrated that pointing-gesture based input in combination with visual and verbal inputs can improve accuracy of object picking tasks. To this end, the work will explore sensor fusion of wearable modalities (e.g., physiological, micro-expressions, lip/jaw movements based on inertial measurements, gaze tracking with smart glasses, etc.) for accurate nonverbal cue sensing and its application in the two example scenarios. One such possible use case is an urban battlefield where the human dismount issues commands to a robotic assistant, for example, to investigate some roadside objects for possible threats. While ongoing work explores such multi-modal instruction comprehension for civilian environments (e.g., factory floors), kinetic military environments are likely to be characterized by higher levels of stress, distraction and time constraints. This in turn will affect the ways in which humans communicate instructions and raise the importance of factoring in non-verbal input (e.g., stress or fear levels) in defining the performance requirement (not just accuracy but factors such as time sensitivity) for such comprehension. We will thus engineer a suite of multimodal features for lightweight detection of higher order non-verbal cues, develop an adaptive system that orchestrates the triggering of multiple modalities based on energy, accuracy and task-specific performance (e.g., how fast can the human and machine converge on instruction comprehension?) trade offs, baselining against purely language-based comprehension.

2.2 Global understanding of location and motion:

For highly mobile situations such as scouting, search-and-rescue, the human and agent require highly accurate spatial and motion awareness. This a key requirement, especially in previously unseen environments such as those envisioned. In our scenarios, we require that the drone or agent be capable of mimicking, or closely follow, the human for situations such as providing blindspot coverage. In situations where the drone detects that the human requires complementary or supplementary scanning, through the combination of various wearable-based attentive states, the drone navigates independent of the human for completing such tasks, and *boomerangs* back to its human accomplice upon completion of the task. Such coordination requires a common grounding of the spatial coordinate systems of the two devices (human-worn wearable and the drone) and continuous tracking.

2.3 Constraint-aware orchestration:

Cooperating human-machine teams can be high efficient in terms of scanning target areas in shorter windows, as opposed to a human-alone or drone-alone baseline. Whilst, small unmanned vehicles, especially aerial vehicles, are extremely restrictive in terms of their size, weight, and hence power and memory, wearable devices such as the Hololens 2 are much more resilient (lasts up to 2-3 hours with a single charge and active rendering throughout the duration).

3 INITIAL PROTOTYPE

Towards realizing our vision for highly interactive human–drone teams, we describe our efforts in prototyping an early version of *TagTeam* that accomplishes real-time motion awareness for the default scenario of *blindspot detection*. We present a simplified architecture in Figure 2. The system consists of the following components (see Figure 3).

A tiny drone: The Crazyflie 2.1 [7], a nano class quadcopter, is an open source flying development platform. It is an example of a SWAP-constrained edge device similar to what we envision in our scenarios, that fits inside the palm of the hand and weighs only a nominal 27 grams. The base design of the quadcopter consists of a STM32F405 main application MCU (microcontroller unit) and communication enabled via a nRF51822 radio and power management MCU. With the long range open USB radio dongle based on the nRF24LU1+ from Nordic Semiconductor [5] (supporting 2.4GHz ISM band and Bluetooth Low Energy), the drone can be controlled from distances as far as 1 km, under ideal conditions. The capabilities of the Crazyflie 2.1 can be extended using a range of off-the-shelf or custom-built expansion decks. In this work,

[5] https://www.bitcraze.io/products/crazyradio-pa/

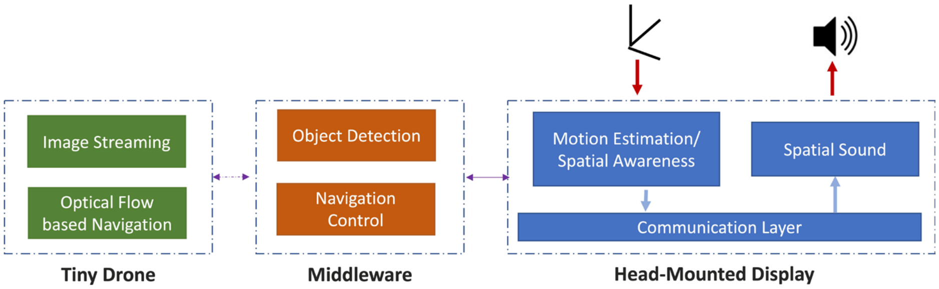

Figure 2: *TagTeam* Components.

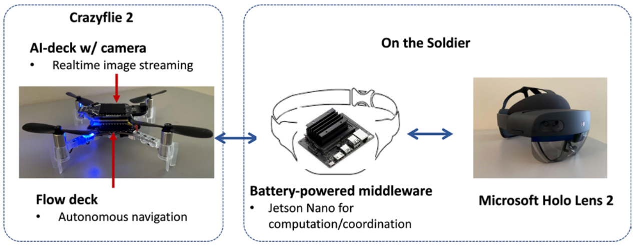

Figure 3: *TagTeam* Implementation.

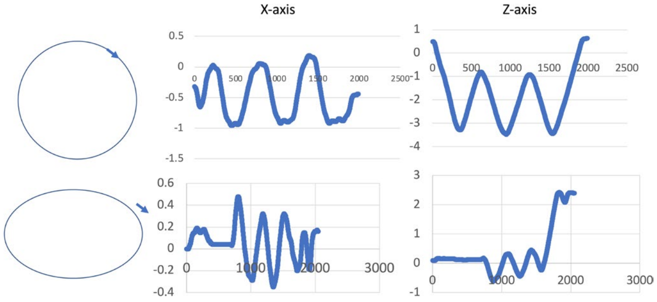

Figure 4: Coordinate variations for exemplar motions: circular (top) and oval (bottom)

we use the AI deck [6] for capturing and wirelessly streaming images and the flow deck V2 [7] for stable flight.

Head-mounted Wearable Display (Smart helmet): We use the Microsoft HoloLens 2, a sensor-rich smart wearable, in our implementation. Through its four grayscale head-tracking cameras, the depth camera and a combination of IMU sensors, the HoloLens device has the capability to maintain highly accurate spatial and motion awareness of the wearer. In the current prototype, we use the Camera.main.transform property of the main camera to localize the user; this property initializes to (0, 0, 0) at the real-world position when the App is launched. In Figure 4, we plot the variation along the x and z directions for two exemplar, controlled motions, circular and oval, performed by a human subject.

Middleware: The Nvidia Jetson Nano[8] acts as the middleware between the drone and the wearable on the soldier. The Nano is a mobile device consisting of a 128-core Maxwell GPU and a Quad-core ARM A57 CPU (clock speed of 1.43 GHz) and 4 GB system memory. In the current implementation, the middleware acts as a broker between the two devices; (a) it ascertains positional information from the wearable device over the MQTT [12] protocol, by subscribing to a certain topic, (b) transforms locational information from the HoloLen's coordinate system to the Crazyflie 2.1's coordinate system, and (c) performs object detection (using the SSD MobileNet V2 detector [26]) on the stream of images transferred by the drone to provide real-time indication of objects in the human's blindspots. Such detections are published as MQTT back to the HoloLens which outputs the information as *Spatial Sound* [9], dependent on the distance and angle of the object relative to the human. During the transformation stage (Step (b)), we take a dead-reckoning based approach [27] where we assume that the initial positions of the HoloLens 2 (or the human) and the drone are known, and we estimate the change in movement needed by the drone, from the change in movement sensed by the HoloLens, assessed periodically, and task the drone to *move* to the new position with a certain velocity (calculated as the distance moved divided the time period between updates). Through experimentation, we found the appropriate coordinate transformation between the two devices to be: $X_{hololens} \rightarrow Z_{crazyflie}$ and $Z_{hololens} \rightarrow -X_{crazyflie}$ (note that in our implementation, the Crazyflie drone faces backwards to provide situation awareness in the human's blindspot regions).

4 PRELIMINARY RESULTS

Figure 5 shows the setup (a 4 ft × 6 ft area) we use in establishing the feasibility of tracking (the human's) movements

[6] https://www.bitcraze.io/products/ai-deck/
[7] https://www.bitcraze.io/products/flow-deck-v2/
[8] https://developer.nvidia.com/embedded/jetson-nano-developer-kit
[9] https://docs.microsoft.com/en-us/windows/mixed-reality/design/spatial-sound

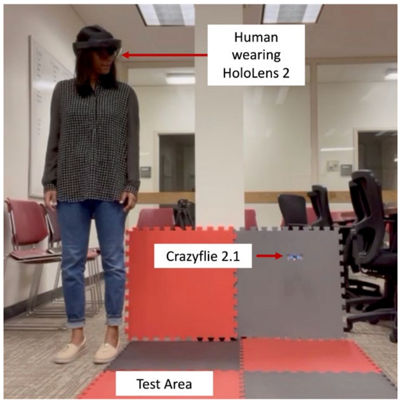

Figure 5: Test setup for location tracking. See https://youtu.be/INQULM5csMU for a video demonstration.

using the dead-reckoning based approach. As seen in the video demonstration, the drone is able to accurately mimic the movements of the human.

To quantify this, we use the vatic.js tool [10] for annotating the bounding box representing both the human's head as well as the drone. In total, the video consists of 503 frames. We use Dynamic Time Warping to capture the similarity in the trajectories of the head and the drone, and observe a high level of synchronization (≈ 0.92 where 1 is the highest).

5 DISCUSSION AND OPEN PROBLEMS

In our current implementation we tackle a fundamental and necessary building block towards building *TagTeam* (i.e., motion awareness between human and drone). As we describe earlier in Section 2, the paradigm of attention-sharing, human-drone teams require several other technical capabilities such as inferring human intent and attention automatically, implicit coordination between the human and drone, and orchestration of the collective resources (e.g., onboard computation on the devices, network bottlenecks, etc.) for effective cooperation and communication. Beyond these immediate requirements, we enumerate a number of open problems for advancing research in this direction.

Extending to multi-human, multi-drone *teams*: While our current scenarios consider a pair of agents, human and otherwise, many of these generalize to multi-human, multi-drone teams – for instance, a SWAT team involved in a high-risk search cooperating with a swarm of drones. Such extensions impose additional challenges such as multi-source information aggregation and dissemination under potentially conflicting sources, data prioritization for processing and exchange under scheduling constraints, and physical-formation aware coordination.

Aspects of trust and psychology: While the form factors of sensing and processing platforms have evolved for *TagTeam*-like technical capabilities to be possible. Whilst works such as those of Hancock et al. [8] study aspects of trust and psychology for human-robot interaction scenarios in general, the human factors related to close-contact human-drone teams remain largely under studied.

6 RELATED WORK

Human–Drone Interaction (HDI): Since recently, researchers have started looking at the technical capabilities HDI requires and novel use cases that it enables [4, 22]. In early works, Christ et al. [4] study the impact of the level of autonomy of the drone on user experience. PFeiffer et al. [22] demonstrate that understanding where humans pay visual attention can aid in improving the navigation capabilities of drones through imitation learning. Several works have specifically focused on exploring *modalities* for interaction such as gesture-based control (e.g., using smart gloves [21], [3]), voice [6, 16, 20], gaze-based teleoperation [9, 30], etc. While these works explore the *explicit control* of the drone through these modalities, in our work, we emphasize the need for *implicit* guidance where the drone automatically infers the intent of the human – for instance, using gaze-based features such as saccades and dwell time to infer areas of uncertainty where the drone should perform a secondary scan to maintain higher accuracy of the scan. Our work is the first, to the best of our knowledge, to articulate a paradigm of humans and drones working in close proxemics to perform cooperative tasks such as indoor scouting and search-and-rescue, whilst addressing challenges in terms of enabling real-time behavior adaptation of the drone leveraging on multi-modal, yet implicit interactions between the drone and its human partner.

Collaborative sense-making at the edge: There have been a number of efforts in enabling lightweight collaboration between machines, especially for compute-intensive tasks such as machine perception using deep neural networks (DNNs). Several recent works have explored optimization techniques for networked sensors to achieve efficient querying [11, 15]. Recent works [1, 13, 17, 24] have also explored the idea of selective activation of nodes in a group of collaborating sensors – e.g., Qiu et al. [24] describe a vehicle tracking scenario where mobile nodes in a hybrid (mobile/infrastructure) camera network are activated selectively, only to resolve ambiguities. Jain et al. [13] provide preliminary examples of the possibility of using inputs from peer, overlapping cameras to utilize such spatiotemporal correlations to optimize the video analytics pipeline. The idea of collaboration among AIoT devices at the edge, and its attendant challenges, has also been mooted more generally recently [1, 2]. Most recently, ComAI [14] demonstrates

[10] https://stefanopini.github.io/vatic.js

concrete mechanisms for low-overhead collaboration for perception. To the best of knowledge, this work is among the first to prescribe the need for lightweight collaboration for highly effective human–machine mixed teams.

ACKNOWLEDGEMENT

We acknowledge the support of the U.S. Army Grant No. W911NF21- 20076.

REFERENCES

[1] T. Abdelzaher, S. Yao, and et al. Eugene: Towards deep intelligence as a service. In *Proc. of IEEE ICDCS'19*, 2019.

[2] H. Barahouei P and T. Nadeem. Collaborative intelligent cross-camera video analytics at edge: Opportunities and challenges. In *Proc. of AIChallengeIoT'19*, 2019.

[3] J. R. Cauchard, J. L. E, K. Y. Zhai, and J. A. Landay. Drone & me: an exploration into natural human-drone interaction. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 361–365, 2015.

[4] P. F. Christ, F. Lachner, A. Hösl, B. Menze, K. Diepold, and A. Butz. Human-drone-interaction: A case study to investigate the relation between autonomy and user experience. In *European conference on computer vision*, pages 238–253. Springer, 2016.

[5] S. Duncan Jr. Nonverbal communication. *Psychological Bulletin*, 72(2):118, 1969.

[6] A. R. Fayjie, A. Ramezani, D. Oualid, and D. J. Lee. Voice enabled smart drone control. In *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, pages 119–121. IEEE, 2017.

[7] W. Giernacki, M. Skwierczyński, W. Witwicki, P. Wroński, and P. Kozierski. Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering. In *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 37–42, 2017.

[8] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[9] J. P. Hansen, A. Alapetite, I. S. MacKenzie, and E. Møllenbach. The use of gaze to control drones. In *Proceedings of the symposium on eye tracking research and applications*, pages 27–34, 2014.

[10] T. Hossain, M. S. Islam, M. A. R. Ahad, and S. Inoue. Human activity recognition using earable device. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 81–84, 2019.

[11] C. Hung, G. Ananthanarayanan, P. Bodik, L. Golubchik, M. Yu, P. Bahl, and M. Philipose. Videoedge: Processing camera streams using hierarchical clusters. In *IEEE SEC'18*.

[12] U. Hunkeler, H. L. Truong, and A. Stanford-Clark. Mqtt-s—a publish/subscribe protocol for wireless sensor networks. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*, pages 791–798. IEEE, 2008.

[13] S. Jain, G. Ananthanarayanan, J. Jiang, Y. Shu, and J. Gonzalez. Scaling video analytics systems to large camera deployments. In *Proc. of ACM HotMobile*, 2019.

[14] K. Jayarajah, D. Wanniarachchige, T. Abdelzaher, and A. Misra. Comai: Enabling lightweight, collaborative intelligence by retrofitting vision dnns. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 41–50.

[15] J. Jiang, G. Ananthanarayanan, P. Bodik, S. Sen, and I. Stoica. Chameleon: scalable adaptation of video analytics. In *Proc. of ACM SIGCOMM'18*.

[16] M. Landau and S. Van Delden. A system architecture for hands-free uav drone control using intuitive voice commands. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction*, pages 181–182, 2017.

[17] J. Lee and a. et. On tracking realistic targets in a megacity with contested air and spectrum access. *MILCOM'18*.

[18] J. MacMillan, C. L. Johnson, M. P. Linegang, C. A. Miller, and R. Parasuraman. Human control of teams of unmanned/robotic vehicles: Exploring the limits of the possible. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 48, pages 523–527. SAGE Publications Sage CA: Los Angeles, CA, 2004.

[19] N. Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

[20] A. Menshchikov, D. Ermilov, I. Dranitsky, L. Kupchenko, M. Panov, M. Fedorov, and A. Somov. Data-driven body-machine interface for drone intuitive control through voice and gestures. In *IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society*, volume 1, pages 5602–5609, 2019.

[21] T. Müezzinoğlu and M. Karaköse. An intelligent human–unmanned aerial vehicle interaction approach in real time based on machine learning using wearable gloves. *Sensors*, 21(5):1766, 2021.

[22] C. Pfeiffer, S. Wengeler, A. Loquercio, and D. Scaramuzza. Visual attention prediction improves performance of autonomous drone racing agents. *Plos one*, 17(3):e0264471, 2022.

[23] J. Prakash, Z. Yang, Y.-L. Wei, H. Hassanieh, and R. R. Choudhury. Earsense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, pages 1–13, 2020.

[24] H. Qiu, X. Liu, S. Rallapalli, A. J. Bency, K. Chan, R. Urgaonkar, B. Manjunath, and R. Govindan. Kestrel: Video analytics for augmented multi-camera vehicle tracking. In *Proc. of IEEE/ACM IoTDI'18*. IEEE, 2018.

[25] C. A. Ruf, D. De Massari, A. Furdea, T. Matuz, C. Fioravanti, L. Van Der Heiden, S. Halder, and N. Birbaumer. Semantic classical conditioning and brain-computer interface control: encoding of affirmative and negative thinking. *Frontiers in Neuroscience*, 7:23, 2013.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[27] U. Steinhoff and B. Schiele. Dead reckoning from the pocket-an experimental study. In *2010 IEEE international conference on pervasive computing and communications (PerCom)*, pages 162–170. IEEE, 2010.

[28] K. Stowers, L. L. Brady, C. MacLellan, R. Wohleber, and E. Salas. Improving teamwork competencies in human-machine teams: Perspectives from team science. *Frontiers in Psychology*, page 1669, 2021.

[29] D. Weerakoon, V. Subbaraju, T. Tran, and A. Misra. Cosm2ic: Optimizing real-time multi-modal instruction comprehension. *IEEE Robotics and Automation Letters*, 2022.

[30] M. Yu, Y. Lin, D. Schmidt, X. Wang, and Y. Wang. Human-robot interaction based on gaze gestures for the drone teleoperation. *Journal of Eye Movement Research*, 7(4):1–14, 2014.