This paper has been accepted for publication at the Workshop on Aerial Robotic Systems Physically Interacting with the Environment (AIRPHARO), Biograd na Moru, 2021.

UAV human teleoperation using event-based and frame-based cameras

J.P. Rodríguez-Gómez, R. Tapia, A. Gómez Eguíluz
J.R. Martínez-de Dios and A. Ollero

***Abstract*— Teleoperation is a crucial aspect for human-robot interaction with unmanned aerial vehicles (UAVs) applications. Fast perception processing is required to ensure robustness, precision, and safety. Event cameras are neuromorphic sensors that provide low latency response, high dynamic range and low power consumption. Although classical image-based methods have been extensively used for human-robot interaction tasks, responsiveness is limited by their processing rates. This paper presents a human-robot teleoperation scheme for UAVs that exploits the advantages of both traditional and event cameras. The proposed scheme was tested in teleoperation missions where the pose of a multirotor robot is controlled in real time using human gestures detected from events.**

***Index Terms*— Robotic perception, teleoperation, UAS, UAV, event-based vision, human detection.**

I. Introduction

In a near future, robots are expected to interact with humans in many collaborative activities. Robots can perform dangerous tasks, execute actions that require very high accuracy, and provide real-time information that might not be available for a human operator. For instance, aerial robots can manipulate objects, provide additional information, or transport assets and tools [1]. However, state-of-the-art aerial robots are potentially dangerous for humans, specially multirotor platforms. Endowing aerial robots with very reactive teleoperation systems is required to guarantee safe human-robot interaction.

Although significant research effort has been devoted over decades to endow robots with the skills required for physical collaboration with humans [2], there is still a need for reliable systems capable of reacting to unexpected events during the human-robot interaction. Event cameras are neuromorphic sensors that capture illumination changes at μ-second time resolution. They are very fast, robust against different illumination conditions, and do not suffer from motion blur. The advent of event cameras occurred during the last decade and the developments of perception systems that fully exploit the sequential and asynchronous nature of event-based vision are still far from those implemented with conventional cameras. However, the potential of event cameras and their application in robotics is very high and can help to develop reactive and reliable perception systems that endow robots with the required capabilities to perceive the environment and interact with humans.

This work was supported by the European Research Council as part of GRIFFIN ERC Advanced Grant 2017 (Action 788247), and the European Commission as part of AERIAL-CORE project (H2020-2019-87147).
The authors are with the GRVC Robotics Laboratory, University of Seville, Seville 41092, Spain email: {jrodriguezg, raultapia, ageguiluz, jdedios, aollero}@us.es

This work presents a hybrid scheme for Unmanned Aerial Vehicle (UAV) teleoperation using event-based and conventional vision. The proposed method detects the human operator and reacts to predefined gestures, which are used to teleoperate the robot. To the knowledge of the authors, this is the first robot teleoperation method relying on event-based vision.

The paper is structured as follows. Section II reviews the existing literature in event-based vision for robotic applications. The proposed method for UAV teleoperation is detailed in Section III. Section IV presents an experimental evaluation of the proposed method using a real multirotor robot. Finally, Section V concludes the paper and presents future directions.

II. Related work

In the last years event-based vision has attracted increasing research interest in the robotics and computer vision communities [3]. Most of the existing works have focused on the development of event-based methods for well known fundamental problems such as feature detection and tracking [4], optical flow estimation [5], depth estimation [6], robot localisation [7], motion and object segmentation [8], object detection [9], feedback control [10], and visual servoing [11], among others.

Although, these works have developed novel methods that cope with the intrinsic nature of event cameras, few of them have focused on real robotics use cases. Some of the existing application-oriented research works have used event-based vision for surveillance tasks [12], pedestrian detection [13], gesture detection [14], autonomous driving [15], and object manipulation [16].

In particular, we are interested in those works that explored the advantages of event cameras on aerial robots. The work in [17] presents a method for detecting and tracking moving objects onboard a Micro Aerial Vehicle (MAV). The global motion is compensated using a model of the affine transformation between two consecutive event images and the moving objects are assumed to be represented by the resulting events. An autonomous landing approach for MAVs is presented in [18] relying on the optical flow from a top-down perspective. The authors perform a comparative study with other landing approaches that use conventional cameras, and their results show their method being the most accurate at high speed. A system for high-speed dodging using a UAV is proposed in [19]. *Event images* are fed to a Deep Learning method that detects and estimates the 3D motion of moving objects and avoids collisions. Another system to perform evasive manuevers is presented in [20], where the

spatio-temporal continuity of events is used to detect and track the moving objects. Then, a potential field method is used to execute the dodging maneuver by the UAV. Recently, [21] uses *event images* in a Visual-Inertial Odometry (VIO) approach in order to perform closed loop autonomous flight subject to failures.

Although the output of event cameras are asynchronous event streams, most of the techniques used onboard robots group the temporally-close received events in frames, i.e. *event images*. The use of *event images* reduces the required computation cost w.r.t. to event-by-event processing and, thus, allows more sophisticated processing schemes and real time computation. However, these approaches do not fully exploit the advantages of event cameras. For instance, grouping events generate trails of events similar to the motion blur effect [19], which implicitly reduce the temporal resolution of the camera.

A variety of event-by-event processing methods exists in the literature for feature detection [4], feature tracking [22], clustering [23], and pose tracking [24], among others. However, few of them have approached event-by-event processing onboard UAVs or considered computational constraints like those on board aerial robots. A tracking method of a UAV 6-DoF pose during high-speed maneuvers is performed in [24] by looking at a black square on a wall. Inspired by how pigeons approacch perching, the work in [11] presents a time-to-contact based visual servoing method using a multirotor UAV. In [25], robot state estimation and attitude tracking are reached using a dualcopter at speeds of 1600 deg/s.

All of the works detailed in this section provide results that contribute to the use of event-based vision onboard UAV for robotic applications. However, to the knowledge of the authors, none of them have approached the human teleoperation of a robot using event-based vision, which is a critical task to allow human-robot interaction for aerial vehicles. This paper presents an event-by-event scheme to allow humans to teleoperate a UAV. The event stream and conventional images provided by a DAVIS 346 camera are combined to detect the human and identify the moving command gestures, which are online executed by the UAV.

III. UAV TELEOPERATION USING MULTIMODAL VISION

Teleoperation onboard UAVs requires a low latency response to improve precision, robustness, and safety. Human gesture detection have been widely studied using traditional cameras. Despite several methods have been reported in the literature [26], their response is limited to the camera frame rate (typically 30 Hz). Event cameras offer μs resolution and high dynamic range (>120dB), which are critical specifications in teleoperation tasks onboard mobile robots. Event cameras return asynchronous events triggered by changes of illumination in the scene. However, since the event representation differs from traditional image format, additional processing is required to use the event stream with classic computer vision algorithms. Thus, hybrid methods using events and frames report a suitable approach to exploit the advantages offered by both sensors.

The proposed method fuses information gathered from both conventional framed images and events. The DAVIS346 sensor provides frames from the Active Pixel Sensor (APS) and asynchronous events from the Dynamic Vision Sensor (DVS). Our scheme uses an image-based object detector configured for people detection in order to provide the location of a human operator in the image plane. Gestures are detected by analyzing the evolution of events in the region reported by the people detector. Gesture information is updated using the event packages provided at high frequency by ASAP [27], which adapts the event packaging to transmit events efficiently. Detected gestures are used to define the new position of the aerial platform in the scenario using a state machine. Each state describes a hovering position for the UAV. The state machine sends the new waypoint position to the robot using the UAL abstraction layer [28], an abstract interface developed to simplify the interaction with aerial robots enabling UAV pose control from different references such a waypoints or velocity commands. Fig. 1 shows the block diagram of the proposed vision-based gesture teleoperation scheme.

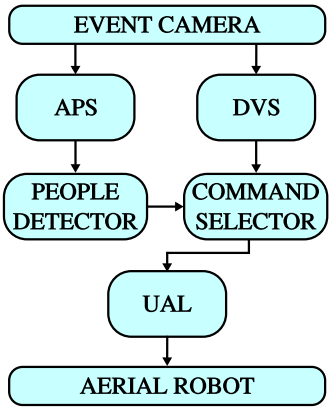


Fig. 1: Block diagram of the event-based gesture teleoperation system.

A. Event-based vision

Event cameras capture illumination changes in the form of events. They respond to the pixel level changes of the log photocurrent $I(\mathbf{x}, t)$ [3]. An event $\mathbf{e} = (\mathbf{x}, t, p)$ with polarity p is triggered at time t in the pixel coordinates $\mathbf{x} = (u, v)$ when the brightness variation in a pixel reaches a threshold c:

$$|logI(\mathbf{x}, t) - logI(\mathbf{x}, t - \Delta t)| > c, \tag{1}$$

where Δt is the time since the last event was triggered, i.e. at the pixel coordinates $\mathbf{x}$. The events triggered by the DAVIS346 camera are packaged using ASAP, which is adopted to provide event packages at high frequency (>200 Hz).

B. People detection using frames

The proposed gesture detector relies in the YOLO [29] object detector. Despite some event-based methods for people detection have been reported, they still require additional development to overcome frame-based methods in terms of accuracy and real-time processing. YOLO returns a bounding box of the detected object location and its detection probability. Bounding box information is preferred as it enables enclosing the region occupied by the detected object instead of a fixed area around the object. The method is configured to use the last layer of the YOLO neural network to perform only people detection. Fig. 2-a shows an example of a person detected using frames from the APS sensor of the DAVIS346.

C. Event-based vision for gesture commands detection

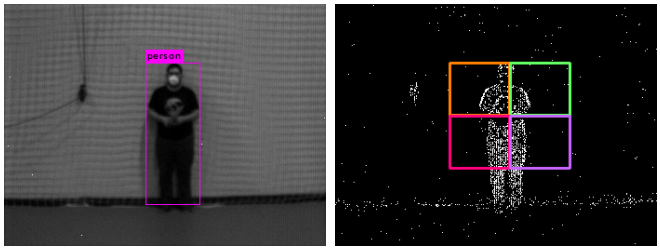
The proposed method detects gestures using event information and the reference bounding box provided by the object detector. Our approach focuses on detecting gestures from events triggered by the movement of the user limbs in specific regions of the image. The set of detection regions $\mathbf{Z} = [Z_0, ...Z_i]$ is defined using the bounding box coordinates provided by the person detector. An event belongs to a region $Z_i \in \mathbb{R}^2$ if it lies inside its boundaries i.e., $\mathbf{x} \in Z_i \Leftrightarrow u_{min} \leq u \leq u_{max}, v_{min} \leq v \leq v_{max}$. The influence of events is determined by the event occurrence in each region Z_i by:

$$\eta_i = \sum_{k=0}^{N} \sum_{\mathbf{x} \in Z_i} \delta(\mathbf{x}_k - \mathbf{x}), \tag{2}$$

where $\mathbf{x}_k$ is the pixel coordinate of event k, N is the number of events analyzed, δ is the Dirac delta, and η_i is the event accumulator of Z_i. Thus, η_i is the sum of all occurrences in region Z_i.

A total of four detection regions around the user body define the locations on the image where the gestures are expected to occur. Fig. 2-b shows an *event-image* by accumulating events (white) during 25 ms within each of the detection regions Z_i, highlighted by squares of different colors: Z_0 (orange), Z_1 (green), Z_2 (pink), and Z_3 (purple).

Gesture commands are defined using the event occurrence in each region. Four types of commands are defined:



(a) (b)

Fig. 2: YOLO person detection and candidate gesture areas during a teloperation mission: (a) person detected using YOLOV3; and (b) candidate zones for gesture recognition. Events are triggered due to the motion of the quadrotor while flying in front of the user.

$$\begin{array}{lll} Right & if & \hat{\eta_0} \geq \kappa, \eta_0 \geq \lambda, \\ Left & if & \hat{\eta_1} \geq \kappa, \eta_1 \geq \lambda, \\ Up & if & \hat{\eta_0} + \hat{\eta_1} \geq \kappa, \eta_0 + \eta_1 \geq \lambda, \\ Down & if & \hat{\eta_2} + \hat{\eta_3} \geq \kappa, \eta_2 + \eta_3 \geq \lambda, \end{array}$$

where λ is the threshold of event occurrence, $\hat{\eta_i} = \eta_i / \eta_T$ is the normalized occurrence, η_T is the number of events in all regions, and $\kappa \in [0, 1]$ is the occurrence priority w.r.t. η_T.

IV. Experimental Results

The proposed method was experimentally validated using a real multirotor platform. In the experiments, a human operator executed different gestures to teleoperate the robot while flying autonomously. The experimental scenario was the GRVC Robotics Lab indoor flight arena of the University of Seville, which equips a motion capture system with 24 *OptiTrack Prime*^{x}13 cameras that provided millimeter-accuracy robot pose estimations.

The experimental platform consisted of a *DJI Flamewheel F450* frame with a *PixRacer* autopilot (see Fig. 3) that equiped an on-board *DAVIS346* event camera. A low-cost *Waveshare NVIDIA Jetson Nano 2GB Developer Kit* board was used for running the different modules of the proposed scheme and for logging the results. *ASAP* was used on top of the UAL abstraction layer using *ROS Melodic* and the *PX4* low-level controller.

A total of 9 locations were selected to move the robot in the flight arena. The robot performed autonomously during the experiments while reacting to the human operator commands. Fig. 4 shows the experimental scenario with the 9 different locations in the scene. Each location describes a state of the state machine along with the UAV position. Fig. 5 shows the state machine used in the experiments. Teleoperation gestures were used to command the robot among the different states. The proposed method was extensively validated by moving the UAV in different positions of the scenario using the input gestures. Parameters κ and λ were set to 0.65 and 250 by empirical testing. Fig. 6 displays

Fig. 3: Aerial robot based on *DJI Flamewheel F450* equipped with a *DAVIS 346* event camera and a low-cost *Waveshare NVIDIA Jetson Nano 2GB Developer Kit* used for on-board computation.

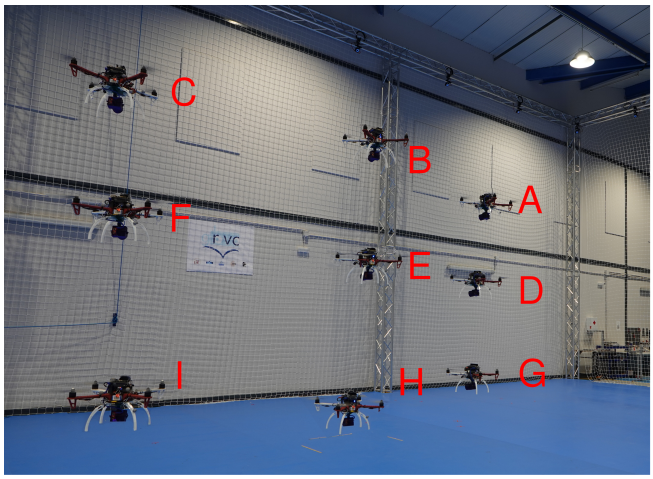


Fig. 4: Positions of the scenario to move the UAV given the input detected gesture. States are laterally-inverted to coincide with the operator perspective.

the trajectory followed by the UAV during an experiment in which the robot was guided between waypoints E, B, D, F, and H using the state machine along with the input gestures. The UAV was capable of reaching each state after correctly identifying the human operator gesture commands.

Fig. 7 shows some *event images* with the detected gesture commands. It is worth noting that the proposed method relies on *event-by-event* processing and, therefore, *event images* are shown only for visualization purposes. In all the performed experiments, the proposed method provided gesture detections at an average frequency of $200Hz$, which is 7 times higher than the typical rates obtained with frame-based methods. Although the location of the human operator in the image plane varied due to the pose configurations at the different states, the proposed method was capable of detecting the correct gesture even when the human was

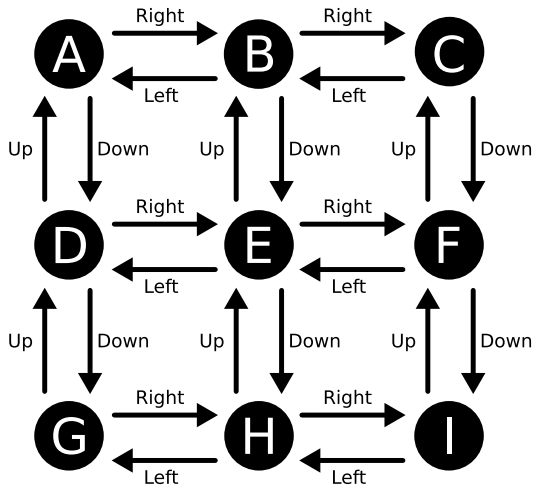


Fig. 5: Diagram of the state machine used during the experiments.

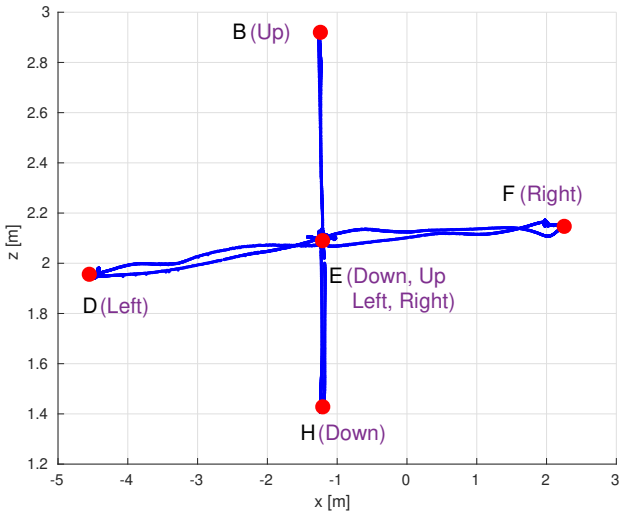


Fig. 6: Drone trajectory during a teleoperation task in the X-Z plane. Red points represent the goal position at each state while the blue line depicts the drone trajectory. The waypoints are shown in black, and the set of commands that led to those poses, in purple.

partially out of the field of view of the camera (as in Fig. 7-c). During the experiments, we observed that the gesture commands were identified in a reliable manner allowing the human operator to effectively guide the multirotor UAV between the different states. Hence, the experimental results suggest the proposed approach is a valid solution for UAV teleoperation and could be integrated in complex systems to enhance human-robot collaboration.

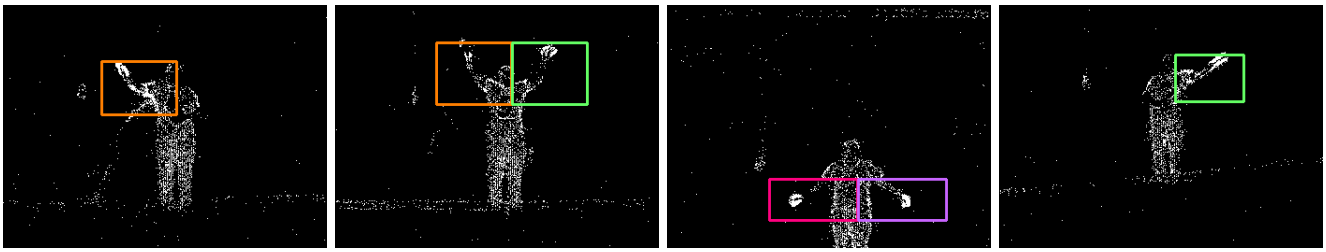
V. Conclusions and Future Work

This paper presents a method for identifying gestures that can be used to teleoperate a multirotor UAV. The method relies on an hybrid approach combining event and conventional cameras. While conventional cameras provide high reliability at human detection, event-based vision provides fast response at gesture recognition. The proposed method has been evaluated on a real multirotor platform being correctly teleoperated and showing reliable gesture identification accuracy.

This research has been conducted in the context of the H2020 AERIALCORE project, which aims at developing integrated cognitive robotic system for aerial co-working with applications in the inspection and maintenance of large linear infrastructures. Our future work will integrate the proposed approach in an integral robotic solution to perform human-robot collaboration activities at complex tasks such as, for instance, inspecting electrical lines.

References

[1] A. Ollero, G. Heredia, A. Franchi, G. Antonelli, K. Kondak, A. Sanfeliu, A. Viguria, J. R. Martinez-de Dios, F. Pierri, J. Cortés, *et al.*, "The aeroarms project: Aerial robots with advanced manipulation capabilities for inspection and maintenance," *IEEE Robotics & Automation Magazine*, vol. 25, no. 4, pp. 12–23, 2018.



(a) *Left* (b) *Up* (c) *Down* (d) *Right*

Fig. 7: Example of gestures commands using event information. (a) left, (b) up, (c) down, and (d) right.

[2] A. Gómez Eguíluz, I. Rañó, S. A. Coleman, and T. M. McGinnity, "Reliable robotic handovers through tactile sensing," *Autonomous Robots*, vol. 43, no. 7, pp. 1623–1637, 2019.

[3] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, *et al.*, "Event-based vision: A survey," *arXiv preprint arXiv:1904.08405*, 2019.

[4] R. Li, D. Shi, Y. Zhang, K. Li, and R. Li, "Fa-harris: A fast and asynchronous corner detector for event cameras," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[5] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Ev-flownet: Self-supervised optical flow estimation for event-based cameras," *Robotics: Science and Systems (RSS)*, 2018.

[6] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2822–2829, 2021.

[7] T. Fischer and M. Milford, "Event-based visual place recognition with ensembles of spatio-temporal windows," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 4, pp. 6924–6931, 2020.

[8] G. Gallego and D. Scaramuzza, "Accurate angular velocity estimation with an event camera," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 632–639, 2017.

[9] Y. Deng, Y. Li, and H. Chen, "Amae: Adaptive motion-agnostic encoder for event-based object classification," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4596–4603, 2020.

[10] J. J. Hagenaars, F. Paredes-Vallés, S. M. Bohté, and G. C. De Croon, "Evolved neuromorphic control for high speed divergence-based landings of mavs," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6239–6246, 2020.

[11] A. Gómez Eguıluz, J. Rodrıguez-Gómez, J. Martınez-de Dios, and A. Ollero, "Asynchronous event-based line tracking for time-to-contact maneuvers in uas," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[12] J. P. Rodríguez-Gómez, A. G. Eguíluz, J. R. Martínez-De Dios, and A. Ollero, "Auto-tuned event-based perception scheme for intrusion monitoring with uas," *IEEE Access*, vol. 9, pp. 44 840–44 854, 2021.

[13] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[14] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, and D. Modha, "A low power, fully event-based gesture recognition system," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7388–7397.

[15] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5419–5427.

[18] B. J. Pijnacker Hordijk, K. Y. Scheper, and G. C. De Croon, "Vertical landing for micro air vehicles using event-based optical flow," *Journal of Field Robotics*, vol. 35, no. 1, pp. 69–90, 2018.

[16] R. Muthusamy, A. Ayyad, M. Halwani, Y. Zweiri, D. Gan, and L. Seneviratne, "Neuromorphic eye-in-hand visual servoing," *IEEE Access*, 2020.

[17] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[19] N. J. Sanket, C. M. Parameshwara, C. D. Singh, A. V. Kuruttukulam, C. Fermuller, D. Scaramuzza, and Y. Aloimonos, "Evdodge: Embodied ai for high-speed dodging on a quadrotor using event cameras," 2019.

[20] D. Falanga, K. Kleber, and D. Scaramuzza, "Dynamic obstacle avoidance for quadrotors with event cameras," *Science Robotics*, vol. 5, no. 40, 2020.

[21] S. Sun, G. Cioffi, C. de Visser, and D. Scaramuzza, "Autonomous quadrotor flight despite rotor failure with onboard vision sensors: Frames vs. events," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 580–587, 2021.

[22] I. Alzugaray and M. Chli, "Ace: An efficient asynchronous corner tracker for event cameras," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 653–661.

[23] J. P. Rodríguez-Gómez, A. G. Eguíluz, J. Martínez-de Dios, and A. Ollero, "Asynchronous event-based clustering and tracking for intrusion monitoring in uas," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8518–8524.

[24] E. Mueggler, B. Huber, and D. Scaramuzza, "Event-based, 6-dof pose tracking for high-speed maneuvers," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2761–2768.

[25] R. S. Dimitrova, M. Gehrig, D. Brescianini, and D. Scaramuzza, "Towards low-latency high-bandwidth control of quadrotors using event cameras," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4294–4300.

[26] H. Liu and L. Wang, "Gesture recognition for human-robot collaboration: A review," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.

[27] R. Tapia, A. Gómez Eguíluz, J. Martınez-de Dios, and A. Ollero, "ASAP: Adaptive scheme for asynchronous processing of event-based vision algorithms," in *IEEE ICRA Workshop on Unconventional Sensors in Robotics*. IEEE, 2020.

[28] F. Real, A. Torres-González, P. Ramón-Soria, J. Capitán, and A. Ollero, "Unmanned aerial vehicle abstraction layer: An abstraction layer to operate unmanned aerial vehicles," *International Journal of Advanced Robotic Systems*, vol. 17, no. 4, p. 1729881420925011, 2020.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.