Revolutionizing Human-Drone Interaction: The Cutting-Edge Interfaces of Tomorrow! State of the Art

Students: Claudio Demaria, Gianluca Galvagni, Gabriele Nicchiarelli, Enrico Piacenti,
Università degli Studi di Genova

***Abstract*—Recent scientific research has introduced innovative approaches to enhance human-drone interaction in various scenarios. These approaches include gesture-based control, voice recognition, Natural Language Processing (NLP), Augmented Reality (AR), Artificial Intelligence (AI), Machine Learning (ML), and collaborative control.**

One of the most promising approaches combines natural language grounding with mixed reality (MR) to create an interface for controlling an autonomous drone. By leveraging MR technology, users can provide high-level tasks and navigational instructions to the drone using natural language commands. These commands are grounded to reward specifications within a Markov Decision Process framework.

Another quite interesting research utilizes Microsoft's HoloLens2 to allow immersive planning of drone paths for photogrammetry and inspection tasks. Users can manipulate waypoints with hand gestures to create paths for drone navigation in a 3D environment. The interface also incorporates collision avoidance by utilizing the HoloLens spatial map to identify potential obstacles ahead of time.

Another approach involves using AR as an interactive tool for human-drone interaction in autonomous navigation. The AR interface displays a reconstructed 3D map from the drone on physical surfaces, enabling users to intuitively set spatial target positions through head gaze and hand gestures.

These approaches represent significant advancements in the field of human-drone interaction by leveraging technologies like mixed reality and augmented reality. In this paper, we will make an overview of these technologies and also analyze them.

***Index Terms*—Human-drone interaction, Augmented reality (AR), Mixed reality (MR), Natural Language Processing (NLP), Gesture-Based Control, Drone navigation**

I. Introduction

With the increasing autonomy of drones, there is a growing need for intuitive and flexible interfaces to facilitate interaction between untrained users and these systems.

Traditional control interfaces, such as joystick controllers or command line APIs [1], often require technical expertise and may not be accessible to non-technical users. To address this challenge, researchers have been exploring innovative approaches to enhance human-drone interaction in different scenarios.

One promising approach makes use of natural language interfaces, which allow users to communicate their intentions to the drone using verbal instructions. By understanding and interpreting these instructions, the drone can engage in autonomous planning while navigating and avoiding obstacles. This natural language interface eliminates the need for users to become proficient in complex system interfaces, making drone control more accessible to a wider range of users [2].

Another avenue of exploration is the development of mixed reality (MR) interfaces specifically designed for drone path planning. Leveraging MR technology, such as the HoloLens2 [3], these interfaces overlay virtual graphics onto the user's physical environment. Users can intuitively draw and modify drone paths using hand gestures, while the MR system ensures collision-free operation through path validation techniques. This immersive and intuitive interaction allows users to plan drone trajectories more effectively and with greater spatial context [4].

Additionally, augmented reality (AR) interfaces offer a compelling solution by displaying the drone's reconstructed 3D map on physical surfaces in front of the operator. Using head gaze and hand gestures, users can set spatial target positions on the 3D map, facilitating precise control and navigation of the drone. This AR interface enhances situational awareness and enables users to interact with the drone and the environment in a more intuitive and immersive manner [5].

By combining natural language processing (NLP), MR technology, and AR interfaces, innovative approaches are revolutionizing human-drone interaction. These advancements empower users to control drones with ease and efficiency, facilitating a wide range of applications such as collaborative photogrammetry, environmental inspection, search, and surveillance activities. Ongoing research and user studies aim to refine these interfaces further, ensuring seamless and effective collaboration between humans and drones in complex real-world environments.

A. Difference between AR and MR

Augmented Reality and Mixed Reality are both immersive technologies, but they differ in their level of interaction with the real and virtual worlds.

Definition I.1 (Mixed Reality (MR)). MR is a combination of both VR and AR, where virtual and real-world elements are seamlessly integrated and interact with each other in real time. MR aims to create a unified and immersive experience by merging virtual objects with the physical environment, enabling the user to interact with both of these realities.

MR systems often involve wearing specialized headsets, such as Microsoft HoloLens, that allow users to see and manipulate

virtual objects within the real world. Thus, MR offers realistic virtual objects that interact with physical objects and respond to real-world lighting and spatial conditions.

Definition I.2 (**Augmented Reality (AR)**)**.** AR overlays virtual content onto the real world, blending digital information with the user's physical surroundings. Unlike VR, AR does not block out the real world but enhances it with additional digital elements.
AR is typically experienced through mobile devices, smart glasses, or heads-up displays (HUDs). It can overlay text, images, 3D models, or other digital information onto the user's view of the real world. AR allows users to interact with virtual objects while still being aware of and able to interact with their physical environment.

In summary, AR overlays digital content onto the real world [6], and MR combines virtual and real-world elements to create a seamless interactive experience between the two worlds. Each technology has its unique use cases and applications in various fields.

In particular, in drone applications, all these technologies (MR, AR, and VR) can be exploited with different benefits. In this research, we will focus on the use of AR and MR, since they significantly improve the way humans can interact with drones and their control in the environment.

B. The main distinctions between MR and AR when using a device like the HoloLens

The same device can work for both technologies, however, it will operate in different ways on the surroundings.

1) **Environment Interaction:**
   - *Augmented Reality* overlays digital content onto the real world, enhancing the user's perception and interaction with their environment. The digital content appears as virtual objects or information superimposed on the real-world view.
   - *Mixed Reality* combines virtual objects with the real world in a way that allows them to interact and coexist with the physical environment. Virtual objects can be anchored to specific locations, respond to physical objects, and maintain a sense of presence within the real world.
2) **Virtual Content Integration:**
   - *Augmented Reality* primarily focuses on overlaying virtual content onto the real world, thus the virtual objects are typically displayed on a flat surface or in a fixed location relative to the user's viewpoint.
   - *Mixed Reality* integrates virtual content seamlessly into the user's environment, allowing virtual objects to interact with the physical world. The virtual content can be occluded by real objects and respond to the user's movements and interactions in a more immersive and realistic way.
3) **Spatial Mapping and Tracking:**
   - *Augmented Reality* typically relies on basic tracking methods, such as marker-based or GPS-based tracking, to determine the user's position and orientation in relation to the virtual content. The spatial mapping may not be much meticulous and the digital content is not aware of the physical environment's detailed structure.
   - *Mixed Reality* devices, like the HoloLens, utilize advanced spatial mapping and tracking technologies to create a detailed understanding of the user's surroundings. This enables the virtual objects to interact with the real environment accurately and maintain their position relative to physical objects.
4) **Interaction Modalities:**
   - *Augmented Reality* devices often use touchscreens, gestures, voice commands, or handheld controllers as input methods to interact with the virtual content. These interactions are typically limited to the device's screen or a specific field of view.
   - *Mixed Reality* devices like the HoloLens offer more immersive and natural interaction modalities. Users can interact with virtual objects using hand gestures, gaze, voice commands, and even hand-held controllers. The spatial mapping and tracking capabilities enable precise hand and gesture tracking, making interactions more intuitive and realistic.

II. Technology Overview

In this section, technological advancements regarding immersive technology and possible implementation are discussed.

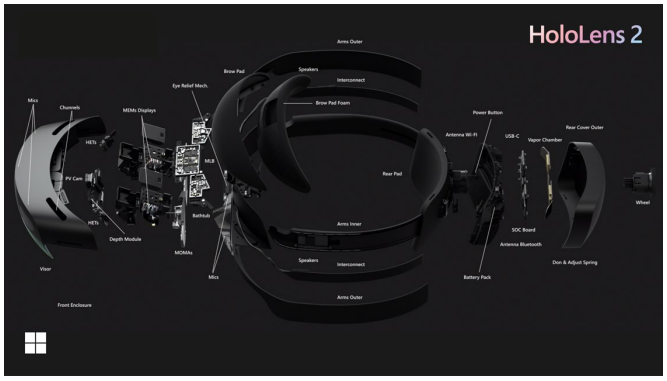

Fig. 1. Design overview of HoloLens2 by Microsoft [3]

*A. **MR for intuitive and flexible ways to control a drone***

As drones become increasingly autonomous, it is necessary to create intuitive and flexible ways for untrained users to interact with these systems: a much higher-level interface, such as natural language and mixed reality, allows the automation of the agent's control in a goal-oriented setting.

1) Natural language interface: A natural language interface is immediately accessible to non-technical users and does not require the user to use a touchscreen or radio control (RC). After a user specifies their goal using language, the robot can understand these instructions and engage in

autonomous planning to follow the instructions while avoiding obstacles.

Current natural language interfaces require a predefined model of the environment including landmarks, which is difficult for a drone to obtain[7][8].
All of these previous studies have required an a priori model of landmarks in the environment.
By contrast, our approach with MR doesn't require previous knowledge of the landmarks, but instead, users can specify virtual landmarks whose groundings are known by the language model: the mixed reality interface allows people to provide landmarks that they can then refer to by using the natural language interface, which enables people to command drones with greater flexibility.
The user can place or remove a virtual landmark at the location they are looking at by voice command or by a tap gesture in the air (it can also be dragged from one place to another with gestures). The landmark facilitates communication, as the user is now able to instruct the drone to navigate to a specific position by saying "Go to the landmark" rather than giving explicit instructions.
This interface enables the landmark objects to be specified in the drone's global frame such that it can interpret commands without a complete model of the environment.
Also, when using the MR interface people can simultaneously observe the drone and the environment while planning the task and giving commands, as compared to being forced to do these sequentially via other 2D and 3D interfaces.

Recent scientific research has done some exploratory user studies where they found that users are able to command the drone much more quickly via both MR interfaces (with and without language) as compared to the web interface, with roughly equal system usability scores across all three interfaces.
MR provides a more intuitive, user-friendly visualization than a 2D visualization tool such as Rviz[9]: in addition, using MR to control an arm or a drone is facilitated by the use of gestures and gazes[10][11].

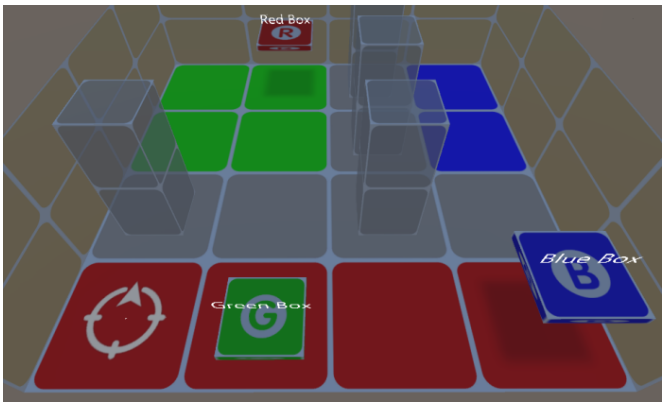

Fig. 2. The environment in MR

2) Hand gesture interface: Despite the multiple benefits provided by the aforementioned natural language interface, such an interface is limited by the lack of spatial information for collocated users when planning paths. There has also been progress in exploring more natural communication methods and making drones safer, so a study using hand natural gestures was conducted as a compromise to obtain a relatively user-friendly experience for the user while also improving the quality of information buffered from the human to the drone. In this section the Drone Brush is introduced, a prototype mixed reality interface designed for immersive planning of drone paths in tasks like collaborative photogrammetry and inspection. The interface utilizes Microsoft's HoloLens 2, enabling users to draw 3D drone navigation paths using hand gestures. By making a simple pinch gesture, users can place waypoints, delete existing waypoints, and move them around the visible space.
To ensure path safety, the interface utilizes the spatial map provided by the HoloLens, allowing for collision detection and prevention during drone navigation. Additionally, the paths are optimized using density-based clustering techniques to avoid complex or redundant drone movements.

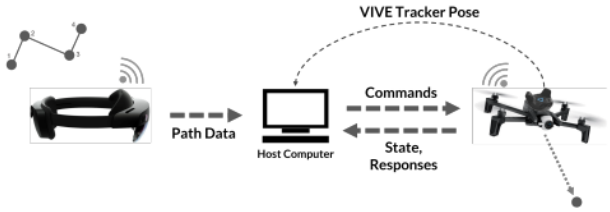

Fig. 3. Architecture of the drone brush

The Drone Brush ecosystem comprises essential components, as illustrated in the figure above: a HoloLens 2 (HL2) Headset for the user interface, a Parrot ANAFI quadrotor drone for navigation, and a host computer for system integration and drone control. Communication between the HL2 and the host computer is facilitated through ROS [12], while the Parrot ANAFI communicates via messages using the Parrot Olympe SDK [13]. Within the mixed reality environment, users can create, manipulate, and validate paths using the HL2 interface. By employing gestures, they can activate a virtual button to send path data to the host computer for execution with the drone. The system relies on real-time pose information of the drone for navigation, and in our prototype, this is obtained from an HTC VIVE tracker 2.0 mounted on the drone. To aid mobile users in navigation, their objective is to develop an easily accessible user interface. A hand-based menu is utilized, which appears when the user presents their palm towards the display, requiring eye gaze and a flat palm to avoid unintended activation. The menu offers options for creating points, editing paths, and transmitting points to the drone. Utilizing hand tracking, the system calculates the distance between finger positions to accurately position points on the path, effectively preventing accidental creation. Constraints are implemented, including a minimum distance between points and the completion of the pinch gesture.

A procedural cylinder mesh is generated to visually connect the points and represent the path. Editing tools are available for point deletion, clearing, and movement. The mesh is updated accordingly as points are added, moved, or removed, and it undergoes periodic processing. Path validation is performed to ensure the safe transmission of the drawn path to the drone, providing users with timely notifications in case of detected collisions with the environment. The interface is constructed using Unity and the Mixed Reality Toolkit (MRTK) [14].

B. AR is used to control a drone that operates beyond the operator's field of view

The drone is equipped with a stereo camera, flight controller, and onboard computer (Fig. 4) to run navigation-related modules. VINS-Fusion [15] a robust and accurate multi-sensor state estimator, is used for self-localization. Using an ESDF (Euclidean signed distance field) map and VIO (visual-inertial odometry), Fast Planner [16] is applied to generate a safe kinodynamic and smooth trajectory. It is capable of autonomously reading the target position and generating a trajectory.

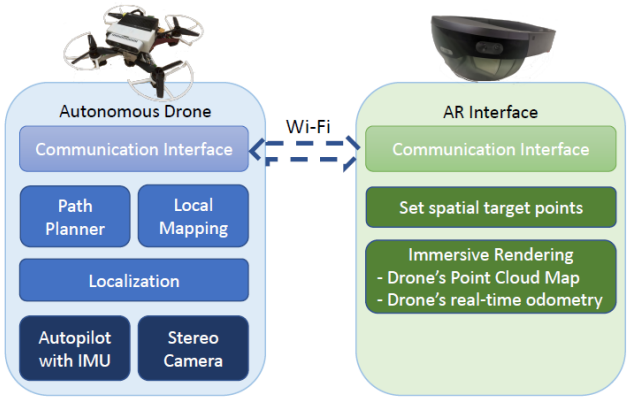

Fig. 4. The structure of a system for using an AR interface and a drone

As a result, the drone can navigate in an unknown environment and generate an occupancy map in real-time. The drone's recognized obstacles are represented by cubic voxels on the map. The interface continuously receives the occupancy map and renders it through the GPU instancing function [17], which allows the HoloLens to efficiently render multiple cubic voxels in one function.

The AR interface unit provides an immersive rendering of environments, presenting the drone's reconstructed 3D map and real-time odometry information to the operator. This dynamic visualization enables a comprehensive understanding of the drone's surroundings. Despite a time delay in updating the drone's information, the interface maintains real-time interaction without interruption. However, users may require instructions and practice to utilize the interface effectively. Another disadvantage of this interface is that it exhibits larger errors and longer response times compared to conventional interfaces like Rviz [18]. However, with practice, the longer response time can be reduced. Additionally, the error can be minimized by incorporating an axis that assists the user in accurately setting the waypoints.

III. Conclusion

In conclusion, this state-of-the-art paper highlights the significant advancements in human-drone interaction through the integration of innovative interfaces.

The combination of natural language processing, Mixed Reality, and Augmented Reality technologies has revolutionized the way users can interact with autonomous drones.

1) The use of natural language interfaces enables non-technical users to communicate their intentions to drones using everyday language. By understanding and interpreting these instructions, drones can engage in autonomous planning while navigating and avoiding obstacles. This approach eliminates the need for users to become proficient in complex system interfaces, making drone control more accessible to a wider range of individuals.
2) The development of MR interfaces, allows users to immerse themselves in the drone control process. Through hand gestures and spatial mapping, users can intuitively draw paths and set waypoints for drone navigation. MR technology provides a more immersive and context-rich interaction, enhancing the effectiveness of drone trajectory planning.
3) AR interfaces offer another compelling solution by displaying the drone's reconstructed 3D map on physical surfaces. Users can set spatial target positions using head gaze and hand gestures, facilitating precise control and navigation of the drone. This AR interface enhances situational awareness and provides an intuitive and immersive interaction with the drone's environment.

By combining these technologies, human-drone interaction has reached new levels of intuitiveness, flexibility, and accessibility.

These advancements have wide-ranging applications in collaborative photogrammetry, environmental inspection, search, and surveillance activities.

Ongoing research and user studies aim to refine these interfaces further, ensuring seamless and effective collaboration between humans and drones in complex real-world environments.

The continued exploration of natural language processing, MR, and AR technologies holds great promise for the future of human-drone interaction, empowering users to control drones with ease and efficiency.

AR and MR allow a strong interaction between the user and the environment since the user is able to interact easily with objects within the field of view, which can be very helpful to operate the drone in a controlled environment.

All the studied cases also prove that AR is an easy-to-use approach also for non-expert users as it requires short training time.

Another possible approach that can be applied to drone control operations is Virtual Reality, which allows full immersion in

the navigation of the drone. It can allow operating over the field of view of the user, enabling a wider operative space. However, VR isn't as easy to use as AR and MR. With this technology, the user is bounded to the use of bigger headsets and can also experience motion sickness.

Thus we can conclude that AR and MR devices provide a more promising and intuitive approach to the field of human-drone interaction, opening up new spaces for the evolution of this subject.

REFERENCES

[1] J. Z. X. L. H. Kang, H. Li and B. Benes, "Flycam: Multitouch gesture controlled drone gimbal photography," in *IEEE Robotics and Automation Letters (RA-L)*, pp. 3717–3724, IEEE, 2018.

[2] D. U. N. G. Baichuan Huang, Deniz Bayazit and S. Tellex, "Flight, camera, action! using natural language and mixed reality to control a drone," in *International Conference on Robotics and Automation (ICRA)*, IEEE, 2019.

[3] Microsoft, "Microsoft hololens 2." https://www.microsoft.com/en-us/hololens/hardware/, 2019.

[4] H. S. A. P. K. L. R. T. Angelos Angelopoulos, Austin Hale and D. Szafir, "Drone brush: Mixed reality drone path planning," in *Late-Breaking Report*, HRI, 2022.

[5] C. Liu and S. Shen, "An augmented reality interaction interface for autonomous drone," in *International Conference on Intelligent Robots and Systems (IROS)*, IEEE/RSJ, 2020.

[6] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *Nature*, vol. 77, no. 12, pp. 1321–1329, 1994.

[7] S. D. M. R. W. A. G. B. S. T. Stefanie Tellex, Thomas Kollar and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *AAAI Conference on Artificial Intelligence*, 2011.

[8] D. A. M. R. N. G. L. L. Siddharth Karamcheti, Edward C. Williams and S. Tellex, "A tale of two draggns: A hybrid approach for interpreting action-oriented and goal-oriented instructions," in *Annual Meeting of the Association for Computational Linguistics Workshop on Language Grounding for Robotics*, 2017.

[9] T. P. Hyeong Ryeol Kam, Sung-Ho Lee and C.-H. Kim, "Rviz: A toolkit for real domain data visualization," vol. 60, no. 2, pp. 337–345, 2015.

[10] D. K. Okan Erat, Werner Alexander Isop and D. Schmalstieg, "Drone-augmented human vision: Exocentric control for drones exploring hidden areas," in *Drone-augmented human vision*, pp. 1437–1446, IEEE, 2018.

[11] E. P. G. C. J. T. G. K. Eric Rosen, David Whitney and S. Tellex, "Communicating robot arm motion intent through mixed reality head-mounted displays," in *International Symposium on Robotics Research*, 2017.

[12] B. G. J. F. T. F. J. L. R. W. M. Quigley, K. Conley and A. Y. Ng, "Ros: an open-source robot operating system," vol. 3, no. 3.2, p. 5, 2009.

[13] P. SA, "Olympe." https://developer.parrot.com/docs/olympe/overview.html, 2021.

[14] M. Corporation, "Mrtk-unity." https://docs.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/?view=mrtkunity-2021-05l, 2021.

[15] P. L. T. Qin and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," in *preprint arXiv:1708.03852*, arXiv, 2017.

[16] L. W. C. L. B. Zhou, F. Gao and S. Shen, "Robust and efficient quadrotor trajectory generation for fast autonomous flight," *Robotics and Automation Letters (RA-L)*, vol. 4, no. 4, p. 3529–3536, 2019.

[17] Unity, "Gpu instancing." https://docs.unity3D.com/Manual/GPUInstancing/, 2019.

[18] "rviz - ros wiki." http://wiki.ros.org/rviz/, 2018.