User-Defined Gesture and Voice Control in Human-Drone Interaction for Police Operations

Julia Hermann
julia.hermann@hs-ruhrwest.de
Institute of Positive Computing, Ruhr West University of Applied Sciences
46236 Bottrop, Germany

Moritz Plückthun
moritz.plueckthun@stud.uni-due.de
Institute for Software Engineering, University of Duisburg-Essen
45127 Essen, Germany

Aysegül Dogangün
ayseguel.doganguen@hs-ruhrwest.de
Institute of Positive Computing, Ruhr West University of Applied Sciences
46236 Bottrop, Germany

Marc Hesenius
marc.hesenius@uni-due.de
Institute for Software Engineering, University of Duisburg-Essen
45127 Essen, Germany

ABSTRACT

Gesture and voice control are increasingly being used in everyday applications, such as tablets and smartphones, but also for controlling smart home systems or even drones. While several studies exist on how users interact with and issue commands to drones, studies using drones in very specific and highly specialized use cases are rare. In a user-centered approach with twelve German police officers, we examined how police forces would trigger drone functions through self-defined gestures and voice commands. For the study, we considered two deployment scenarios with a total of 21 functions that were developed together with the police. The focus here is on use in large crowds and the pursuit of suspects. We identify sets of custom gestures and possible voice commands.

CCS CONCEPTS

• **Human-centered computing → User centered design**; *Empirical studies in interaction design*; User interface design.

KEYWORDS

Human-Drone Interaction, Gesture Control, Voice Control, Natural User Interfaces, Drones, UAVs, Police, User-Defined, User centered

ACM Reference Format:
Julia Hermann, Moritz Plückthun, Aysegül Dogangün, and Marc Hesenius. 2022. User-Defined Gesture and Voice Control in Human-Drone Interaction for Police Operations. In *Nordic Human-Computer Interaction Conference (NordiCHI '22), October 8–12, 2022, Aarhus, Denmark.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3546155.3546661

1 INTRODUCTION

The field of human-drone interaction has grown rapidly in recent years, because drones – also referred to as Unmanned Aerial Vehicles (UAVs) – are becoming more and more visible in different fields of application. Today's drones offer a variety of functions, most prominently the possibility to create video or photo recordings from lofty heights, but drones can also follow a defined target or carry small loads. Such features open up a wide range of applications and interesting use cases for drones have been suggested in a variety of professional settings: for example, applications in taking inventory [11], firefighting [4], and police work [10] have been discussed.

In police work, a drone could follow potentially violent persons and support officers with information about their actions and whereabouts or make video recordings to document crime scene investigations. In addition, drones can perform flexible reconnaissance. Predefined flight routes, such as circling the current position with a defined radius or heading for different waypoints, can be used for fast and precise sighting of the surroundings. Besides hazard detection, this can also be used for more precise intervention planning, for example to support special forces. Another application for reconnaissance work is the observation of groups of people, for example at sporting events or demonstrations. This allows people to be viewed from different perspectives and possible sources of danger to be identified.

To control drones, physical remote controls or apps running on mobile devices, e.g., smartphones or tablets, are often used, but both approaches are unsuitable for police work: since officers would be busy controlling the drone with a device in their hands, their scope for other actions would be limited. Especially when working in smaller teams (e.g., police cars are often manned with two officers), remote control of a drone is disruptive because users are limited in their ability to perform other tasks. A more autonomous drone is thus desirable, functioning as an additional colleague with supporting functions, and are commercially available. This autonomous behavior greatly simplifies control in particular, since only changes in position or function are transmitted to the drone [26], yet it still requires a way to receive commands. Gestures or voice commands come quickly into mind as appropriate interaction modalities: they

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
NordiCHI '22, October 8–12, 2022, Aarhus, Denmark
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9699-8/22/10…$15.00
https://doi.org/10.1145/3546155.3546661

allow for direct control without additional devices and are also becoming more and more common in mainstream apps, but most prominently they do not require police officers to use additional devices and can thus be embedded – more or less – easily into everyday work.

Designing drone control in such a way requires to select appropriate gestures or voice commands for a specific purpose [9, 21] with regard to the use case and the targeted user group. Individual control profiles are technically feasible, would just need a short training sequence with user and drone, and might offer the potential to simplify drone interaction for one specific user, but they are not feasible in most practical settings. Most prominently, they would require specific configuration for all potential users, but drones could not be flexibly taken over by other officers during operations. Thus a uniform set of gestures and voice commands is necessary that can be intuitively understood and used by every police officer. We understand *intuition* in the sense that the interaction must be adapted to how the targeted user group (here: police officers) are accustomed to working so they can easily learn to handle a drone from their everyday routine. However, such an approach requires additional knowledge from the users' domain, because the subtleties of everyday police work may be unknown to developers.

We aimed to study how police officers would command a drone during operations and opted for a user-centered development approach in the form of an elicitation study to identify suitable gestures and voice commands for drone control in police operations. We presented 12 German police officers with several tasks in two different usage scenarios, in total a set of 21 possible drone referents, and asked for potential gestures and voice commands. From this data, we derived a user-defined set of gesture and voice commands that can serve as a foundation for designing appropriate interaction measures when developing drones for such specialized user cases. In this paper, we present the elicited gestures and voice commands.

We will first review related work with regard to drone control and our co-creation approach in Section 2. We then describe our study design in Section 3 and the results in Section 4. We discuss the results and their implications in Section 5 and conclude the paper with an outlook on future research in Section 6.

2 RELATED WORK

User Interfaces (UIs) using gesture and voice control have been researched intensively in recent years. Studies with users to derive potential gestures, often dubbed *user-defined gesture sets*, have been conducted for different use cases, devices, and sensors. Also, the interaction with robots in general and drones in particular has been investigated and various approaches been developed, showing how different modalities can be used to exchange information and issue commands. We will review the state of the art in three areas related to our work: Interaction between human and drones in general and with gestures and voice commands in particular, collecting user-defined gesture sets, and using drones in professional settings.

Different researchers dealt with the various command options in human-drone interaction, such as using a remote control [3], wearables [26], or voice control [14, 15, 32]. However, a large number of studies have focused on gesture control [7, 17, 19–21]. Cauchard et al. [8] use a projected graphical UI to interact with a drone that also employs gestures to issue commands. Other studies deal with direct interaction with the drone through sensors, detecting the user's facial poses [6, 18], or touching the drone [2]. Suárez Fernández et al. [28] experimented with several UIs using voice, posture, and hand gestures for drone control. Pfeil et al. [23] studied upper body gestural interaction in controlled laboratory conditions where users were given specific interaction metaphors. Wuth et al. [32] report that participants perceived robots using synthesized voices as transparent and efficient compared with beeps known from robots in popular movies. Their participants furthermore preferred voice commands over textual input when issuing instructions to the robot. Seeing a drone as a type of robot makes it feasible to simply transfer such results to human-drone interaction. This efficient communication relies on knowing the context for the specific task being tackled by the drone. Overall, the context in human-drone environments plays an important role to achieve effective control of the UAV. There are already efforts to control UAVs with simple voice commands. Landau and van Delden [14] describe a suitable architecture for such systems. Abioye et al. [1] combine speech and gesture control into a model of multimodal control for aerobots operating at higher autonomy levels in the context of a patrol, search, and rescue application.

These works show the versatility of human-drone interaction and emphasize the potential for using different modalities, but often focus on technical aspects (i.e., sensors and processing algorithms) and validate whether the employed hard- and software works as intended. Implementing use cases in highly constrained application domains, however, requires to thoroughly analyse how the interaction must be designed and how it fits the type of work in the application domain. Such applications require detailed knowledge of the application domain, which can be obtained by including future users in the development. For our chosen application domain, police work, we make a first step in creating the foundation for suitable interaction concepts, which requires to collect suitable gestures and voice commands from the intended user group.

Wobbrock et al. [31] introduced the blueprint for deriving gestures with users and showed how to develop *user-defined gesture sets* in *elicitation studies*. While they focused on surface gestures for simple computing tasks, their approach has been widely adopted by various research groups: Villarreal-Narvaez et al. [30] review over 200 gesture elicitation studies. Several gesture elicitation studies dealt specifically with gesture control of drones. For example, Obaid et al. [21], E et al. [9], and Cauchard et al. [7] studied gesture control of drones for use by private consumers. Here, different user-defined gesture sets were elicited: for example, for how to use the drone for self-photography or to execute basic commands that enable the drone to be positioned as desired. Further work examined the control of drones with gestures during specific leisure activities, e.g., when running accompanied by a drone [27]. The studies by Cauchard et al. [7] and E et al. [9] take cultural differences into account when defining gesture sets and thus highlight the influence of additional factors on gesture definition. The bulk of this work was aimed at common lay users and everyday tasks that can be performed with drones available to consumers, and while some of the basics might also be valid for professional users, they might also have more distinct requirements: for example, lay people will probably reuse common gestures from everyday interaction with others (in the context of their respective cultural background), but professional users from a domain actively using gesture-based communication (e.g., military) may resort to prefer this type of gesture to reduce learning barriers.

Others have investigated how humans and drones interact via gestures in professional scenarios that require fast decision making, even under circumstances where information is crucial, but not completely available. Drones have been hailed as a means for providing previously hard to acquire information. Besides police work, possible use cases originate in firefighting and military support and several authors investigated the use of gestures in such scenarios. Taralle et al. [29], for example, deal with supporting infantrymen during missions in hostile environments via drones and the corresponding interaction. Mission planning takes on a higher priority and the gestures are primarily used to activate individual waypoints. Other user-centric studies have examined the integration of drones in firefighting. Alon et al. [4] studied firefighting tasks for drones and appropriate custom gestures that can be performed in highly constrained environments. The focus was on human-drone interaction in emergency response and other complex professional environments. Medeiros et al. [16] interview several firefighters and derive from the results that they prefer to use a combination of gestures and voice commands. Most notably, the authors' results hint that pointing to a target and uttering a voice command in combination, thus using multimodality, is the favored way of interaction. Our initial discussions with police officers before setting up the study, however, indicate that multimodality is not suited to this type of work (see Section 3.2 for details), emphasizing the need for specialized, domain-specific gesture sets.

3 STUDY DESIGN

For the study, we developed useful scenarios for police operations in co-creation with representatives from the targeted user group, police officers. Within these scenarios, the drone supports police officers in their work. These tasks range from general video documentation to targeted reconnaissance of the area. We used this tasks as the foundation for the next steps, i.e., studying which gestures and voice commands would police officers define for interacting with the drone. To identify a set of recommended gestures and voice commands, we followed the procedures used in earlier studies, e.g., Obaid et al. [21] and E et al. [9], as well as further studies on gesture control in robotics [24–26] and aspects of gesture recognition over larger distances [12]. We will first review the tasks used in Section 3.1 and then describe the general setup in Section 3.2. We then describe our participants in Section 3.3 and the procedures for data analysis in Section 3.4.

3.1 Scenarios and Tasks

To define realistic scenarios and interesting tasks for using drones in police work, we first evaluated potential use cases with a domain expert: A police officer from the area of *Advanced Training* supported us to find a suitable foundation for our study. In cooperation with him, possible deployment scenarios were designed and their potential for a gesture- and voice-controlled drone evaluated. From the various scenarios, the evaluation resulted in two most interesting areas: big public events and patrol. Big public events might include demonstrations, sport events (e.g., soccer games), or music festivals and police officers thus have to deal with a (potentially) large area and a large crowd of people, leading to situations that are hard to oversee. On patrol, while there are less people involved, situations might change quickly and thus require a different perspective on a scene to, e.g., collect evidence or track suspects with thermal images to identify hiding spots. Furthermore, drones can support with reconnaissance during operations on impassable terrain where visibility is restricted by obstacles.

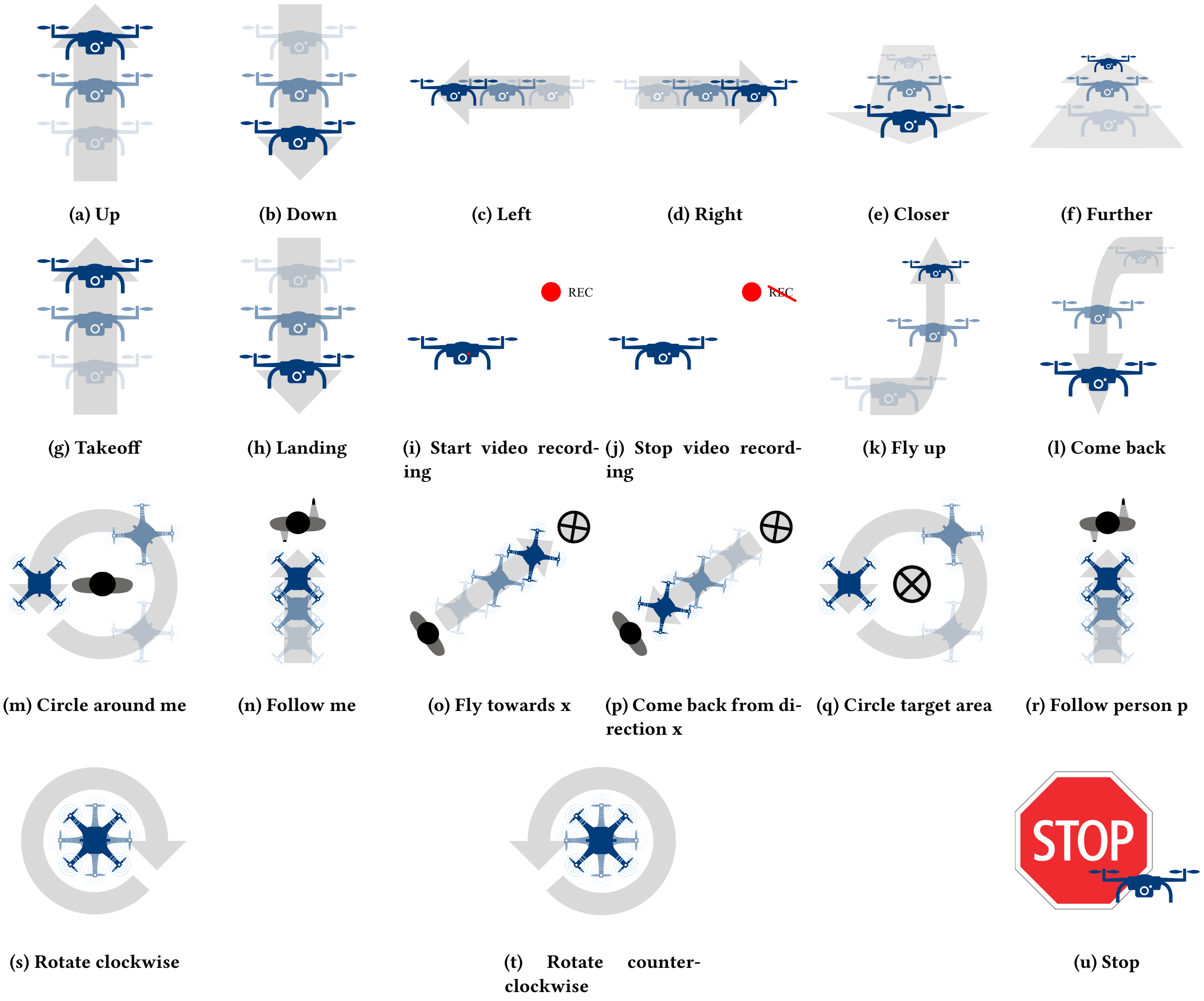
Based on these two scenarios, we continued to develop a set of suitable tasks in which a drone could be beneficial in co-creation with our consulting police officers. From these tasks, we derived 21 referents that drones should provide and that police officers can command the drone to perform. In addition to basic control elements, such as *Take Off* and *Landing*, other more sophisticated referents, like *Fly Towards Something* or *Circle the Target Area*, are required. In Figure 1, all referents we used for our study are listed.

Initially, we considered to use the area of an old coal mine for the study, which would have allowed us to demonstrate the various referents to the participants and have them interact with a real drone, but unfortunately we had to abandon this idea after the firsts tests for several reasons. The area and the surrounding buildings from the coal mine negatively affected our drones sensors and made especially navigation and communication error-prone. Another problem was the distance to the site. While the police management initially allowed their officers to participate in our study during their hours of service, they would have been required to travel to the study site, prolonging the necessary time frame for each session beyond the limits possible. In order to avoid time-consuming travel, we decided to move the study to seminar rooms in the police office. For safety reasons, flying a drone indoors was forbidden so we had to show the previously filmed drone actions via video and gave graphical information and verbal explanations.

3.2 Study Setup

We met the participants alone in unique sessions with no further person present. The participants first received a fact sheet detailing several aspects of the study and gave their consent to participation and the creation, evaluation, and subsequent use of video and audio recordings for scientific purposes. We then handed them a questionnaire asking for demographic data (age, gender, professional position, and left or right handed) and previous experience with gesture and voice control in general as well as gesture control of drones were asked. Furthermore, questions were asked about technology affinity and hobbies. If gesture- or language-intensive hobbies are carried out during free time, such as that of a soccer coach, it must be assumed that this also has an effect on the type of operation. This questionnaire also presented participants some statements about their expectations regarding the use of drones in police work and their thoughts on using gesture and voice control during operations. The participants then received a description of all scenarios and illustrations of the various tasks the drone could perform to gain a first look of what we would discuss with them.

After the paper work, we started the collection of gestures and voice commands. For this purpose, the current participant was positioned at a distance of about five meters facing a Microsoft Kinect.



(a) Up (b) Down (c) Left (d) Right (e) Closer (f) Further

(g) Takeoff (h) Landing (i) Start video recording (j) Stop video recording (k) Fly up (l) Come back

(m) Circle around me (n) Follow me (o) Fly towards x (p) Come back from direction x (q) Circle target area (r) Follow person p

(s) Rotate clockwise (t) Rotate counter-clockwise (u) Stop

Figure 1: List of all Referents

A second video camera was positioned slightly offset to provide the clearest possible view of the participant's movements. The camera also recorded the session's audio for later analysis following the idea of *think aloud*. Kinect was used to collect additional data and gather better insight into the nature of the movement and its components and also offered, thanks to precise timing, the possibility to divide and measure the movements in their individual phases in detail. There was no direct control of the drone due to the local conditions. In addition to the graphical information and the verbal explanation, we showed a video recording of each referent and the corresponding drone action in order to eliminate possible ambiguities from any written or oral explanation as far as possible. Figure 2 shows a schematic diagram of the experimental setup. Although we provided participants a list of all referents at the beginning of the session, during data collection referents were presented at random to remove potential learning effects and biases. Participants also had no time restrictions and could progress through the various referents on their own speed. Directly after defining each gesture and voice command, we asked the participants to rate how difficult they found it to define appropriate gestures, which corresponds to the way Obaid et al. [21] designed their study.

We asked the participants to perform a gesture for each referent and to name a voice command they thought to be suitable. Responding data were collected separately, do not build on each other, and are not interlinked. We thus specifically refrained from collecting multimodal commands. While multimodality offers interesting capabilities and has been tied especially to combining speech and gestures (most notably since the famous *Put That There* by Bolt [5]), it has also been accompanied by several myths, e.g., that users, although they often state a strong preference for multimodal

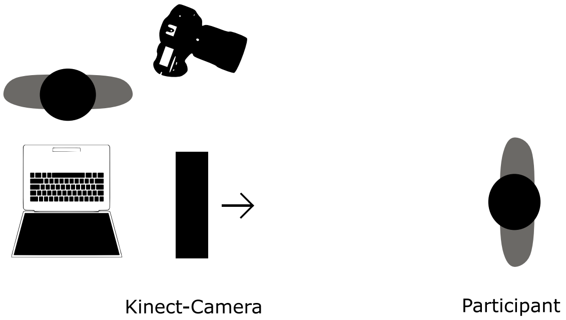


Figure 2: Sketch of the experimental setup

interaction, will also mainly use multimodal interaction [22]. As we designed potential scenarios and tasks in collaboration with police officers before collecting data, the discussions also showed that multimodality would bring only little value to our intended user group. Most prominently, chances were deemed high that police officers may simply be unable to combine speech and gestures into multimodal commands, because one of the modalities is restricted (e.g., both hands are holding something or the surroundings are too noisy).

3.3 Participants and Previous Experience

The group of participants consisted of twelve persons (three female and nine male). Nine participants were right handed and three left handed[1]. Eight participants worked at the *Directorate for Central Tasks* and four at the *Directorate for Riot Police*, the latter one being particularly relevant for the events scenario. Handedness has a distinct role, because there are slight differences in reaction time when using the weak hand, which is especially important in stressful situations. Only two participants mentioned a hobby somewhat related to a more intense use of gestures, thus we did not consider this area any further. Possible differences in the choice of a gesture could not be identified due to the small number. Participants were about 36 years old on average with a below average median of 31.5 years. Within the group of participants there was one person who already had experience with a gesture-controlled drone, but most participants stated that they had no experience with voice or gesture control. All participants were fluent in German, the language for all sessions.

3.4 Result Analysis

Collected data were analyzed in two steps. In a first step, data were aggregated to determine the majority choices of all participants for each defined gesture and voice command. In a second step, we determined agreement scores using Equation 1 from E et al. [9], where P is the set of all defined interactions for the current referent and P_i the subset of identical interactions.

$$A = \frac{|P|}{|P-1|} \sum_{P_i \subseteq P} \left(\frac{|P_i|}{|P|} \right)^2 - \frac{1}{|P|-1} \qquad (1)$$

4 RESULTS

For evaluation, we examined the gestures and voice commands that our participants defined. We grouped the gestures according to form, type, and body area used. Furthermore, we identified the commands chosen most frequently for a certain action. We also asked the participants to subjectively evaluate for each referent how hard they found it to find a suitable gesture or command. We will first review the results for defined gestures (Section 4.1) and then for voice commands (Section 4.2). We then focus on the participants' perception of the difficulty when defining gestures and voice commands (Section 4.3) and the calculation of agreement scores (Section 4.4). We will conclude this section with some general remarks collected, especially regarding the participants' impression of gesture and voice controlling drones (Section 4.5).

4.1 Defined Gestures

We reviewed all defined gestures and assigned them to categories for different forms, types of gestures, and body areas based on the classification developed by Obaid et al. [21]. The form of a gesture can be divided into static and dynamic gestures. In general, gestures consist of three phases: an initial preparation phase at the beginning (mainly to bring body parts into a starting position), a core phase (the semantically meaningful part of the gesture), and a retraction phase where users would relax again. Static gestures are motionless (and sometimes referred to as poses) during the core phase, while dynamic gestures contain movement. This classification is important in order to identify requirements for the technical implementation, i.e., suitable hard- and software. A gesture's type consists of four subcategories:

- deictic – position or direction (e.g., pointing somewhere)
- iconic – a visual depiction similar to the expected action (e.g., rising an arm when instructing the drone to rise)
- metaphoric – an abstract visual depiction (e.g., clapping hands for landing)
- emblematic – an artificial symbol (e.g., lifting the right fist for stop).

Table 1 shows the results. Looking at the share of each group within the data set, our results are close to Obaid et al. [21] for form and gesture type: our participants also preferred dynamic, deictic gestures. However, we observe a stronger preference for emblematic gestures. For the body parts group, our results differ: our participants preferred to use one hand, while the results from Obaid et al. [21] are evenly distributed for both hands.

Table 2 contains an overview of all referents and the selected gestures. The listed movements were selected by at least two participants, thus the threshold for inclusion in this list is 16.7 % of occurrences. We chose this threshold because some gestures, e.g., *Follow Me*, produced so many different results that we deemed it reasonable. Repeating the study with a larger group of participants could provide clearer agreements, or would show that there is no recognizable preferred gesture of the target group if there is very little agreement.

[1] It is sheer coincidence that the ratio of male/female and lefthanded/righthanded participants is identical.

Table 1: Distribution of Gestures based on the classification from Obaid et al. [21]

Group	Type	Share of Data Set
Form	dynamic	88 %
	static	12 %
Gesture Type	deictic	73 %
	iconic	4 %
	metaphoric	5 %
	emblematic	19 %
Body Parts	one hand	68 %
	two hands	29 %
	full body	3 %

At the same time, the listing also shows gestures used for different referents by various participants. While the participants consistently avoided double use of gestures within their own definitions, the same gesture could be used for different purposes by different participants. For example, the use of one-handed and two-handed gestures varied between participants, especially for referents like *climb* and *start*. Furthermore, the results show that gesture choices tend to vary less between participants when deictic gestures (e.g., pointing) lend themselves, but differ greatly for referents requiring other types of gestures.

4.2 Defined Voice Commands

Before describing the defined voice commands in detail, we must note that the choice of voice commands might be influenced by the description and the verbal explanation given each participant before asking for suitable voice commands and gestures. Still, some common patterns emerged between the majority of participants, thus we think the voice commands provide some interesting insights. In Table 3, the expressions chosen by the participants[2] are listed and sorted according to their occurrence, accompanied with their share of the set as a percentage. Again, we decided for a threshold of 16.7 % and thus exclude results mentioned only by one participant, but must emphasize that this is a decision for the sake of brevity: in contrast to the defined gestures, we see more unique voice commands from our participants in the data set. The results also show that the participants struggled to find suitable commands for more sophisticated referents, such as *start video recording*. Many participants deliberately tried to simplify such commands to find a match with the short and crisp commands used for more simple referents.

We also observed that a lot of the voice commands resemble the descriptions and explanations we used to describe the referents to our participants (hence the initially mentioned bias), but some variations are visible. Furthermore, as we already noted with the definitions of gestures, some participants used the same voice command for different referents. Although they do not have a high occurrence, such overlaps can lead to problems in choosing the right command and may also yield a confounding experience for some users. Examples are instructing the drone to rise higher while flying and starting up the drone from the ground. In both cases the command *Up* is used to have the drone move upwards.

4.3 Perceived Difficulty of Gesture and Voice Command Definition

After the participants processed a referent, we immediately asked them to subjectively assess the difficulty of finding a proper gesture and voice command for that specific referent. We aimed for finding a metric of how intuitive the chosen gesture or voice command could be for experts in a real application scenario. We assume that if the participants find it particularly easy to identify a gesture or voice command, they can also be recalled quickly when running such an application. Table 4 shows the average difficulty according to the participants' assessment on a 7-level Likert scale ranging from 1 (very difficult) to 7 (very easy). In general, our participants rated the finding of gestures and voice commands in the higher parts of the *easy* side on the Likert scale, yet we can observe a high degree of variations between the defined gestures (leading to low agreement scores, see Section 4.4).

4.4 Agreement Scores

To calculate agreement scores between the participants' choices, we used the approach from E et al. [9]. Table 5 shows the results, split by gesture and voice commands. The results range from 0 to 1; 0 indicating a low agreement (i.e., each participant chose a different gesture or voice command) and 1 a complete match (i.e., all participants decided for the same gesture or voice command). Agreement scores differ notably between all referents. Some have very low agreement scores for both gesture and voice commands, e.g., *Start video recording* (0.06 gesture/0.05 speech) and *Stop video recording* (0.08 gesture/0.08 speech). Others have great variations between gestures and voice commands, most notably *Left* (0.14 gesture/1.00 speech) and *Right* (0.17 gesture/1.00 speech), but also others show this pattern.

4.5 General Remarks

We also asked participants to assess the suitability of using drones within their everyday police work and especially rate the usability of gesture and voice control. The results are based on how participants agreed to a set of statements asking how they perceive the usefulness of gesture and voice controlling drones during operations. They rated all aspects on a 5-level Likert scale ranging from *Do Not Agree* (1) to *Fully Agree* (5). While all participants agreed that drones offer interesting capabilities for police work and also add value to several tasks (average rating 4.8), they also acknowledged that they will become more common in the near future (average rating 4.1), but were more skeptical regarding the interaction with the drones. Overall, our participants preferred to control drones via voice commands. The average rating for using voice commands to handle drones was 3.75, while gesture control was rated at around 3.1. In our data, this difference cannot be correlated or traced back to previous experience. From analyzing the audio recordings, we see that participants mainly doubted the technical feasibility of gesture control and could better imagine voice commands within everyday working life. However, it must be noted that participants did not interact with a drone themselves during our sessions (due to safety reasons, see Section 3.1). Thus, their assessment could change once they have the chance to actively engage with a drone.

[2] Please note that all results have been translated from German to English.

Table 2: Results for Gestures

Referent	Gesture	Percentage	Form	Gesture Type	Body Parts
Up	One hand up	50.00 %	dynamic	deictic	One hand
	Two hands up	25.00 %	dynamic	deictic	Two hands
Down	One hand down	41.70 %	dynamic	deictic	One hand
	Two hands down	25.00 %	dynamic	deictic	Two hands
Left	Pointing left	33.30 %	dynamic	deictic	One hand
	Right hand in front of the body to the left	25.00 %	dynamic	deictic	One hand
Right	Pointing right	41.70 %	dynamic	deictic	One hand
	Left hand in front of the body to the right	16.70 %	dynamic	deictic	One hand
Closer	Draw one hand to the body	66.70 %	dynamic	deictic	One hand
	Draw two hands to the body	25.00 %	dynamic	deictic	Two hands
Further	Pushing away with one hand	41.70 %	dynamic	deictic	One hand
	Pushing away with two hands	25.00 %	dynamic	deictic	Two hands
	Pointing away with one hand	25.00 %	dynamic	deictic	One hand
Takeoff	One hand up	33.30 %	dynamic	deictic	One hand
	Two hands up	16.70 %	dynamic	deictic	Two hands
	Drone launches from platform	16.70 %	dynamic	iconic	Two hands
Landing	One hand down	33.30 %	dynamic	deictic	One hand
	Two hands down	33.30 %	dynamic	deictic	Two hands
	Drone lands on platform	16.70 %	dynamic	iconic	Two hands
Start video recording	Lens gesture	25.00 %	static	iconic	One hand
	Flashlight	16.70 %	dynamic	metaphoric	One hand
Stop video recording	Cross wiping	25.00 %	dynamic	emblematic	Two hands
	Fist clenched	16.70 %	static	emblematic	One hand
	Clap	16.70 %	dynamic	emblematic	Two hands
Fly up	One hand up (from shoulder)	41.70 %	dynamic	deictic	One hand
	Two hands up (from shoulder)	25.00 %	dynamic	deictic	Two hands
	One hand pointing up	16.70 %	static	deictic	One hand
Come back	Draw one hand to the body (from above)	25.00 %	dynamic	deictic	One hand
	Draw two hands to the body (from above)	16.70 %	dynamic	deictic	Two hands
Circle around me	Circular movement above the head	75.00 %	dynamic	deictic	One hand
Follow me	Pump arm (Trucker salute)	16.70 %	dynamic	emblematic	One hand
	Hand on the chest	16.70 %	dynamic	deictic	One hand
	Waving behind	16.70 %	dynamic	deictic	One hand
	Hand knocks on head	16.70 %	dynamic	emblematic	Two hands
Fly towards x	Pointing towards direction	83.30 %	dynamic	deictic	One hand
Come back from direction x	Draw one hand to the body	50.00 %	dynamic	deictic	One hand
	Draw two hands to the body	25.00 %	dynamic	deictic	Two hands
Circle target area	Circling in front of the body pointing to the sky	41.70 %	dynamic	deictic	One hand
	Circling in front of the body pointing to the ground	16.70 %	dynamic	deictic	One hand
	Upright circle in front of the body	16.70 %	dynamic	deictic	One hand
Follow person p	Pointing with one hand	58.30 %	dynamic	deictic	One hand
	Pointing with two hands	16.70 %	dynamic	deictic	Two hands
Rotate clockwise	Circling in front of the body (finger up)	41.70 %	dynamic	deictic	One hand
	Circling in front of the body (finger down)	16.70 %	dynamic	deictic	One hand
Rotate counterclockwise	Circling in front of the body (finger up)	41.70 %	dynamic	deictic	One hand
	Circling in front of the body (finger down)	16.70 %	dynamic	deictic	One hand
Stop	Arms crossed	33.30 %	static	emblematic	Two hands
	Crossing arms apart	16.70 %	dynamic	emblematic	Two hands
	Flat hand in front of the body	16.70 %	static	emblematic	One hand

5 DISCUSSION

We collected user-defined gestures and voice commands for the interaction with drones in police work. We will first derive a set of user-defined gestures and voice commands from the results (Section 5.1). We then discuss some general aspects (Section 5.2) and review potential threats to validity (Section 5.3).

5.1 User-Defined Gesture Set

Figure 3 provides an overview of the gesture set resulting from the study. For each referent, we also list alternative gestures. We removed the referent *Follow me* due to the low level of agreement between our participants (see Section 4.4) in this overview. The captions show the referent and percentage of occurrence of the gesture.

Table 3: Results for Voice Commands (Translated from German to English)

Referent	Voice Command	Percentage
Up	Rise	75.00 %
	Up	16.70 %
Down	Sink	83.30 %
	Down	16.70 %
Left	Left	100.00 %
Right	Right	100.00 %
Closer	Closer	58.30 %
	To me	16.70 %
	Approach	16.70 %
Further	Back	58.30 %
	Further	25.00 %
Takeoff	Takeoff	91.70 %
Landing	Land	83.30 %
Start video recording	Start video	16.70 %
	Start video recording	16.70 %
	Recording	16.70 %
Stop video recording	Stop recording	25.00 %
	Stop video recording	16.70 %
	Stop video	16.70 %

Referent	Voice Command	Percentage
Fly up	Fly up	41.70 %
	Up	25.00 %
	Height x	16.70 %
Come back	Come back	58.30 %
Circle around me	Circle me	83.30 %
Follow me	Follow me	66.70 %
	Follow	33.30 %
Fly towards x	Fly towards x	41.70 %
	Fly to x	16.70 %
Come back from direction x	Come back	83.30 %
Circle target area	Circle target area	50.00 %
	Circle	25.00 %
Follow person p	Follow person	83.30 %
Rotate clockwise	Rotate clockwise	66.70 %
	Turn right	16.70 %
Rotate counterclockwise	Rotate counterclockwise	66.70 %
	Turn left	16.70 %
Stop	Stop	75.00 %

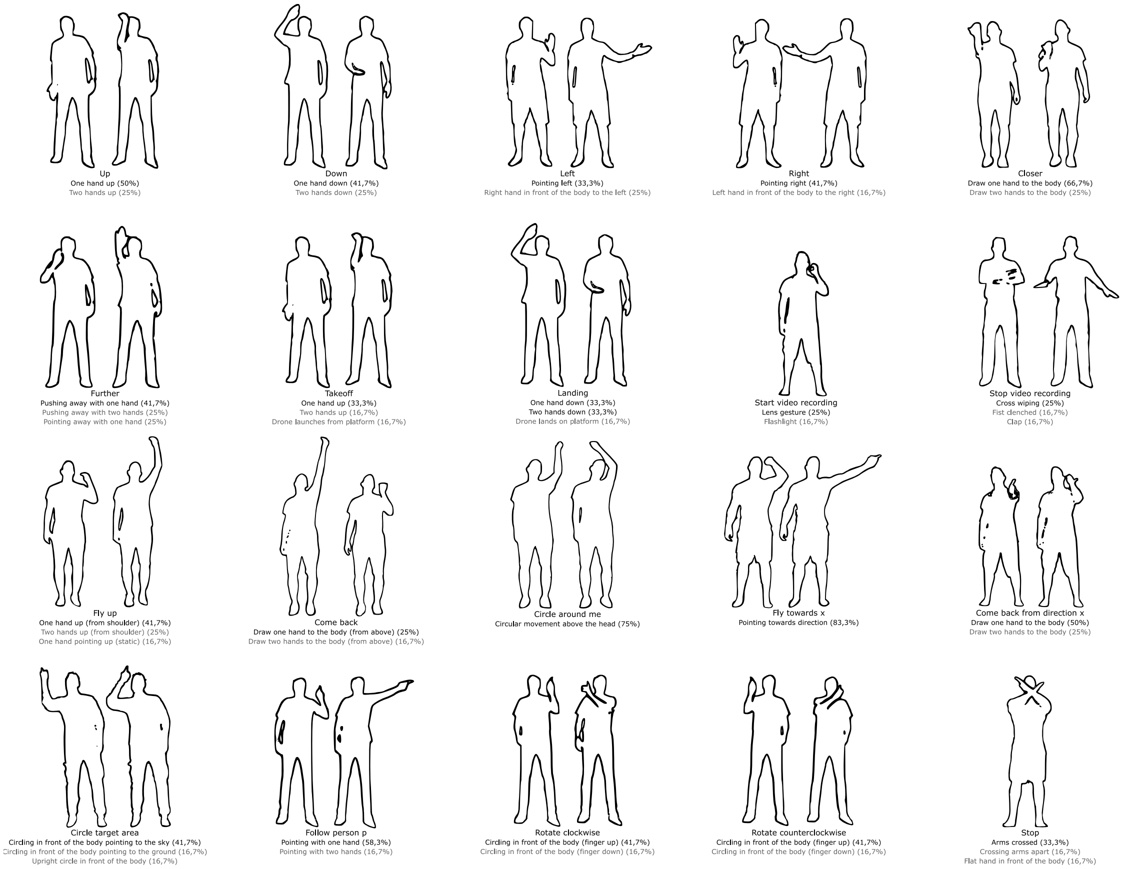

Up
One hand up (50%)
Two hands up (25%)
Down
One hand down (41,7%)
Two hands down (25%)
Left
Pointing left (33,3%)
Right hand in front of the body to the left (25%)
Right
Pointing right (41,7%)
Left hand in front of the body to the right (16,7%)
Closer
Draw one hand to the body (66,7%)
Draw two hands to the body (25%)
Further
Pushing away with one hand (41,7%)
Pushing away with two hands (25%)
Pointing away with one hand (25%)
Takeoff
One hand up (33,3%)
Two hands up (16,7%)
Drone launches from platform (16,7%)
Landing
One hand down (33,3%)
Two hands down (33,3%)
Drone lands on platform (16,7%)
Start video recording
Lens gesture (25%)
Flashlight (16,7%)
Stop video recording
Cross wiping (25%)
Fist clenched (16,7%)
Clap (16,7%)
Fly up
One hand up (from shoulder) (41,7%)
Two hands up (from shoulder) (25%)
One hand pointing up (static) (16,7%)
Come back
Draw one hand to the body (from above) (25%)
Draw two hands to the body (from above) (16,7%)
Circle around me
Circular movement above the head (75%)
Fly towards x
Pointing towards direction (83,3%)
Come back from direction x
Draw one hand to the body (50%)
Draw two hands to the body (25%)
Circle target area
Circling in front of the body pointing to the sky (41,7%)
Circling in front of the body pointing to the ground (16,7%)
Upright circle in front of the body (16,7%)
Follow person p
Pointing with one hand (58,3%)
Pointing with two hands (16,7%)
Rotate clockwise
Circling in front of the body (finger up) (41,7%)
Circling in front of the body (finger down) (16,7%)
Rotate counterclockwise
Circling in front of the body (finger up) (41,7%)
Circling in front of the body (finger down) (16,7%)
Stop
Arms crossed (33,3%)
Crossing arms apart (16,7%)
Flat hand in front of the body (16,7%)

Figure 3: User-Defined Gesture Set

Table 4: Perceived Difficulty

Function	Gestures	Voice Commands
Up	6.42	6.75
Down	6.58	6.75
Left	6.67	6.92
Right	6.67	6.92
Closer	6.08	6.33
Further	5.92	6.25
Takeoff	6.00	6.92
Landing	5.83	6.75
Start video recording	5.42	6.83
Stop video recording	5.17	6.00
Fly up	5.75	5.83
Come back	5.75	5.75
Circle around me	6.08	6.67
Follow me	5.92	6.58
Fly towards x	5.42	5.00
Come back from direction x	5.83	5.75
Circle target area	4.75	6.42
Follow person p	3.83	6.17
Rotate clockwise	5.42	5.83
Rotate counterclockwise	5.00	5.75
Stop	6.83	7.00

Table 5: Agreement Scores

Function	Gestures	Voice Commands
Up	0.27	0.56
Down	0.20	0.70
Left	0.14	1.00
Right	0.17	1.00
Closer	0.47	0.34
Further	0.24	0.36
Takeoff	0.12	0.83
Landing	0.20	0.68
Start video recording	0.06	0.05
Stop video recording	0.08	0.08
Fly up	0.21	0.21
Come back	0.06	0.32
Circle around me	0.55	0.68
Follow me	0.07	0.52
Fly towards x	0.68	0.14
Come back from direction x	0.27	0.68
Circle target area	0.18	0.27
Follow person p	0.33	0.68
Rotate clockwise	0.17	0.44
Rotate counterclockwise	0.17	0.44
Stop	0.12	0.55

Furthermore, Table 6 shows the user-defined set of voice commands. Here, the referent *Start Video Recording* had to be removed due to low agreement scores. Interestingly, agreement was already low for the definition of gestures for this referent and participants predominantly used iconic or metaphoric gestures here.

Table 6: User-Defined Voice Commands Set (Translated from German to English)

Referent	Voice Command	Percentage
Up	Rise	75.00 %
Down	Sink	83.30 %
Left	Left	100.00 %
Right	Right	100.00 %
Closer	Closer	58.30 %
Further	Back	58.30 %
Takeoff	Takeoff	91.70 %
Landing	Land	83.30 %
Stop video recording	Stop recording	25.00 %
Fly up	Fly up	41.70 %
Come back	Come back	58.30 %
Circle around me	Circle me	83.30 %
Follow me	Follow me	66.70 %
Fly towards x	Fly towards x	41.70 %
Come back from direction x	Come back	83.30 %
Circle target area	Circle target area	50.00 %
Follow person p	Follow person	83.30 %
Rotate clockwise	Rotate clockwise	66.70 %
Rotate counterclockwise	Rotate counterclockwise	66.70 %
Stop	Stop	75.00 %

5.2 General Observations

The gestures our participants defined show that not only aspects like cultural differences affect the choice of gestures E et al. [9], but also that the professional background has a strong influence. They reused elements of tactical gestures used for communication between police officers and repurposed them to interact with the drone. Apparently, they could access an existing repertoire of gestures, making their choices easier and giving them an entry point into designing the interaction. Such a gesture set presumably supports learning to interact with the drone quickly, making the interaction more intuitive to experienced officers and thus more accessible for everyday use. While this seems obvious and a good *user-centered* approach on first glance, it puts interaction designers into a dilemma: such gestures can lead to misunderstandings between officers due to the communication ambiguity between human-drone and human-human interaction and thus jeopardize actual operations in the worst case. As a consequence, such applications require a close look into usage scenarios and also an extensive validation. We furthermore argue that such user-defined gesture sets provide a good starting point for designers to gather an understanding of how their users perceive the interaction and their everyday reality, but a dedicated interaction design is still required to resolve such conflicting situations.

Another interesting aspect is the choice of non-deictic gesture types for more sophisticated referents. If a referent cannot be indicated by a deictic gesture, the chosen gestures vary increasingly between participants, showing the broad spectrum of possibilities, which in turn emphasizes the need for careful interaction design. Comparing these results with those from Obaid et al. [21], some similarities emerge, but their resulting gesture set has a strong focus on deictic gestures. The agreement in the occurrence of the individual gestures was also low, which can also be explained by the subdivision into the body areas used. For the referant *sink*, for example, our participants preferred a one-handed gesture, whereas a two-handed gesture predominates in the study by Obaid et al. [21]. Besides this distinction, there are other cases where the results are similar, but not identical.

5.3 Threats to Validity

Our study is subject to some limitations potentially threatening the validity of our results. We will follow the suggestions from Kitchenham et al. [13] and differentiate between general (Section 5.3.1), internal (Section 5.3.2), and external validity (Section 5.3.3).

5.3.1 General Validity. This study was designed for the German-speaking region and the participants were also German-speaking and mainly of German nationality. There are two aspects threatening the generalizability of the study: Cultural differences and our translation of results. It is likely that our results are biased with regard to cultural differences that exist in Human Computer Interaction (HCI) and have already been reported with regards to interaction with drones [9], replication studies involving participants with other cultural background are a potential future work. This aspects must be considered for gestures and voice commands alike, because both might be subjected to cultural biases. For reporting this study at an international venue, we also translated the results for voice commands from German to English and thus cannot exclude any biases coming from our translation.

We furthermore specifically focused on police work, but other professional groups that seem similar (e.g., firefighters, see Alon et al. [4]) might have similar requirements and face similar problems. However, a direct transfer of our results must be tested first with a specific design of scenarios and tasks. This might also be true for other lines of police work.

5.3.2 Internal Validity. During the sessions, we described the drone actions with illustrations, captions, and oral explanations for each command. It is especially unclear whether this approach influenced the participants' voice command choices: we did not require participants explicitly to choose alternative terms, although it was possible.

Furthermore, we had to relocate the study from an outdoor site to an indoor conference room, accompanied most notably with restrictions on a live demonstration of a drone. Although we showed video demonstrations of a drone in action for each task, we cannot exclude that the artificial study setup influenced our participants' experience during the session. While a direct interaction between participant and drone via gestures and speech would not have been possible in the live setting either, we could have imitated the drone's response in a *Wizard-of-Oz*-style experiment, which might have given our participants a deeper insight into how working with a drone feels like, leading to other results.

5.3.3 External Validity. To define scenarios and tasks prior to the study, we cooperated with police officers working in training and educating other officers. A resulting influence, e.g., by using scenarios from the trainers favorite operations, is possible. The same tasks in other scenarios could yield different gestures and voice commands.

Furthermore, some police units work with animals, most notably dogs. We did not specifically take this into account (also, because we had no officer trained in this area available), but it seems possible that such knowledge would yield different results.

6 CONCLUSION AND FUTURE WORK

We studied how police officers would interact with a drone using gestures and voice commands. With 12 police officers participating, we derived a user-defined gesture set and a user-defined voice-command set for 21 functions used in several tasks across two scenarios (public events and patrol).

For future work, several interesting tasks and open questions remain. We aim to convert the user-defined command sets into a real application, which would allow to gather more experience with the requirements for interacting with drones in police work. This also requires a more in-depth analysis of suitable sensors and processing algorithms. We refrained from including multimodal commands in this study, i.e., the combination of gesture and voice commands, due to practical concerns from consulting officers. However, these were mainly related to the chosen scenarios; other scenarios might well benefit from multimodality. Furthermore, other modalities besides speech and gesture exist that were not considered in our study. As an alternative, police officers could be equipped with smartwatches carried on the wrist for interacting with the drone, which would allow for a similar control option like a mobile device, yet at least temporarily free the hands. We also looked an transmitting information from the human operator to the drone but not vice versa. However, police officers would use a drone to gather information of, e.g., their surroundings, thus investigating how such information can be conveyed during operations is an open yet interesting task.

REFERENCES

[1] Ayodeji O Abioye, Stephen D Prior, Glyn T Thomas, Peter Saddington, and Sarvapali D Ramchurn. 2018. The Multimodal Speech and Visual Gesture (mSVG) Control Model for a Practical Patrol, Search, and Rescue Aerobot. In *Towards Autonomous Robotic Systems*, Manuel Giuliani, Tareq Assaf, and Maria Elena Giannaccini (Eds.). Springer International Publishing, Cham, 423–437.

[2] Parastoo Abtahi, David Y. Zhao, Jane L. E., and James A. Landay. 2017. Drone Near Me: Exploring Touch-Based Human-Drone Interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 34 (sep 2017), 8 pages. https://doi.org/10.1145/3130899

[3] Jassim Al-Fadhli, Mustafa Ashkanani, Abdulwahab Yousef, Issam Damaj, and Mohammad El-Shafei. 2014. RECON: A remotely controlled drone for roads safety. In *2014 International Conference on Connected Vehicles and Expo (ICCVE)*. 912–918. https://doi.org/10.1109/ICCVE.2014.7297688

[4] Omri Alon, Sharon Rabinovich, Chana Fyodorov, and Jessica R. Cauchard. 2021. *Drones in Firefighting: A User-Centered Design Perspective.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3447526.3472030

[5] Richard A. Bolt. 1980. "Put-that-there": Voice and Gesture at the Graphics Interface. *ACM SIGGRAPH Computer Graphics* 14, 3 (1980), 262–270. https://doi.org/10.1145/965105.807503

[6] Jake Bruce, Jacob Perron, and Richard Vaughan. 2017. Ready—Aim—Fly! Hands-Free Face-Based HRI for 3D Trajectory Control of UAVs. In *2017 14th Conference on Computer and Robot Vision (CRV)*. 307–313. https://doi.org/10.1109/CRV.2017.39

[7] Jessica R. Cauchard, Jane L. E, Kevin Y. Zhai, and James A. Landay. 2015. Drone & Me: An Exploration into Natural Human-Drone Interaction. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 361–365. https://doi.org/10.1145/2750858.2805823

[8] Jessica R. Cauchard, Alex Tamkin, Cheng Yao Wang, Luke Vink, Michelle Park, Tommy Fang, and James A. Landay. 2019. Drone.io: A Gestural and Visual Interface for Human-Drone Interaction. *ACM/IEEE International Conference on Human-Robot Interaction* 2019-March (2019), 153–162. https://doi.org/10.1109/HRI.2019.8673011

[9] Jane L. E, Ilene L. E, James A. Landay, and Jessica R. Cauchard. 2017. Drone & Wo: Cultural Influences on Human-Drone Interaction Techniques. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6794–6799. https://doi.org/10.1145/3025453.3025755

[10] Bart Engberts and Edo Gillissen. 2016. *Policing from Above: Drone Use by the Police*. Vol. 27. 93–113. https://doi.org/10.1007/978-94-6265-132-6_5

[11] Alexander Freistetter and Karin Anna Hummel. 2019. Human-Drone Teaming: Use Case Bookshelf Inventory. In *Proceedings of the 9th International Conference on the Internet of Things* (Bilbao, Spain) *(IoT 2019)*. Association for Computing Machinery, New York, NY, USA, Article 39, 4 pages. https://doi.org/10.1145/3365871.3365913

[12] DoHyung Kim, Jaeyeon Lee, Ho-Sub Yoon, Jaehong Kim, and Joochan Sohn. 2013. Vision-based arm gesture recognition for a long-range human–robot interaction. *The Journal of Supercomputing* 65, 1 (2013), 336–352. https://doi.org/10.1007/s11227-010-0541-9

[13] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg. 2002. Preliminary Guidelines for Empirical Research in Software Engineering. *IEEE Transactions on Software Engineering* 28, 8 (2002), 721–734. https://doi.org/10.1109/TSE.2002.1027796

[14] Megan Landau and Sebastian van Delden. 2017. 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* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 181–182. https://doi.org/10.1145/3029798.3038329

[15] Xiaoling Lv, Minglu Zhang, and Hui Li. 2008. Robot control based on voice command. In *2008 IEEE International Conference on Automation and Logistics*. 2490–2494. https://doi.org/10.1109/ICAL.2008.4636587

[16] Anna C S Medeiros, Photchara Ratsamee, Yuki Uranishi, Tomohiro Mashita, and Haruo Takemura. 2020. Human-Drone Interaction: Using Pointing Gesture to Define a Target Object. In *Human-Computer Interaction. Multimodal and Natural Interaction*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 688–705.

[17] Mani Monajjemi, Sepehr Mohaimenianpour, and Richard Vaughan. 2016. UAV, come to me: End-to-end, multi-scale situated HRI with an uninstrumented human and a distant UAV. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4410–4417. https://doi.org/10.1109/IROS.2016.7759649

[18] Jawad Nagi, Alessandro Giusti, Gianni A. Di Caro, and Luca M. Gambardella. 2014. Human Control of UAVs using Face Pose Estimates and Hand Gestures. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 1–2.

[19] Jawad Nagi, Alessandro Giusti, Luca M. Gambardella, and Gianni A. Di Caro. 2014. Human-swarm interaction using spatial gestures. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3834–3841. https://doi.org/10.1109/IROS.2014.6943101

[20] Tayyab Naseer, Jürgen Sturm, and Daniel Cremers. 2013. FollowMe: Person following and gesture recognition with a quadrocopter. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 624–630. https://doi.org/10.1109/IROS.2013.6696416

[21] Mohammad Obaid, Felix Kistler, Gabrielė Kasparavičiūtė, Asim Evren Yantaç, and Morten Fjeld. 2016. How Would You Gesture Navigate a Drone? A User-Centered Approach to Control a Drone. In *Proceedings of the 20th International Academic Mindtrek Conference* (Tampere, Finland) *(AcademicMindtrek '16)*. Association for Computing Machinery, New York, NY, USA, 113–121. https://doi.org/10.1145/2994310.2994348

[22] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. *Commun. ACM* 42, 11 (1999), 74–81. https://doi.org/10.1145/319382.319398

[23] Kevin Pfeil, Seng Lee Koh, and Joseph LaViola. 2013. Exploring 3d Gesture Metaphors for Interaction with Unmanned Aerial Vehicles. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 257–266. https://doi.org/10.1145/2449396.2449429

[24] Vasil L. Popov, Kostadin B. Shiev, Andon V. Topalov, Nikola G. Shakev, and Sevil A. Ahmed. 2016. Control of the flight of a small quadrotor using gestural interface. In *2016 IEEE 8th International Conference on Intelligent Systems (IS)*. 622–628. https://doi.org/10.1109/IS.2016.7737492

[25] Siddharth S Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review* 43, 1 (2015), 1–54. https://doi.org/10.1007/s10462-012-9356-9

[26] Carine Rognon, Stefano Mintchev, Fabio Dell'Agnola, Alexandre Cherpillod, David Atienza, and Dario Floreano. 2018. FlyJacket: An Upper Body Soft Exoskeleton for Immersive Drone Control. *IEEE Robotics and Automation Letters* 3, 3 (2018), 2362–2369. https://doi.org/10.1109/LRA.2018.2810955

[27] Matthias Seuter, Eduardo Rodriguez Macrillante, Gernot Bauer, and Christian Kray. 2018. Running with Drones: Desired Services and Control Gestures. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 384–395. https://doi.org/10.1145/3292147.3292156

[28] Ramon A. Suárez Fernández, Jose Luis Sanchez-Lopez, Carlos Sampedro, Hriday Bavle, Martin Molina, and Pascual Campoy. 2016. Natural user interfaces for human-drone multi-modal interaction. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*. 1013–1022. https://doi.org/10.1109/ICUAS.2016.7502665

[29] Florent Taralle, Alexis Paljic, Sotiris Manitsaris, Jordane Grenier, and Christophe Guettier. 2015. A Consensual and Non-Ambiguous Set of Gestures to Interact with UAV in Infantrymen. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 797–803. https://doi.org/10.1145/2702613.2702971

[30] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. *A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?* Association for Computing Machinery, New York, NY, USA, 855–872. https://doi.org/10.1145/3357236.3395511

[31] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1083–1092. https://doi.org/10.1145/1518701.1518866

[32] Jorge Wuth, Pedro Correa, Tomás Núñez, Matías Saavedra, and Néstor Becerra Yoma. 2021. The Role of Speech Technology in User Perception and Context Acquisition in HRI. *International Journal of Social Robotics* 13, 5 (aug 2021), 949–968. https://doi.org/10.1007/s12369-020-00682-5