How Would You Gesture Navigate a Drone? A User-Centered Approach to Control a Drone

Mohammad Obaid[1], Felix Kistler[2], Gabrielė Kasparavičiūtė[4],
Asım Evren Yantaç[3], Morten Fjeld[4]
[1]Department of Information Technology, Uppsala University, Uppsala, Sweden
[2]University of Augsburg, Augsburg, Germany
[3]KUAR, Koç University, Istanbul, Turkey
[4]t2i Lab, Chalmers University of Technology, Gothenburg, Sweden

ABSTRACT

Gestural interaction with flying drones is now on the rise; however, little work has been done to reveal the gestural preferences from users directly. In this paper, we present an elicitation study to help in realizing user-defined gestures for drone navigation. We apply a user-centered approach in which we collected data from 25 participants performing gestural interactions for twelve drone actions of which ten are navigational actions. The analyses of 300 gesture data collected from our participants reveal a user-defined gestural set of possible suitable gestures to control a drone. We report results that can be used by software developers, engineers or designers; and included a taxonomy for the set of user-defined gestures, gestural agreement scores, time performances and subjective ratings for each action. Finally, we discuss the gestural set with implementation insights and conclude with future directions.

CCS Concepts

•**Human-centered computing** → **Gestural input;**

Keywords

Gesture; user-defined; drone; quadcopter; interaction; study.

1. INTRODUCTION

Gestural interactions have been introduced as a way of communication in several domains [15, 6, 28, 22, 7]; and are considered to be a form of a non-verbal communication that is intuitive and fluent. Designing gestural interactions requires the understanding of the users' preferences and needs [21]; conforming to the interaction design cycle. Designing gestural interaction to control drones is no exception, and requires identifying the users' preferences. In the past this has mainly been looked at from the technological point of view, where researchers looked at how tracking technologies and control devices can be used to navigate a drone such as the work presented in [4, 14].

Investigating the control of drones with human gestures opens up exciting possibilities and interesting research challenges. To leverage some of these possibilities, recent projects have approached novel domains including surface computing [29], public displays [10] and mobile interaction [20]. Moreover, research is required to produce validated algorithms capable of recognizing full body gestures and postures, in real time, to tele-operate and guide robots and hence enhance the user's natural experience and engagement with the robot [25, 26]. Typically, algorithms that use body gestures to control robots are based on gesture design paradigms that are defined by their developers. However, if users are not involved in the process, the gestures designed may not be the most intuitive and may not correspond to users' natural behavior.

Recently, Cauchard et. al. [1] depicted a user-centred approach to explore interaction modalities with drones through an elicitation study. Their results showed that gestural interaction was generally the more preferred modality. In addition, the work by Obaid et al. [18] defined a gestural set to navigate a humanoid robot.

This has motivated us to build on the work of Cauchard et. al. [1] to fully investigate with details how users would use their body gestures to navigate a drone located in their field of view in a similar approach to Obaid et al. [18].

The aim is to systematically study what kind of gestures a user can employ to tell the drone to, for example, take off, move left/right, turn left/right (Figure 1), and also advanced operations such as taking pictures or recording video. We follow a user-centered approach to prioritise the needs, wants, and limitations of the user. That is, we examine the design parameters directly related to users' behavior. A body of such parameters is often referred to as gesture taxonomy; and here we study how to establish a user-defined gestural set to navigate a drone. We quantify the agreement scores of the gestural set to ensure that they really correspond to users' preferences.

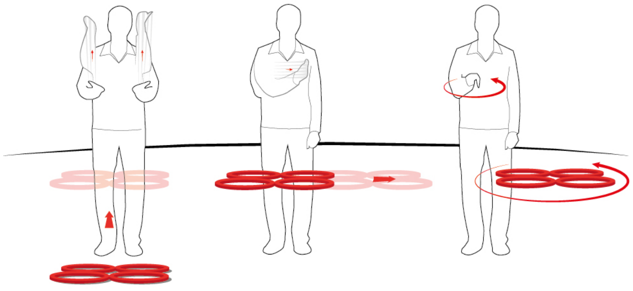

Figure 1: Example illustration of the research motivation

AcademicMindtrek'16, October 17-18, 2016, Tampere, Finland
©2016 ACM. ISBN 978-1-4503-4367-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2994310.2994348

2. RELATED WORK

Investigating and designing novel gestural interactions to control a drone are no exception and require a user-centered approach to understand the user's needs and preferences. Thus, our research is based on three main topics from related literature that include understanding of the human gestural categories, gesture control of drone/robotic platforms, and designing gestural interactions. In the following section we present the related literature and background of our research.

2.1 Human Gestures

In the field of non-verbal behaviours the term 'Kinesics' refers to the study of human gestures and postures. This is a subject that has attracted many researchers and has been studied rigorously over the decades. These studies have revealed several gestural taxonomies that are commonly used. Efron [2] classified gestures into five categories: physiographics, kinetographics, ideographics, deictics, and batons. Another categorization came from Ekman and Friesen [3] who looked at the communication functions to define four categories: emblems, illustrators, regulators and adaptors. Another popular form of presenting gestures was proposed by McNeill [12], who defined five types of gestures: cohesive, beat (emblematic), deictic, iconic, and metaphoric gestures.

In the presented work, we adopt the definition types presented by McNeill due to their popularity [12, 13], and we further explain each of these gestural types. The beat and cohesive gestures are related to rhythmic movement and continue during speech, but do not necessary have a direct meaning. Emblematic gestures are representations that do not have a real-world property, but represent meaning that needs to be learned and is often culture specific. Examples are the thumbs-up sign or a head nod, which either can have positive or negative meanings depending on the cultural background. Deictic gestures refer to pointing at a reference in space, while iconic and metaphoric refer to depicting an action or concept visually; the difference between the two being that iconic gestures are concrete and directly represent the action (such as gesturing if a plate is large or small), while metaphoric gestures help explain an abstract idea (such as the V shape gesture to represent victory). In addition, McNeill defined four phases for each gesture that includes preparation, stroke, hold, and retraction. The preparation is the phase that brings the body from its rest to a position that is suitable for executing the gesture. The stroke phase is the main information contained in the gesture, while the retraction is the phase where the body goes to its rest position again. The hold phase can happen after the stroke phase when the gesture remains in its position an extended time period (such as pointing to the right).

2.2 Drone Control

Generally, controlling drones is a subject that has been studied vastly in the technical domain, where engineers and software developers have looked at different ways of using interaction technologies to control a drone. For example, the classical way of controlling a drone is with traditional control devices such as a joystick controller or a mobile device; for example the Parrot AR. Drone 2.0[1] can be controlled using a mobile device.

Several researchers looked at other ways to control drones using different interaction modalities. For example, Hasan et al. addressed the use of eye gaze to control a drone [5], while Teixeira et al. [27] investigated the teleoperation of a drone using head positions and gestures. Many other interaction techniques have been suggested by researchers such as LaFleur et al. [11] use of a brain user interface to control a drone. Moreover, several have proposed computer vision algorithms, the use of tracking technologies (such as Microsoft Kinect) and other related multimodal techniques to control a drone using body gestures [19, 23, 16].

In summary, most research does not involve the user when designing the gestural interaction, and the user is rather asked to adopt what the engineers and developers have already implemented. In this paper, we contribute by proposing a user-defined gestural set to control a drone based on a user-centered approach. The defined gestures can be used by designers and system engineers to develop a system that is suitable for the users' needs. In the next section, we further address this user-centered approach.

2.3 A User-Centered Approach

Putting the user in the center to define gestural preferences when interacting with technologies has gained momentum in the past years, allowing users to define interactions they find intuitive and easy. Recently, the method used by Wobbrock et al. [29] has inspired many researchers to follow a similar user-centred approach when defining gestures for interactive technologies. The method follows a guessability study that depends on the users' gesture elicitions when executing a specific task with an interface (i.e. surface-top). For example, Obaid et al. [17, 18] have adopted their method to define gestural interactions to navigate a humanoid robot. Wongphati et al. [30] adopted a similar method to investigate gestural characteristics when controlling an end effector of a robotic arm. Kistler et al. [8] identified user-defined full body gestures for an interactive storytelling in a virtual reality scenario. Ruiz at el. [20] investigated a set of user-defined motion gestures when interacting with a smartphone, while Kray at el. [9] and Kurdyukova at el. [10] identified user-defined gestures that relate to communicating between multiple devices, such as mobile-phones, public displays, tablets and tabletops. As one can see, the approach by Wobbrock et al. was used in different domains to understand users' actions when interacting with an interface.

In a similar way, Cauchard et. al. [1] presented a study with 19 participants to explore natural human interactions with a drone to define the modalities users would use when interacting with a drone, in addition to what gestures they would use to control a drone. Their results revealed that the majority of users preferred gestures (86%) and sound (38%) modalities to other interaction techniques. Cauchard et. al. investigated body gestures, sounds, and combinations of these and provided agreement scores for the different modalities per task (including navigational actions). While Cauchard et. al. examined body gestures, sounds, and combinations of these, in our work, we examine body gestures only. The reason for our choice was to narrow in the design space for human-drone navigation gesture design and investigate it further. In addition, Cauchard et. al. relate human-drone interaction to gesture design metaphors, which might help gesture designers to better understand key aspects of those metaphors. In our work we focus on the implementation insights for drone gesture navigation, which can also be useful for the practice of human-drone gesture design.

Thus, we go further and build on the work of Cauchard et. al., and follow a similar approach to Obaid et al. [18], to define a set of gestures to navigate a drone. Our contributions are:

- a user-defined gesture set to navigate a drone,
- a gesture taxonomy of the user-defined set,
- quantitative outcomes (per gesture candidate): agreement scores, ease of use rating, occurrence frequencies, as well as time performance.

[1] http://ardrone2.parrot.com/

- a set of implementation insights for drone gesture navigation, which can be useful for the practice of human-drone gesture design.

3. USER STUDY

The main aim of the study is to investigate user-define gestural interactions to navigate a drone. We address the main navigational commands for a drone that include *move left, move right, move forward, move backward, go up, go down, turn left, turn right, take off*, and *land*. In addition, we add two supporting commands that are actively used with drones which are *take picture* and *record video*. The drone used for the study is named Parrot AR.Drone 2.0, which was controlled through an iPhone 5S application called FreeFlight 2.4.19, which was created by the same company that produced the drone.

We use a Wizard-of-Oz technique to control the drone throughout the study sessions to give the user the impression of the feedback they get when they control the drone and allows them to exibit their actions based on the physical attributes of a flying drone.

3.1 Apparatus

The study was conducted in a space that is part of an open floor that is 5 meters by 10 meters and arranged as shown in Figure 2.

Duct tape was used to mark the starting position for the participant for each session. The starting position of the drone was four meters in front of the participant. The starting points were kept constant throughout the study. The space was equipped with a camera recording the frontal view of the participant and the study space. In addition, a table with a computer screen on it was placed about one meter on the right side and behind the participants. The computer screen was used to play the demonstration videos of the drone actions. The instructor had a position behind the participant to control what was being displayed on the computer screen and to activate the drone's actions during the study; the instructor was not in the field of view of the participant. Finally, a poster was placed on the wall that the participant was facing as a point of reference for some of the actions to be performed.

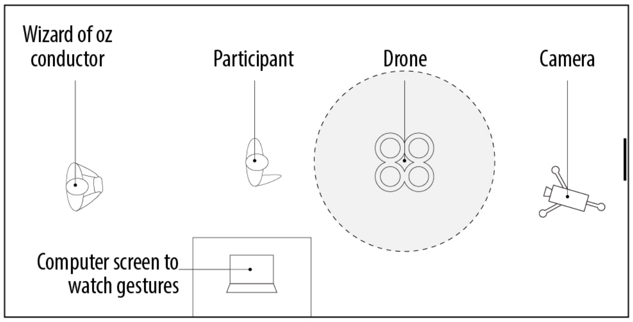

Figure 2: The experimental setup

3.2 Participants

Data was collected from 25 participants, all came from Sweden[2]. These participants (10 female, 15 male) had an average age of 37.6 (SD=12) and most had a background in Interaction Design, Engineering, and Computer Science. Participants were asked to specify whether they were right handed or left handed and to share whether they had information or experience with gestural interaction with a drone or other remote controlled aircrafts. Both scales were a 5-point Likert scales ranging from 1 (no experience) to 5 (very experienced). The 25 participants had a low average experience with gestural interaction = 1.88 and with drones or other remote controlled aircrafts = 1.32. All participants except three were right-handed.

3.3 Procedure

To begin, each participant was given a description of the study and was told that the study was investigating drone control with body gestures. Having read and understood the study, they were asked to sign a consent form agreeing to participate. After this step, the participant stood over a marked region facing the drone and the following steps were carried out: (1) on a computer screen, the participant watched a demonstration video of a drone in action, (2) when the participant said that he/she was ready and understood the action, they were asked to perform a gesture that controlled the drone to behave in similar way to the action presented on the screen, (3) once the gesture was performed by the participant, the wizard (instructor) activated the corresponding actions for the participant to watch, (4) Finally, the participant was asked to rate how easy it was to find/come up with a gesture for that particular action.

Each participant watched the twelve actions presented in random order, of which ten were navigational actions (*move left, move right, move forward, move backward, go up, go down, turn left, turn right, take off*, and *land*) and two were supporting actions (*record video* and *take picture*). For most actions (except *take-off*), the participant was informed that every drone's action would start from a hovering status, which means that the drone would be hovering one meter above ground. Finally, when the gesture and drone action were done, the instructor landed the drone and placed it back to its starting position - making it ready for the next action.

3.4 Measurements

3.4.1 Subjective Measures

After each action the participant had to answer the following 7-point Likert scale question: "How easy was it to think of the right gesture for this action?". The 7-point Likert scale reached from 1 = Extremely hard to 7 = Extremely easy.

3.4.2 Objective Measures

The video recordings of all participants, from a camera videotaping the frontal view of the user, were annotated using the ELAN[3] annotation tools [24]. For each participant video, the twelve actions were annotated and for each action the Stroke phase was highlighted (the gestural phases are described in the Related Work section). In addition, using the annotations of the twelve actions, the time performances of each Stroke phase were extracted and recorded for further analysis.

To make sure annotations were done correctly we did a cross check between two coders, where Coder A did the whole set of videos from the 25 participates and Coder B independently annotated six videos (24% of the data).

4. RESULTS

In the next section, we present the results of our gesture elicitation study, including our gesture taxonomy, the created gesture set, user ratings, and agreement scores.

[2] We considered only one nationality to eliminate any cultural effects on the performed gestures.

[3] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands (https://tla.mpi.nl/tools/tla-tools/elan/).

Table 1: Taxonomy of full body gestures for controlling a drone

Form	static	A static body posture is held after a preparation phase.
	dynamic	The gesture contains movement of one or more body parts during the stroke phase.
Gesture Type	deictic	The gesture is indicating a position or direction.
	iconic	The gesture visually depicts an icon and directly represents a real-world property.
	metaphoric	The gesture visually depicts an icon and describes a real-world property in an abstract way.
	emblematic	The gesture is an artificial symbol that does not represent a real-world property, but represents meaning, which needs to be learned and is often culture specific.
Body Parts	one hand	The gesture is performed with one hand.
	two hands	...with two hands.
	full body	...with at least one other body part than the hands.

4.1 Taxonomy

The videos of the users performing gestures for the twelve actions of the drone were analyzed by extracting the stroke phases of all proposed gestures. The gestures were further classified according to our taxonomy of full body gestures with the three dimensions: *form*, *gesture type*, and *(involved) body parts*. Each dimension consists of multiple items, shown in Table 1. The taxonomy is based on the one presented by Obaid et. al. [18]; however, we modified the gesture type dimension (previously named nature dimension) to be closer oriented to McNeill's [12]. We removed the viewpoint dimension, as we did not observe any "robot-centric" gestures. This is probably due to the fact that the drone has no clear front and back, so the participants did not consider the view-point.

Form distinguishes between static and dynamic gestures (without and with movement respectively). Static gestures have a preparation phase at the beginning, in which the user moves into the gesture space, but the core part of the gesture is after the preparation phase. Therefore, the gesture is kept for a certain amount of time before the user releases it again in the retraction phase. In opposite, dynamic gestures have a clear stroke phase including the movement of body parts between the preparation and retraction phases.

The *gesture type* dimension is oriented according to the taxonomy by McNeill [12]. It uses four of McNeill's categories in the following way: Deictic gestures indicate a position or direction. These gestures can be either static, e.g. pointing to the right, or dynamic, e.g. waving to the right. They can be performed with one hand, two hands, or even other body parts, e.g. tilting the head. We also consider indicating an angular direction as a deictic gesture, e.g. drawing a circle in the air in clock-wise direction and the transverse plane for turning left. Iconic gestures convey information by visually depicting an icon that directly represents a physical, spatial or temporal property of a real-world referent. Especially for full body gestures, iconic gestures are often very direct and a real-world action is described by actually performing it like a pantomime, e.g. stepping left for moving to the left. Metaphoric gestures visually depict an icon as well. However, they describe the real-world property in a more abstract way. An example is performing a grabbing gesture with one hand with the meaning that the drone should take a picture. Emblematic gestures are artificial symbols that do not represent a real-world property, but represent meaning, which needs to be learned and is often culture specific. Therefore, they are not directly related to the meant action, but one needs to have background knowledge, e.g. that snapping the fingers should mean that the drone should take a picture.

The *body parts* dimension should be self-explanatory. It distinguishes between one hand, two hands, and full body gestures that involve at least one other body part.

Figure 3 shows the taxonomy distribution for the 300 gestures. Similar to [18], we found many more dynamic than static gestures, mostly deictics and hand gestures; however, the number of two-hand gestures is almost equal to the number of one-hand gestures.

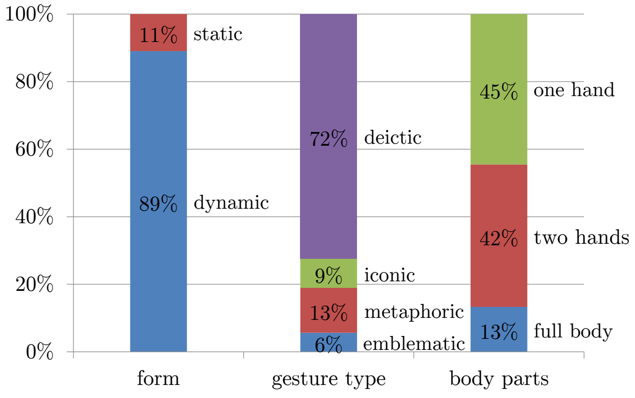

Figure 3: Overall taxonomy distribution

4.2 Gesture Set

To find gesture candidates we adopt and modify the process by Wobbrock et al. [29] by allowing multiple levels of candidates:

- For each system action a, a set $M(a)$ is identified, which contains all proposed gestures.
- The proposed gestures in $M(a)$ are then grouped into subsets of identical gestures $M_i(a)$, with $i \in 1..n_a$ and n_a being the total number of identified subsets for action a.
- The representative gesture candidates $c_j(a)$ for action a are identified by selecting the subsets $M_i(a)$ with the largest sizes, i.e.: $c_j(a) = MAX_{i \in 1..n_a, M_i(a) \neq c_k(a) \forall k<j}(M_i(a))$

Table 2: Gesture Candidates for the twelve actions

Action	Gesture Candidates	Occurrences	Form	Gesture type	Body parts
Move left	Swipe Left	48%	dynamic	deictic	one-hand
Move right	Swipe Right	48%	dynamic	deictic	one-hand
Move forward	Two Hands Push Front Push Front Step Forward	44% 24% 24%	dynamic dynamic dynamic	deictic deictic iconic	two-hands one-hand full-body
Move backward	Two Hands Pull Back Pull Back Step Backward	40% 28% 20%	dynamic dynamic dynamic	deictic deictic iconic	two-hands one-hand full-body
Go up	Two Hands Move Up Move Up	48% 40%	dynamic dynamic	deictic deictic	two-hands one-hand
Go down	Two Hands Move Down Move Down	60% 32%	dynamic dynamic	deictic deictic	two-hands one-hand
Turn left	Draw Circle CCW	42%	dynamic	deictic	one-hand
Turn right	Draw Circle CW	42%	dynamic	deictic	one-hand
Take picture	Click Button Click Button Holding Camera	24% 20%	dynamic dynamic	metaphoric metaphoric	one-hand two-hands
Record video	Clapperboard Hands Clap Hands Crank Camera	17% 13% 13%	dynamic dynamic dynamic	metaphoric emblematic metaphoric	two-hands two-hands two-hands
Take off	Two Hands Move Up	64%	dynamic	deictic	two-hands
Land	Two Hands Move Down	48%	dynamic	deictic	two-hands

As there can be multiple sets $M_i(a)$ of equal size, there can also be multiple gesture candidates $c_j(a)$. As we look at more gesture candidates, we have alternatives in the case the first candidate cannot be used, e.g. for technical reasons. As not all alternative gesture candidates c_j are similarly often represented in the set $M(a)$, we further propose that an alternative candidate should only be taken if its size is not much smaller than the size of the first candidate, i.e. an alternative is only taken into account if its size is at least half the size of the first candidate. The decision of whether two gestures are identical or not can vary depending on the criteria that are important in the specific case. For example, in most cases, it is not important how many times a repetitive gesture has actually been repeated or whether a pointing gesture has been performed with the left or right hand.

Table 2 depicts the gesture candidates for each of the twelve actions, what percentage of the participants proposed each candidate and its taxonomy. Second or third candidates are only included when they occurred at least half as often as the first candidate. Except for *record video*, we almost always got a clear first candidate, only for *go up* and *take picture*, the participants were undecided whether to perform the gesture with one or two hands. Apart from those three actions, *go down*, *move forward* and *move backward* were proposed as additional candidates that could be considered, while the remaining six actions only had one candidate. A detailed visualisation of the gestural candidates is represented in Figure 6.

4.3 User Ratings

Figure 4 shows rated easiness of finding a gesture for the twelve actions on a scale from 1=Extremely hard to 7=Extremely easy.

A one-way repeated measures ANOVA revealed that the ratings differed significantly between the actions with $F(11,24) = 19.03$, $p < 0.001$, $\eta^2 = 0.44$. In particular, the actions *record video* and *take picture* received significantly lower ratings than the other actions, $p_{recordVideo} < 0.001$, $p_{takePicture} < 0.05$ (all Bonferroni corrected), except for the comparison between *take picture* and *turn left*. Further, *take picture* still received a significantly higher rating than Record video, $p < 0.05$ (again Bonferroni corrected).

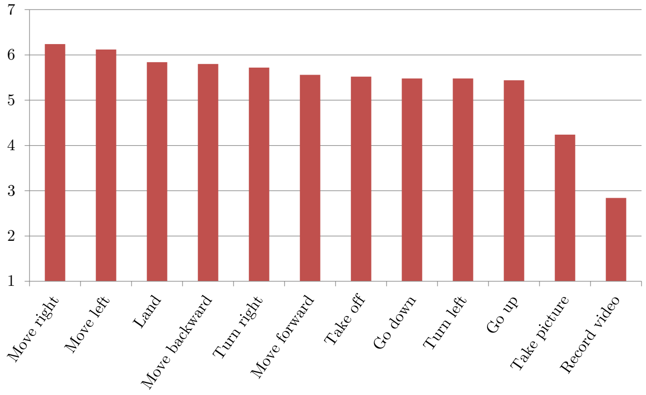

Figure 4: Rated easiness to find a gesture for the twelve actions

4.4 Agreement Scores

To evaluate the degree of agreement among participants with the proposed gestures, we use the formula by Wobbrock et al. [29] to calculate an agreement score $AS(a)$ corresponding to an action a:

$$AS(a) = \sum_{i \in 1..n_a} \left(\frac{|M_i(a)|}{|M(a)|} \right)^2$$

Table 3: The time performance for the Stroke phase of the 12 actions performed by the 25 participants; the times given are the mean and standard deviation (SD) in seconds

	Move left	Move right	Move froward	Move backward	Go up	Go down	Turn Left	Turn right	Take picture	Record video	Take off	Land
Mean	1.96	2.07	1.59	1.66	2.03	1.83	2.37	2.9	1.7	2.2	1.96	1.45
SD	1.4	1.23	0.9	0.89	1.52	1.17	1.36	1.68	3.47	2.54	1.47	0.74

An agreement score $AS(a)$ is therefore represented by a number in the range $[1/|M(a)|, 1]$ with a higher value corresponding to a higher agreement, 1 representing a perfect agreement (all participants chose the same gesture for this action) and $1/|M(a)|$ representing no agreement (all participants chose different gestures for this action). The agreement score can be used as an additional measure for the quality of the gesture candidates. When there is a high agreement, the study participants had a very similar concept on how to represent the action with a gesture, whereas with a low agreement there was no common concept, but the participants really had to be creative to come up with an appropriate gesture for this action.

Figure 5 displays the agreement scores for the twelve actions ordered from high to low agreement. Again, *take picture* and *record video* got the lowest scores. In addition, *turn left* and *turn right* also received slightly lower agreements than most of the other actions, while *go down*, *take off* and *go up* had higher agreements. The mean agreement score was $\overline{AS} = 0.29$. The agreement between the participants further correlates to the rated easiness to find a gesture candidate for the corresponding action (Pearson's $r = 0.598$, $p < 0.05$). When the participants thought it was easy to find a gesture for an action, more same gestures were proposed, and when the participants thought it was difficult to find a gesture for an action, we got a higher variation of gestures as well.

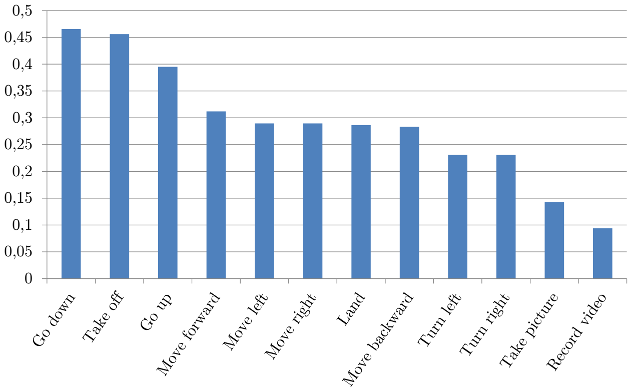

Figure 5: Agreement scores for user-defined gestures

4.5 The Stroke Phase Time Performances

We consider the timing collected from the stroke phase as the most valuable when it comes to the development of a gestural recognition system [18]; thus, we report in Table 3 the average times for the twelve drone actions presented in Table 2.

5. DISCUSSION

The taxonomy distribution when using gestures to control a drone clearly reveals that dynamic gestures is the form that is most preferred (89%), while the overall gesture type was deictic (72%), and both one hand (45%) and two hands (42%) were used in general. These findings directly compare to the findings in Obaid et. al. [18] (Figure 1 (b); Technical, as most of our participants had a technical background), where a very similar gesture form (dynamic; 79.1%), gesture nature (deictic; 77%) and the used body parts (one hand; 54%, and two hands; 33%) were found when navigating a humanoid robot. In addition, all of the ten navigational actions were performed with dynamic deictic gestures and generally came with a high agreement between the users and a higher rating on how easy it was to think of them. This suggests that navigating a drone using body gestures can be associated to interacting with an object that is considered a living being (i.e. when we perceive robots as humanlike [31]). This extends and confirms the feedback comments Cauchard et. al. [1] received from their participants.

On the other hand, the two supporting actions (*take picture* and *record video*) were dynamic and metaphoric/emblematic gestures. Participants found it hard to think of a suitable gesture and had to elaborate further to find a suitable one. It was less agreed on between users due to the nature of those two actions as they are too abstract. It can be suggested that other interaction modalities might be more suitable for actions that are considered metaphoric or emblematic or are performed in that fashion. This suggestion comes from the work by Cauchard et. al. [1] who states that taking a selfie photograph with a drone is preferred with sound (speech input) or both sound/gesture. However, associating more abstract actions with other modalities requires further investigations.

In general, one hand or two hands were used to perform the gestures, however, we found that the use of two hands actually follows symmetry from one hand to another. Thus, when designing for drone gestural interaction, we suggest that to make the gestures set one-handed and account for their symmetry. For example, *record video*, *take off* and *land* were preferred to be done with two hands, however, both hands were doing exactly the same movements. In addition, users preferred two hands for *move forward*/*move backward* but the hands were performing a symmetrical action. In our findings the use of one hand or two hands can suggest to help distinguish between *go up*/*go down* (one handed) and *take off*/*land* (two-handed) for which the first options are the same.

Another aspect that plays an important role in gestural interactions is the 'task' the drone does. In our study, we kept the role of the drone neutral; however, users' feedback reflected on the drone's task. For example, P8 answered "Yes" when asked "Would you have liked to control a drone using a device?", and reflected on their answer saying that they would use "[A] controller, so that [there] is an explicit mapping between action and output", where the action they wanted to perform is "photo/video to make sure that the result is what you expect. Photographing without a view finder is hard"; while P10 said "No" when asked if they want a device to control the drone and reflected on it saying that "it is easier to control by body. Even better for health, we need more of movement (physical)".

Finally, a challenge to consider is environmental factors when flying a drone. We have defined possible candidates as a set of gestures to navigate a drone (Figure 6), those were performed in an indoor setup and gestural interactions might have to be updates in an outdoor setup; the outdoor environment is less controlled and a larger flying range, thus, the users' gestures might be different.

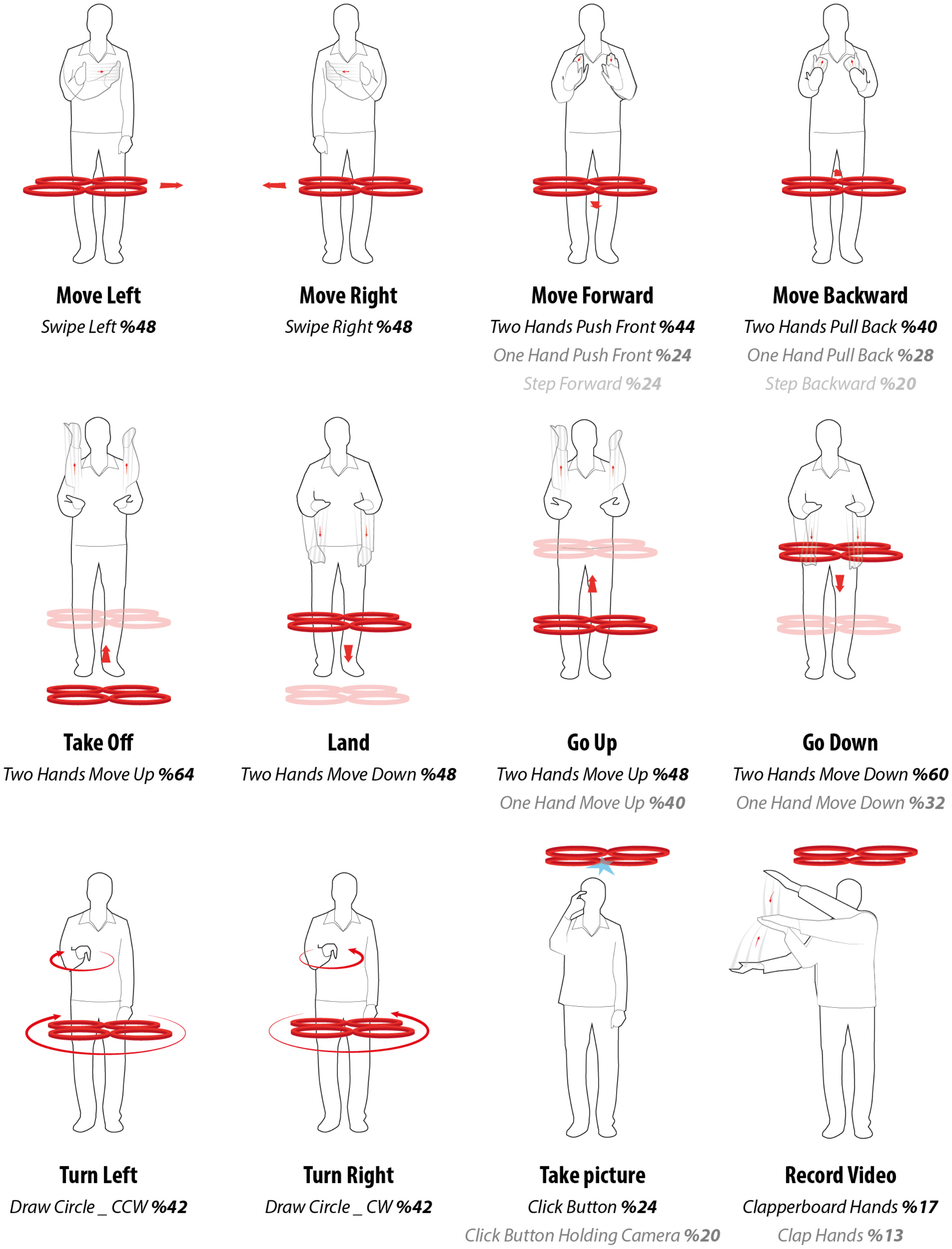

Figure 6: Visual representations of the features for gestural candidates of each action

5.1 Implementation Insights

One might think how can the obtained results be used to implement a recognition system that allows the user to operate a drone. Though we do not describe the implementation aspects of such system as it is outside the scope of this paper but we do give insights to our impressions and experiences for future work.

Our first consideration for a recognition system was to use the opensource Full-Body Interaction Framework (FUBI) that was used by Obaid et. al. [18]. We have setup an AR Drone 2.0 Parrot to listen to commands coming from body motions recognised with FUBI with the use of a Microsoft Kinect device as shown in Figure 7.

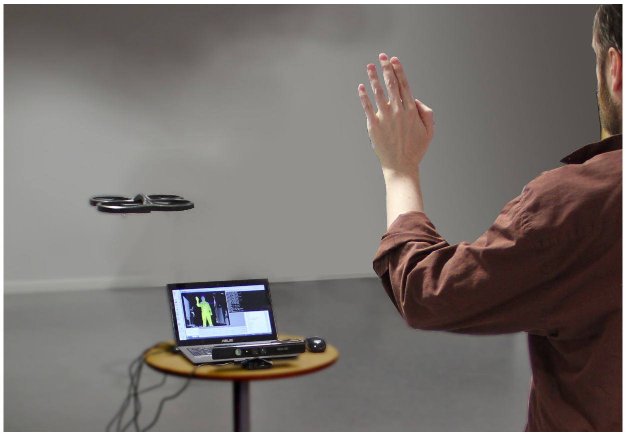

Figure 7: An example of the FUBI setup with a Microsof Kinect to control a drone

However, using Microsoft Kinect does limit the user to be fixed in one position and the user has to be facing towards the device. In a realistic outdoor scenario this might not be feasible when navigating a drone around, as the user needs to be flexible to move around when giving commands to the drone. Thus, we propose to use inertial sensors such as the Xsens Motion trackers[4] to allow for immersing the full body motion into a recognition system that offers higher freedom of movements and a wider tracking range.

We deem it beneficial to further exclusively explore how body gestures can be used in drone navigation; before including other modalities such as human speech. Exploring speech-control for a drone, such as in Cauchard et. al. [1], is indeed interesting, however, using speech-control can comes at a cost, consisting of one or more of having lower classification accuracy, need for reliable speech recognition, drone- and natural speech interference, and potentially low social acceptance of speaking in urban public places.

We use the action *move left* to demonstrate how a developer could use the results obtained in a recognition system. First, the developer should consider how many possible gesture candidates are possible for the command *move left* which can be obtained from Table 2. In this case it is Swipe Left and we can see that it is of a Dynamic form, which means that the gesture is in motion and may also include repetitions. The developer should also consider the body part involved in the motion which in this case it is one-hand. For the *move left* action, we can see that it is of a deictic type, thus it is generally easier to implement as deictic gestures are directional while metaphoric gestures might require an intelligent system to recognise their pattern. Finally, the developer should consider the time performances when setting up a recognition system, where in for the action *move left* the one-hand swipe left motion took on average 1.83 seconds (SD=1.17) as shown in Table 3. The performance details of gesture motions are also important when two actions have candidates of a similar form and style, such as *go down* and *land*. We can see from Table 2 that the two actions share the gesture candidate Two Hands Move Down, however, from Table 3 we can see that the gesture was performed faster in the action *land*.

The results obtained are limited to the point-of-view of the user, where all actions are considered while the drone was facing in the same direction as the user. We anticipate that the gestural set might have to be updated if the drone was considered with multiple view points (not just the frontal point view of the user).

We finish off the section with a scenario depicting how a user might apply the gestural actions to control a drone. In a typical scenario, the user would stand facing the drone, which he would lift from the ground by moving both hands up. The drone takes off and moves up to the user's waist level. In order to move the drone away from him, he pushes both hands to the front. He keeps repeating the same gesture until the desired position is reached. When the drone reaches the aimed for position, the user starts swiping his left hand repeatedly to right which then moves the drone to the right. At this point, the user makes a counter-clockwise movement with his left hand to turn the drone around. Finally he makes a click button gesture to take a picture of him. Finishing this real action, the user raises both of his hands and pulls both hands repeatedly until the drone gets closer to him. The user ends the task by moving two hands down to land the drone on the ground.

6. CONCLUSION AND FUTURE WORK

We have presented a study to investigate user-defined gestures to control a drone. Twenty five participates took part in the study where they were asked to watch a drone motion and elicit how they would want to gesture control the drone in a similar way. In the study, participants watched twelve actions of which ten were navigational actions and two were supporting actions. The participants answered a questionnaire to reveal how easy it was to think of corresponding control gestures. Our results reported on the analysis of 300 gestures that resulted top twelve gestures based on agreement scores among the elicited gestures. We presented the user-defined gestural set to control a drone, along with agreement scores for each action. We constructed a gesture taxonomy of the user-defined set and outlined the time performances for each of the gestural motions. The study presented useful details on gestural interaction for controlling a drone, which can be used by system designers or engineers to develop technologies to enable its functionality.

Future directions of the presented work aim to provide an expanded gestural set when considering the user-drone interactions from multiple angles. We also plan to investigate a more universal gesture set across different cultures, as the current set is defined with users from a single cultural background (Swedish). Finally, the actions chosen were based on navigational actions in a neutral task; however, future drones will be investigated with task driven actions (such as drones for sports, rescue, and security).

7. REFERENCES

[1] Jessica R. Cauchard, Jane L. E, Kevin Y. Zhai, and James A. Landay. Drone & me: An exploration into natural human-drone interaction. In *the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 361–365, New York, NY, USA, 2015. ACM.

[2] David Efron. *Gesture and Environment*. King's Crown Press, Morningside Heights, New York, 1941.

[4] https://www.xsens.com

[3] Paul Ekman and Wallace Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.

[4] Jakob Engel, Jürgen Sturm, and Daniel Cremers. Camera-based navigation of a low-cost quadrocopter. In *the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2815–2821, Oct 2012.

[5] John Paulin Hansen, Alexandre Alapetite, I. Scott MacKenzie, and Emilie Møllenbach. The use of gaze to control drones. In *the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 27–34, New York, NY, USA, 2014. ACM.

[6] Eleanor Jones, Jason Alexander, Andreas Andreou, Pourang Irani, and Sriram Subramanian. Gestext: Accelerometer-based gestural text-entry systems. In *the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2173–2182, New York, NY, USA, 2010. ACM.

[7] Maria Karam and M.C. Schraefel. A taxonomy of gestures in human computer interactions. University of Southampton, 2005.

[8] Felix Kistler and Elisabeth André. User-defined body gestures for an interactive storytelling scenario. In *Human-Computer Interaction - INTERACT 2013*, volume 8118 of *Lecture Notes in Computer Science*, pages 264–281. Springer Berlin Heidelberg, 2013.

[9] Christian Kray, Daniel Nesbitt, John Dawson, and Michael Rohs. User-defined gestures for connecting mobile phones, public displays, and tabletops. In *the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '10, pages 239–248, New York, NY, USA, 2010. ACM.

[10] Ekaterina Kurdyukova, Matthias Redlin, and Elisabeth André. Studying user-defined ipad gestures for interaction in multi-display environment. In *International Conference on Intelligent User Interfaces*, pages 1–6, 2012.

[11] Karl LaFleur, Kaitlin Cassady, Alexander Doud, Kaleb Shades, Eitan Rogin, and Bin He. Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface. *Journal of Neural Engineering*, 10(4):046003, 2013.

[12] David McNeill. So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371, 1985.

[13] David McNeill. *Head and Mind: What Gestures Reveal About Thought*. University of Chicago University of Chicago Press, 1992.

[14] Jawad Nagi, Alessandro Giusti, Gianni A. Di Caro, and Luca M. Gambardella. Human control of uavs using face pose estimates and hand gestures. In *the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 252–253, New York, NY, USA, 2014. ACM.

[15] Jamie Ng, Tze-Jan Sim, Yao-Sheng Foo, and Vanessa Yeo. Gesture-based interaction with virtual 3d objects on large display: What makes it fun? In *the SIGCHI Conference on Human Factors in Computing Systems*, CHI EA '09, pages 3751–3756, New York, NY, USA, 2009. ACM.

[16] Wai Shan Ng and Ehud Sharlin. Collocated interaction with flying robots. In *20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2011)*, pages 143–149, July 2011.

[17] Mohammad Obaid, Markus Häring, Felix Kistler, René Bühling, and Elisabeth André. User-defined body gestures for navigational control of a humanoid robot. In *the 4th International Conference on Social Robotics*, volume 7621 of *Lecture Notes in Computer Science*, pages 367–377. Springer Berlin Heidelberg, 2012.

[18] Mohammad Obaid, Felix Kistler, Markus Häring, René Bühling, and Elisabeth André. A framework for user-defined body gestures to control a humanoid robot. *International Journal of Social Robotics*, 6(3):383–396, 2014.

[19] Kevin Pfeil, Seng Lee Koh, and Joseph LaViola. Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles. In *the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 257–266, New York, NY, USA, 2013. ACM.

[20] Jaime Ruiz, Yang Li, and Edward Lank. User-defined motion gestures for mobile interaction. In *the 2011 annual Conference on Human Factors in Computing Systems*, CHI '11, pages 197–206, New York, NY, USA, 2011. ACM.

[21] Dan Saffer. *Designing Gestural Interfaces*. O'Reilly Media, Sebastopol, 2009.

[22] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[23] Andrea Sanna, Fabrizio Lamberti, Gianluca Paravati, and Federico Manuri. A kinect-based natural interface for quadrotor control. *Entertainment Computing*, 4(3):179 – 186, 2013.

[24] Han Sloetjes and Peter Wittenburg. Annotation by category: Elan and iso dcr. In *the 6th International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), 2008.

[25] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In *the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004).*, volume 3, pages 2422 – 2427 vol.3, 2004.

[26] Halit Bener Suay and Sonia Chernova. Humanoid robot control using depth camera. In *the 6th International Conference on Human-Robot Interaction*, HRI '11, pages 401–402, New York, NY, USA, 2011. ACM.

[27] Joao M. Teixeira, Ronaldo Ferreira, Matheus Santos, and Veronica Teichrieb. Teleoperation using google glass and ar, drone for structural inspection. In *the 2014 XVI Symposium on Virtual and Augmented Reality*, pages 28–36, May 2014.

[28] Radu-Daniel Vatavu. User-defined gestures for free-hand tv control. In *the 10th European Conference on Interactive TV and Video*, EuroiTV '12, pages 45–48, New York, NY, USA, 2012. ACM.

[29] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. User-defined gestures for surface computing. In *the SIGCHI Conference on Human Factors in Computing Systems*, pages 1083–1092. ACM, 2009.

[30] Mahisorn Wongphati, Hirotaka Osawa, and Michita Imai. User-defined gestures for controlling primitive motions of an end effector. *Advanced Robotics*, 29(4):225–238, 2015.

[31] Jakub Zlotowski, Ewald Strasser, and Christoph Bartneck. Dimensions of anthropomorphism: From humanness to humanlikeness. In *the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, pages 66–73, New York, NY, USA, 2014. ACM.