Exploring User-Defined Gestures and Voice Commands to Control an Unmanned Aerial Vehicle

Ekaterina Peshkova(✉), Martin Hitz, and David Ahlström

Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
{ekaterina.peshkova,martin.hitz,
david.ahlstroem}@aau.at

Abstract. In this paper we follow a participatory design approach to explore what novice users find to be intuitive ways to control an Unmanned Aerial Vehicle (UAV). We gather users' suggestions for suitable voice and gesture commands through an online survey and a video interview and we also record the voice commands and gestures used by participants' in a Wizard of Oz experiment where participants thought they were manoeuvring a UAV. We identify commonalities in the data collected from the three elicitation methods and assemble a collection of voice and gesture command sets for navigating a UAV. Furthermore, to obtain a deeper understanding of why our participants chose the gestures and voice commands they did, we analyse and discuss the collected data in terms of mental models and identify three prevailing classes of mental models that likely guided many of our participants in their choice of voice and gesture commands.

1 Introduction

Over the past decade, researchers have shown increased interest in developing interfaces to interact with a UAV (Unmanned Aerial Vehicle) that is partially explained by recent availability of low-cost UAVs (e.g., Parrot AR.Drone), which became affordable to a larger range of researchers. Much of the current works investigated novel input modalities in order to provide more natural ways of interaction such as speech [1], gestures (including head [2–4], hand [5–8], and upper body movements [9, 10]), face detection [11], gaze direction [12, 13], and even brain activity [14].

When designing a human-machine interface, the way users imagine interaction with the system has to be carefully investigated as it is of crucial importance for the intuitiveness of the resultant interaction vocabulary. Accordingly, a good understanding of users' mental models is believed to be a key aspect to consider in a user-centered design. A number of authors have explored users' mental models with the aim to improve interface usability [15] and to design intuitive interfaces, both for human-computer interaction [16, 17] and human-robot interaction [18, 19]. Regarding interaction with a UAV, to our best knowledge, our work is the first attempt to investigate users' truly 'natural' behavior (i.e., unguided and without instructions and a set of predefined and imposed commands) with their associated mental models when navigating a single UAV.

R. Poppe et al. (Eds.): INTETAIN 2016, LNICST 178, pp. 47–62, 2017.
DOI: 10.1007/978-3-319-49616-0_5

The goal of our study is to create a collection of user-defined gestures and voice commands to control a UAV and to gain insights regarding any possible underlying mental models. Accordingly, we let users define their own input (voice and gesture) vocabulary for ten main commands to operate a drone: *takeoff*, *land*, *up*, *down*, *left*, *right*, *rotate left*, *rotate right*, *forward*, and *backward*. We target our study towards an interface for users with no previous experiences in navigating UAVs and explore the behavior of novice users in order to understand what gestures and voice commands are intuitive for users whose mental models are not influenced by previous knowledge and experiences with UAVs.

Our exploration of user-defined commands to navigate a UAV is based on three data sources: (1) an online survey where we collected suggestions for voice commands, (2) a video interview where we recorded suggested gestures, and (3) a Wizard of Oz experiment where we observed and recorded participants' behavior and spontaneous voice and gesture command choices while piloting a drone (a skilled drone pilot, the 'wizard', secretly interpreted participants' commands and operated the drone accordingly using the standard touchscreen-based interface). With our study, we aim to address the following questions:

- What are intuitive gestures and voice commands to navigate a UAV?
- Are there commonalities in users' behavior and command choices?
- What mental models do novice users have regarding the navigation of a UAV?
- Do novice users rely on one coherent mental model when they navigate a UAV or do they rely on concepts and notions drawn from different mental models?

The main contribution of our work is a collection of user-defined gestures and voice commands that can serve as a source of inspiration and guidance for future research and projects on interaction with unmanned agents (Unmanned Ground Vehicles or Unmanned Underwater Vehicles). We also provide insights regarding the mental models novice users rely on when navigating a drone. Although our exploration is focused on commands for a UAV, we anticipate that the obtained results could easily be extended to the more general case of human-robot interaction and related navigation tasks. Finally, we also see a minor contribution in our methodological approach by demonstrating the strengths of combining three techniques to elicit user-defined voice commands and gestures to allow interaction designers to come up with close to optimal natural user interfaces.

Next, we provide a brief overview of related work and then describe the used materials and procedures of the online survey, the video interview, and the Wizard of Oz experiment. After that we report our findings, discuss mental models in the context of UAV control, and then analyse our findings in terms of mental models. We conclude with a summary and an outline of promising directions for future work.

2 Related Work

Over at least two past decades, most research has investigated unconventional input modalities such as speech, gestures, gaze direction, and facial expressions to interact with a robot [20]. It is a widely held view that the use of natural cues peculiar to human-human interaction could contribute to the development of a more natural human-robot interaction. In some sense, much human-robot interaction research and work on interfaces for UAV control, including ours, are guided by the motto of designing robots or vehicles capable of perceiving and interpreting human behavior, not the reverse. Previous work has also demonstrated strong advantages of interfaces that allow the user to interact through multiple "natural" input modalities, such as body movements and speech. We review inspiring related work on multimodal control interfaces and work on speech and gesture interaction.

In early work, Hauptmann and McAvinney [21] investigated the use of gestures and speech to rotate, translate, and scale a 3D cube on a screen. They reported that the study participants preferred to use a combination of gestures and speech rather than a single modality. Similarly, Sharma et al. [22] pointed out that multimodal interaction promotes natural conversation and presented several arguments for multimodality. Among them was an argument that a fusion of multiple senses is inherent in human nature. In particular, the combination of speech and gestures has received close attention [23, 24]. In the context of navigation, Jones et al. [25] explored the use of speech and gestures to control the flight of a group of UAVs and found that many study participants found it intuitive to interact with the system using both speech and gestures.

Quigley et al. [26] compared several user interfaces that employ different modalities to interact with a UAV. These were: *physical interfaces* (an altitude joystick, a physical icon controller or "phicon" – a real object used to interact with a system, and an altitude TrackPoint$^{\mathrm{TM}}$), *direct manipulation* (PDA- and laptop-based interfaces), *voice-based interface*, and an interface with *numerical parameter entry*. The study found that users prefer simple intuitive interfaces with a lower precision level rather than numerical parameter-based interfaces that require entering exact flight data. The main reasons were that high-precision interfaces result in unreasonably high workload and that the provided precision would not be needed in most situations. The study also showed that a future user interface should require reasonable levels of mental and physical demands and leave out unnecessary complications that impede the operators' work. Quigley et al. stressed the effectiveness of voice-based control when a UAV is within an operator's field of view, as the operator could fully focus visual attention on the operated UAV instead of continuously switching visual focus between the UAV and the input controls.

In more recent years, numerous projects on gesture-based interaction, both with a single UAV and a group of UAVs, have been presented. These projects have compared gesture control to standard touchscreen-based interfaces [9] and joystick input [2]. Pfeil et al. [9] explored five different gesture sets to pilot a UAV (e.g., a user spreads the arms to the sides and navigates the UAV by bending and rotating the upper body; a user holds an imagined UAV and its movements are mapped to movements of the actual UAV) along a predefined path. Pfeil et al. reported that ten of their fourteen participants preferred gestures over the touchscreen-based interface and that six participants rated the touchscreen-based interface as the least fun technique to use. Higuchi and Rekimoto [2] evaluated and compared their gesture-based interface to the use of joystick input. The suggested head-gesture interface synchronized the position and orientation of the operator's head with those of a UAV. In the study, participants had to control a UAV and take pictures of both stationary and moving objects with the on-board camera. A better performance, in terms of time required to complete the tasks and in terms of the accuracy of taken pictures (i.e., how close photographed objects were to the center of a picture), was achieved with the gesture-based interface. In addition, post-questionnaires revealed that participants found it simpler to control the UAV with the head-gestures than with the joystick.

Most of the suggested gesture-based interaction techniques were defined by designers [3, 5, 6, 8, 10–12, 27] and only in some of them the vocabulary was further evaluated by users [2, 4, 7, 9]. Jones et al. [25] investigated users' natural behavior using speech and gestures to control the flight of a group of UAVs in their Wizard of Oz study conducted in a simulated environment. In a single UAV case, there are two works where users were let to define the vocabulary. Burke and Lasenby [28] provided five persons with verbal descriptions of UAV's responses and recorded the suggested gestures, and Cauchard et al. [29] used a Wizard of Oz study to explore users' spontaneous command suggestions. None of the two works provided a list of collected user-defined commands. However, Cauchard et al. reported that most of the nineteen participants used a "human to human-like" interaction style. This finding makes sense considering that the participants did knew about the human operator who was interpreting and translating their commands. However, this knowledge might also have influenced participants' behavior and command choices. Nevertheless, the study by Cauchard et al. clearly demonstrates the feasibility of exploring users' natural behavior and intuitive gesture and voice commands for UAV interaction.

While the review above shows a great deal of work on natural interaction with UAVs, the main novelty brought by our work consists in the identification of mental models [30] for UAV navigation and their association with natural interaction gestures and voice commands, especially for novice users.

3 Data Collection

For our exploration we picked the eight basic motion commands that are necessary to manoeuvre a UAV: *up*, *down*, *left*, *right*, *rotate left*, *rotate right*, *forward*, and *backward*. We also included the *takeoff* command and the *land* command. Each of these two commands is a combination of a functional command, *turn on* respectively *turn off* the rotors, and a motion command, *up* respectively *down* (until a certain height is reached or until the ground is touched). We collected gesture and voice command suggestions for the ten commands from 110 persons: 50 persons (aged between 25 and 68 years (mean 37, s.d. 12), 19 female) participated in an online survey, 27 persons (aged between 21 to 52 years (mean 31, s.d. 8), 12 female) participated in a video interview, and 33 persons (aged between 21 and 62 years (mean 30, s.d. 9), 11 female) participated in a Wizard of Oz session. Only five experiment participants and twelve interview participants indicated having previous (childhood or more recent) experiences in operating remote-controlled toys such as cars, boats, or drones. Among the participants there were no experienced UAV pilots. Survey respondents were not questioned regarding their previous experiences. Five persons participated in both the experiment and in the interview.

The online survey was open during five weeks and consisted of ten questions (excluding instructions and demographic questions), one for each of the ten commands. Each question showed one image to illustrate the outcome of one of the ten commands and the participant was asked to provide one word, or a short phrase, that he/she thought would be a suitable voice command for the depicted outcome. The used images are shown in Fig. 1(a to j).

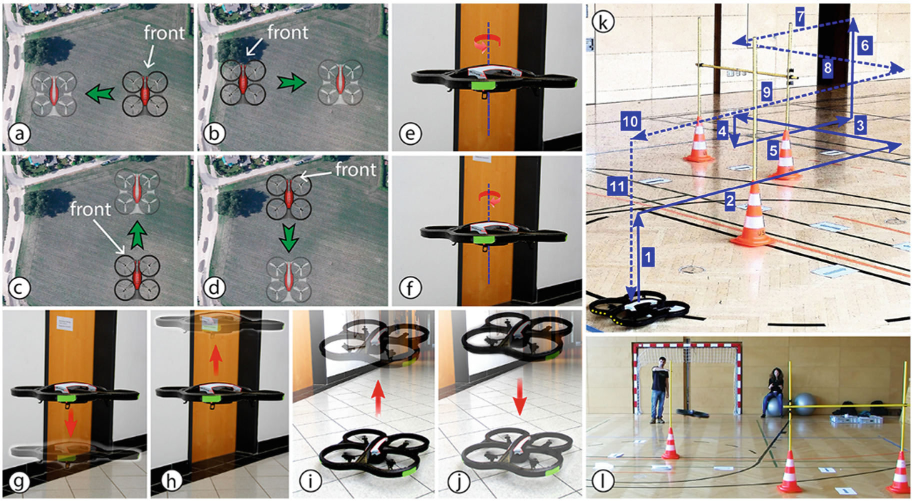


Fig. 1. Materials used in the survey (a–j) and experiment (k and l). Survey images to illustrate commands: (a) *left*, (b) *right*, (c) *forward*, (d) *backward*, (e) *rotate left*, (f) *rotate right*, (g) *down*, (h) *up*, (i) *takeoff*, and (j) *land*. (k) Obstacle path. (l) Experiment setup with participant (left) and wizard (right).

In the video interview participants watched ten short video clips (2–6 s long, one for each command) of a drone performing each of the ten commands (e.g., a drone *taking-off* from the ground, a drone turning *left* in the air). After having watched a clip the participant was asked to suggest a gesture that would be suitable to command the drone to perform the shown action. Participants' gestures were captured on video for later analysis and demographic data were collected through a short questionnaire.

The experiment was conducted in a gymnasium. The participant's task was to use voice and gesture commands to navigate a UAV (Parrot AR.Drone 2.0) along a pre-defined path that was indicated by arrows on the floor, as shown in Fig. 1k and l. To complete the path all of the ten commands under study had to be used. First, the participant had to command the drone to *takeoff* (Fig. 1k, segment 1) from a fixed start position. After takeoff, the drone had to fly *forward* (Fig. 1k, segment 2) to the right of the first vertical pole and then fly *left* (Fig. 1k, segment 3) without changing its orientation. When the drone was approximately in the middle between the two most distant vertical poles, the participant had to command the drone to fly *down* (Fig. 1k, segment 4) until it could pass below the horizontal pole that connected the vertical poles.

After having flown below the horizontal pole (Fig. 1k, segment 5) the drone had to fly *up* (Fig. 1k, segment 6) higher than the horizontal pole and then fly *backward* (Fig. 1k, segment 7) above the pole. Once the horizontal pole was passed, it had to fly sideways to the *right* (Fig. 1k, segment 8). Next, the drone had to continue *backwards* (Fig. 1k, segment 9) until it was approximately above its initial position. On reaching the initial position (Fig. 1k, position 10), the participant had to *rotate* the drone by 180° (either *rotate left* or *rotate right*) so that its front faced the participant. After that, the drone had to be rotated back in the opposite direction by 180°. Finally, the drone had to *land* at the start position (Fig. 1k, segment 11).

The participant was told that the drone could be manoeuvred through both voice commands and gestures and that the purpose of the task was to verify the accuracy of the voice and gesture recognizers. Two video cameras on tripods facing the participant (Fig. 1l was taken by one of the cameras) served both as visual props to make voice and gesture detection seem realistic and to record the participant's voice and gesture commands for later analyses. The 'wizard' who secretly interpreted and translated the participant's commands and manoeuvred the drone using a tablet was sitting behind the participants (Fig. 1l). The participant was told that the 'wizard's' role was to take notes about the recognizers' performance during the task. The navigation task required a constant attention of the participants on the operated drone due to its dynamic nature. One experimenter was standing next to the participants to clarify potential doubts of the participants. In that way, the participants did not have time and need to look at the "wizard".

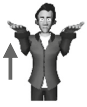
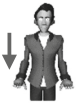
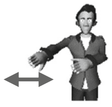
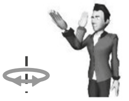
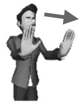
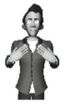
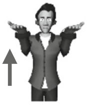
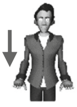
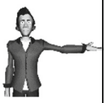
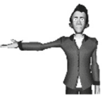
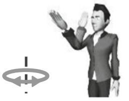
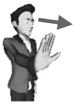
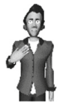
In order to avoid giving away hints about possible voice and gesture commands while explaining the task and the required path, the experimenter showed the path by carrying the drone along the path. After that, the participant did the same to confirm that he/she had understood the required path. The participant was asked to stand during the task and to remain within a 2 × 2 m large square that was marked on the floor.

4 Results

We divide our presentation of the results in three parts: we first report on results regarding voice commands (obtained from the survey and from the experiment). Next, we present our findings regarding gestures (obtained from the experiment and from the interview), and finally we discuss how our participants combined voice and gesture commands during the experiment. The results are summarised in Table 1[1]. The table lists the most frequently suggested and used voice commands and gestures. The numbers in the table represent the rounded relative frequency (%) with which the listed words and gestures occurred among all word suggestions, respectively used gestures, for the specified command (columns). Numbers in italics (row 4–6) show the number of experiment participants who used a voice command (*Voice*), a gesture (*Gesture*), or a combination (*V + G*) for the specified command.

[1] Please contact the authors for the full collection of voice and gesture commands.

Table 1. The most frequently used/suggested voice and gesture commands.

<table>
<tr><th></th><th></th><th colspan="2"><i>Up &amp; Down</i></th><th colspan="2"><i>Left &amp; Right</i></th><th colspan="2"><i>Rotate Left & Rotate Right</i></th><th colspan="2"><i>Forward & Backward</i></th><th colspan="2"><i>Takeoff &amp; Land</i></th></tr>
<tr><td><b>Survey</b></td><td><b>Voice</b></td><td colspan="2">up (81)
upward (8)
down (83)
downward (4)</td><td colspan="2">left (82)
west (6)
right (84)
east (6)</td><td colspan="2">turn left (30)
rotate left (26)
rotate counter clockwise (17)
turn right (31)
rotate right (25)
rotate clockwise (17)</td><td colspan="2">forward (58)
go (13)
straight (8)
north (6)
back/backward (42)
reverse/south (6)</td><td colspan="2">up (28)
takeoff (21)
lift/start (17)
land (56)
down (23)</td></tr>
<tr><td rowspan="2"><b>Experiment</b></td><td><b>Voice</b></td><td colspan="2">up (88)
higher (12)
down (100)</td><td colspan="2">left (80)
turn left (12)
right (92)
turn right (8)</td><td colspan="2">turn (48)
turn left (12)
turn (50)
turn right (12)</td><td colspan="2">forward (56)
straight (32)
back (76)
backward (16)</td><td colspan="2">up (42)
start (32)
land (40)
stop (33)
down (13)</td></tr>
<tr><td><b>Gestures</b></td><td>
31</td><td>
58</td><td colspan="2">
24 (left) and 26 (right)</td><td colspan="2">
35</td><td>
23</td><td>
35</td><td colspan="2">–</td></tr>
<tr><td colspan="2"><i>Voice</i></td><td><i>7</i></td><td><i>9</i></td><td><i>8</i></td><td><i>10</i></td><td><i>10</i></td><td><i>9</i></td><td><i>7</i></td><td><i>7</i></td><td><i>9</i></td><td><i>15</i></td></tr>
<tr><td colspan="2"><i>Gesture</i></td><td><i>9</i></td><td><i>10</i></td><td><i>9</i></td><td><i>9</i></td><td><i>8</i></td><td><i>9</i></td><td><i>9</i></td><td><i>9</i></td><td><i>2</i></td><td><i>3</i></td></tr>
<tr><td colspan="2"><i>V+G</i></td><td><i>17</i></td><td><i>14</i></td><td><i>16</i></td><td><i>14</i></td><td><i>15</i></td><td><i>15</i></td><td><i>17</i></td><td><i>17</i></td><td><i>22</i></td><td><i>15</i></td></tr>
<tr><td><b>Interview</b></td><td><b>Gestures</b></td><td>
33</td><td>
37</td><td>
30</td><td>
30</td><td colspan="2">
22</td><td>
22</td><td>
37</td><td colspan="2">–</td></tr>
</table>

4.1 Voice Commands

The first row in Table 1 lists the most frequently suggested voice commands in the survey. The second row lists the most common voice commands that were used among the experiment participants who either used a voice command or a combination of speech and a gesture for the corresponding command.

In the study we observed many commonalities between the voice commands suggested in the survey and those being used in the experiment. The majority of all suggestions for the *up* and *down* commands in the survey were ‘up’ (81 %) and ‘down’ (83 %). The second most frequent words were ‘upward’ (8 %) respectively ‘downward’ (4 %). Also in the experiment, ‘up’ was the most frequent (88 %) voice command used to manoeuvre the drone *up* among the seven participants who used a voice command. In the experiment, ‘higher’ was the second most frequent (12 %) voice command for *up*. All nine participants who used a voice command for *down* used the word ‘down’.

In the survey, the most frequent suggestions to command a UAV to fly *left* or *right* were the words ‘left’ (82 %), respectively ‘right’ (84 %). When a voice command was used for flying *left* or *right* in the experiment (8 resp. 10 participants), the mostly used words were ‘left’ (80 %) respectively ‘right’ (92 %).

The short phrases ‘turn left’ (30 %), ‘rotate left’ (26 %), and ‘rotate counter clockwise’ (17 %) were the most frequent suggestions in the survey for the *rotate left* command. For the *rotate right* command, an almost identical distribution was observed between the phrases ‘turn right’ (31 %), ‘rotate right’ (25 %), and ‘rotate clockwise’ (17 %). In the experiment, nine participants used a voice command for *rotate left* and for *rotate right.* Nearly half of the participants (48 %) used either ‘turn’ or ‘turn around’ in a combination with a gesture to specify the direction to rotate. ‘Turn left’ and ‘turn right’ were used by 12 % of the participants.

Both in the survey and in the experiment, more than half (58 %) of the participants used ‘forward’ to command the UAV to fly *forward.* ‘Go’ and ‘straight’ were other frequently used commands provided in the survey and in the experiment.

In the survey, the same number of the respondents suggested ‘back’ (42 %) and ‘backward’ (42 %) to command a UAV to fly backward. In the experiment, a majority of the participants who used a voice command preferred the shorter version ‘back’ (76 %) instead of ‘backward’ (16 %). This difference indicates the practical use of the experiment that allowed us to observe natural behavior of novice users when navigation a real vehicle in a concrete scenario.

In the survey, the suggestions for the *takeoff* command were divided between ‘up’ (28 %), ‘takeoff’ (21 %), ‘lift’ (17 %), and ‘start’ (17 %). In the experiment, the most frequently used voice commands for *takeoff* were ‘up’ (42 %) and ‘start’ (32 %). The survey respondents and the experiment participants frequently suggested, and used, ‘land’ (56 % resp. 40 %) and ‘down’ (23 % resp. 13 %) to command the UAV to *land*. In the experiment, ‘stop’ was also the second most frequently (33 %) used voice command to *land* the drone.

Interestingly, the respondents and the participants used ‘up’ both for the *up* and *takeoff* commands and ‘down’ for the *down* and *land* commands. This observation suggests that there was a tendency to neglect the functional difference of the *takeoff* and *land* commands.

Overall, the strongest agreement on voice commands was for the *up*, *down*, *left*, and *right* commands. For the *forward* and *backward* commands, the suggestions were mainly divided between two options. There were other interesting ideas such as ‘let’s go*’* for the *takeoff* command; ‘park’ and ‘that’s it’ to command *land*; a phrase ‘ready, steady, go’ was used for the first forward command.

In the experiment, over half of the participants (58 %) used ‘stop’ to command the UAV to stop. In addition, the participants kept interacting with the UAV using ‘go on’, ‘again’, ‘(a bit) more’, ‘(it’s) ok’, ‘yes’, and ‘no’ to continue or discourage the current movement of the UAV. Two survey participants and four participants in the experiment specified attributes for commands such as ‘1 m up’ or ‘180° counter clockwise’.

4.2 Gestures

Row 3 and 7 in Table 1 show the most frequently observed gesture for each of the eight motion commands (*up*, *down*, *left*, *right*, *rotate left*, *rotate right*, *forward*, and *backward*) during the experiment and in the video interview, respectively. Overall, we observed a variety of gestures, including small finger gestures, such as pointing with a finger, and full-body gestures, such as taking a step towards in a desired direction. However, hand gestures were the most common. The majority (52 %) of interview participants and the majority (62 %) of the experiment participants who used hand gestures tended to consistently use either one or both hands for all the commands; the others mixed one-handed and two-handed gestures. In case of two-handed gestures, it is interesting to note that most of the time the involvement of the second hand was redundant as it simply duplicated the movements of the other. Furthermore, we did not observe any overlaps between the gestures used for the eight motion commands.

In order to command a UAV to fly up, 33 % of the interviewees and 31 % of the participants showed the following gesture: arms bent at the elbows in front of the person point upwards with palms facing the ceiling. The same number of the interviewees moved either one (37 %) or two hands (37 %) down with palms facing the floor. In the experiment, 68 % of those who showed gestures used both hands for the down command and the remaining 32 % used one hand.

The most frequently observed gestures for *left* and *right* in the experiment were to put out both hands in front of the torso and then move the hands to the left, respectively to the right. Most interview participants (30 %) suggested the same gestures, but used only one hand, the left for the *left* command and the right hand for the *right* command.

The most popular gestures, both in the experiment and in the video interview, for the *rotate left* and *rotate right* commands were to put out the dominant hand, bend the elbow, and then circle the hand/forearm counter clockwise or counter clockwise (experiment 35 %, interview 22 %).

In the experiment, the most frequently used gesture (23 %) for the *forward* command among participants who used a gesture or combined a gesture with speech was to extended both arms in front of the upper chest and "pushed" the hands away. The same gesture, but with only one hand, was also the most popular (22 %) gesture for *forward* in the video interview. For the opposite direction, the *backward* command, experiment participants most frequently (35 %) used both hands and "waved" towards themselves. Interview participants used only one hand (37 %).

4.3 Combination of Voice Commands and Gestures

Our analysis of the video recordings during the experiment revealed that 11 participants (33 %) mostly (at least for 8 out of 10 commands) used a combination of voice and gesture commands, six participants (18 %) mostly used a gesture only (at least for 8 out of 10 ten commands), and seven participants (21 %) mostly used voice commands only. For each command, the usage of voice commands only, gestures only, and a combination of both modalities was approximately the same: 25 %, 25 %, and 50 %, respectively. The exceptions were the *takeoff* (27 %, 6 %, and 67 %) and *land* (45 %, 9 %, and 45 %) commands, for which only few participants used gestures only. The reason might be the functional difference: while the other eight commands suggest only a certain movement of the UAV, the *takeoff* and *land* commands involve an additional function: turning on and off the rotors, which is possibly more easily communicated through a voice command.

In a combination with voice commands, gestures often served to explicitly indicate the direction, e.g., ‘go up’ accompanied with a hand pointing upwards. The exceptions were the *rotate left* and *rotate right* commands for which 45 % of the participants used gestures to specify the direction e.g., ‘rotate’ accompanied with a gesture indicating the direction of rotation.

4.4 The Takeoff and Land Commands

Gestures used for the *takeoff* and *land* commands are left out from Table 1. The reason is their sameness with the gestures shown for the *up* and *down* commands, respectively.

In the experiment, the wizards were asked to hesitate with taking off and landing the UAV to give the participants time to figure out gestures different from those used for the *up* and *down* commands. For most participants, in the end, after the ineffective insisting on using the *up/down* gesture to command the UAV to takeoff/land, most of the participants used a voice command such as ‘start’ and ‘stop’. The fact that the participants used voice commands instead of inventing a separate gesture for the *takeoff* and *land* commands suggests that speech fits more naturally for the given commands.

Those who participated in the video interview were asked to suggest only gestures and, as a result, several variations of *land* gestures different from *down* gestures were observed. Almost one fifth of the participants (22 %) separated the *land* and *down* commands either by using a totally different gesture, as visualized in Fig. 2a, or by augmenting the down gesture (Fig. 2b and c). The gestures shown in Figs. 2b and c were used both individually and in a combination with the most frequently used down gesture (Table 1). However, these gestures give an impression of being more “forced” rather than being natural and intuitive since their mental models are less obvious.

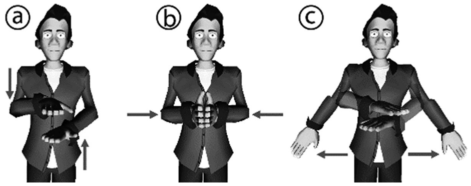


Fig. 2. Gestures used to distinguish the land from the down command: (a) vertical clap, (b) a horizontal clap, (c) crossed hands move to the sides.

5 Discussion

We set out this study to provide insight into intuitive interaction using voice and gestures to operate a UAV. The most frequently used voice commands and gestures listed in Table 1 are apparently the most intuitive ones for novice users and the answer to our question “What are intuitive gestures and voice commands to navigate a UAV?” We now present and discuss our main findings that answer the remaining three questions:

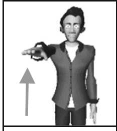
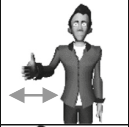
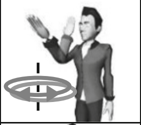
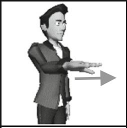
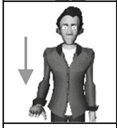
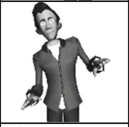
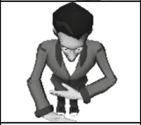
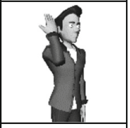
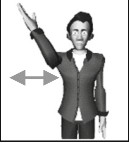
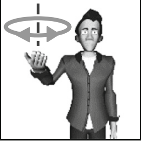
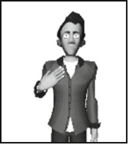
5.1 Commonalities in Users' Behavior and Command Choice

The analysis of the observed gestures and voice commands within each part of the study individually showed that less diversity of voice commands and gestures was observed in the experiment compared to the survey and the interview, respectively. With regard to the survey, this finding may partly be explained by the number of the respondents that is higher than the number of those participated in the experiment (50 vs. 33). As for the interview, though the number of the participants was slightly lower than in the experiment (27 vs. 33), the same tendency was detected. There is, however, another possible reason to which we are inclined. The observed diversity is likely to be related to a number of "degrees of freedom". Survey and interview participants had full discretion to suggest any relevant idea. Whereas participants in the experiment were also encouraged to use any voice command and/or gesture, we cannot exclude that experiment participants tried to give commands that they thought could be more easily recognized by the pretended voice and gesture recognizer or that they tried to guess and use the commands chosen by the developer. This finding suggests that the survey and interviews are mutually complementary with the experiment as the limitation of one is compensated by the strong side of the other. Accordingly, we emphasize the importance of considering all three sources of data to make the picture complete.

In terms of gestures, the data obtained from the interview and the experiment revealed that the absolute majority of the participants indicated the direction to fly with their hands and specified the direction to rotate by rotating the forearm. Likewise, there were many commonalities among the collected voice commands. These findings suggest that, in the considered context, there is indeed a potential in the development of gestures and voice commands that are intuitive for novice users.

Overall, the obtained gesture set (Table 1) included from 9 to 17 options for each of the eight commands. However, if we neglect the orientation of the palms and the cases where the second hand was used explicitly, meaning that a command would have been

Table 2. Final gesture set, suitable for all participants.

Up & Down	*Left & Right*	*Rotate Left & Rotate Right*	*Forward & Backward*
			
			
			

clear even if only one hand was involved, then an interface able to correctly interpret the set of gestures presented in Table 2 would allow all our experiment participants to navigate a UAV without any initial instructions.

5.2 Mental Models

To answer our third and fourth questions “What mental models do novice users have regarding the navigation of a UAV?” and “Do novice users rely on one coherent mental model when they navigate a UAV or do they rely on concepts and notions drawn from different mental models?” we first briefly explain what we mean with ‘mental models’. The concept of mental models originates from cognitive science and its introduction is most often attributed to Craik [30], more than seventy years ago. Since then it has been widely used within the field of Human-Computer Interaction [31] to reason about, and to understand, human behavior. However, less HRI-related work can be found on mental models. A mental model about the functionality and behavior of a system is formed mainly by previously acquired knowledge (e.g., from documentation or instructions) and experiences with similar systems. Moreover, a system is often considered to be intuitive to use if the way the system works matches the user’s expectations, which in turn are dictated by his/her mental model of how the system works. Accordingly, and since we are interested in designing intuitive interactions using gestures to manoeuvring a UAV, we tried to identify any patterns regarding mental models in participants’ behavior.

Of course, one could argue that it is unlikely that an interface designer can design an interface that meets the expectations of all the potential users. While it is true that it is hardly possible to predict the users’ behavior, mental models might help to guide a user to the desired behavior. In many cases, a hint can be enough to guide a user to a certain mental model. The following example illustrates the key idea behind the concept. For example, a user is asked to guide a UAV by imitating the desired actions using hand movements, as illustrated in Fig. 3. Following this scenario, the user could intuitively control the flight of the UAV without further instructions simply by adhering to the mental model “my hand represents the UAV” (imitative class of mental models in our classification; cf. below). Therefore, the use of mental models that define a behavior that is intuitive for an individual or a group of users under a certain scenario could help in stimulating the desired behavior. Accordingly, we aimed to identify user-defined mental models, which seemed to guide each participant’s behavior, in order to come up with intuitive interaction techniques.

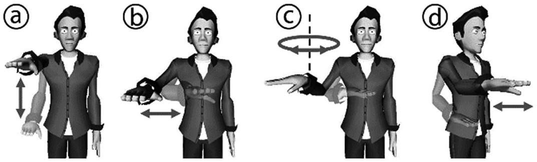


Fig. 3. Gestures associated with the imitative class of mental models with the hand imitating the required commands: (a) *up/down*; (b) *left/right*; (c) *rotate left/rotate right*; (d) *forward/backward*.

Our analysis of mental models in terms of their commonalities and differences led to clustering of related mental models into three classes: *imitative*, *intelligent,* and *instrumented*. In the *imitative* class, a part of the operator's body e.g., a hand, as in Fig. 3, serves as a surrogate of the UAV and thus movements of this body part are directly mapped to movements of the UAV. Gestures include those where the UAV follows the motions of the head, one hand, two hands, the upper-body, or the whole body.

In the *intelligent* class, an operator expects a certain level of intelligence of a UAV that enables the UAV to interpret a given command correctly. Gestures include those where an indication of the direction is given with the index finger, thumb, hand, forearm, arm, and "come to me" or "go away" gestures.

In the *instrumented* class, an operator gives the flight instructions using an imaginary tool. This class is represented through four mental models. We call these the 'virtual UAV', the 'puppet ruler', the 'joystick', and the 'super power'. With the 'virtual UAV' the operator holds in the hands a virtual UAV whose movements are mapped directly to the movements of the real UAV. With the 'puppet ruler' the operator holds an imaginary UAV right in front of the body that is 'connected' with the UAV through two invisible 'strings', the real UAV copies the movements of the 'puppet' UAV. With the 'joystick' the operator associates the forearm with a joystick and tilts the forearm as if it was a joystick: forward, backward, left or right to command a UAV to go forward, backward, left, and right, respectively. Finally, with the 'super power' the operator keeps the arms in front of the body at chest-height with the palms facing forward and 'pushes' or 'pulls' the UAV in the desired direction, as if having a magic super power.

Most of our participants seem to have preferred an *intelligent* type of mental model and seem to have expected that the UAV could correctly interpret all the given commands, including high-level commands such as the "come to me". A reason behind this preference might be the strong resemblance to human-to-human interaction.

Another important finding is the tendency participants had to stick to one and the same mental model while navigating the UAV. Particularly, almost a quarter (24 %) of all participants who used gestures in the experiment used at least six gestures (out of 8) that were associated with one mental model and more than one third (38 %) of the participants used at least four gestures that were associated with one mental model. In the survey, almost one fifth (19 %) of the interview participants suggested gestures associated with one mental model and the majority (63 %) suggested at least five gestures belonging to the one and same mental model. These observations underline the importance of considering mental models to define a coherent gesture set.

A possible reason for using more mental models could be the inability of the 'main' model to cover all commands. For example, tilting the forearm as if it was a joystick could be intuitively associated with the *forward*, *backward*, *left*, and *right* commands, however, the rest of the studied commands are not that intuitively mapped to any such forearm movement. Another reason could involve aspects related to physical ergonomics. For example, imitation of UAV movements with one hand with the palm facing downwards can naturally cover all the studied commands except for the *rotate left* and *rotate right* as rotation of the hand at the wrist about vertical axis is not physically comfortable (Fig. 3c).

The gesture vocabulary composed of the most frequently used gestures (Table 1) does not necessarily warrant the coherence of its components. For instance, each gesture considered individually seems to make sense. However, the gesture set on the whole seems to be inconsistent. In particular, the *up* gesture looks like if the operator holds a virtual UAV in the hands. The *down* gesture can be seen as the imitation of the required UAV's motion with both hands. An operator simply indicates the direction to fly with a fully extended arm when commanding a UAV to fly left and right as well as the direction to rotate by rotating the forearm. The *forward* gesture looks like if the operator has the 'super force' to push the UAV forward. The *backward* gesture resembles the "come to me" gesture. Although the gestures might be interpreted differently, it is worth mentioning that it is not possible to give an operator only one single hint that uniquely defines the set of gestures presented in Table 1 and that allows the operator to navigate the UAV without further instructions. In addition, it seems illogical to use both hands to command a UAV to fly up and down while only one hand is used for the remaining commands. In such a case, an operator would have to learn by heart when to use one hand and when to use both hands. This inconsistency makes the gesture set shown in Table 1 harder to remember and might lead to a confusion of gestures. For the set of commands studied, this issue does not seem to cause serious problems as the vocabulary is small and could be easily learned. However, in cases where a larger gesture vocabulary is necessary, such inconsistencies might cause confusion and notably increase the user's mental workload. In order to avoid such problems, we recommend avoiding a mixture of mental models, as far as possible. It is our strong believe that an interaction technique composed of gestures belonging to one mental model would lead to a lower mental workload.

6 Conclusion

The present exploration used three sources (an online survey, a Wizard of Oz experiment, and a video interview) to elicit *intuitive* gestures and voice commands for controlling a UAV from novice users. With our approach we could identify: (1) the most intuitive gestures and voice commands; (2) important commonalities in users' behavior and command choices; and (3) various mental models that guided our novice users in their voice and gesture choices. With our exploration, we also provided evidence that indicates that novice users tend to rely on concepts and notions drawn from one mental model rather than drawn from different mental models.

The main outcome of our exploration is a collection of user-defined voice commands and gestures to manoeuvre a UAV. Although the focus of the exploration was on UAVs, the proposed collection of gestures and voice commands, as well as, the underlying methodology (using several methods to elicit what users' find intuitive) might serve as a source of inspiration for other researchers and interface designers in their development of natural and intuitive interactions for a broader range of unmanned agents, including Unmanned Ground Vehicles and Unmanned Underwater Vehicles.

For future work we plan to compare the use of the user-defined gestures and voice commands to other interaction modalities. Further studies are also needed to investigate the influence of mental models on various usability-related interface aspects, such as learnability, memorability, ergonomics, satisfaction, and intuitiveness. In addition, it would be interesting to compare experiences of individuals within the same task but with gesture sets associated with different mental models.

References

1. Supimros, S., Wongthanavasu, S.: Speech recognition - based control system for drone. In: Proceedings of the 3rd ICT International Student Project Conference (ICT-ISPC 2014), pp. 107–110 (2014)
2. Higuchi, K., Rekimoto, J.: Flying head: a head motion synchronization mechanism for unmanned aerial vehicle control. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 2029–2038 (2013)
3. Teixeira, J.M., Ferreira, R., Santos, M., Teichrieb, V.: Teleoperation using google glass and AR.Drone for structural inspection. In: Proceedings of the XVI Symposium on Virtual and Augmented Reality (SVR 2014), pp. 28–36 (2014)
4. Pittman, C., LaViola, J.J.: Exploring head tracked head mounted displays for first person robot teleoperation. In: Proceedings of the 2014 ACM International Conference on Intelligent User Interfaces (IUI 2014), pp. 323–328 (2014)
5. Mashood, A., Noura, H., Jawhar, I., Mohamed, N.: A gesture based kinect for quadrotor control. In: Proceedings of the 2015 International Conference on Information and Communication Technology Research (ICTRC 2015), pp. 298–301 (2015)
6. Naseer, T., Sturm, J., Cremers, D.: FollowMe: person following and gesture recognition with a quadrocopter. In: Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013), pp. 624–630 (2013)
7. Ng, W.S., Sharlin, E.: Collocated interaction with flying robots. In: Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2011), pp. 143–149 (2011)
8. Soto-Guerrero, D., Ramírez Torres, J.G.: A human-machine interface with unmanned aerial vehicles. In: Proceedings of the 10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2013), pp. 307–312 (2013)
9. Pfeil, K., Koh, S.L., LaViola, J.: Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI 2013), pp. 257–266 (2013)
10. Ikeuchi, K., Otsuka, T., Yoshii, A., Sakamoto, M., Nakajima, T.: KinecDrone: enhancing somatic sensation to fly in the sky with kinect and AR.Drone. In: Proceedings of the 5th Augmented Human International Conference (AH 2014), no. 53 (2014)
11. Nagi, J., Giusti, A., Gambardella, L.M., Di Caro, G.A.: Human-swarm interaction using spatial gestures. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014), pp. 3834–3841 (2014)
12. Monajjemi, V., Wawerla, J., Vaughan, R., Mori, G.: HRI in the sky: creating and commanding teams of UAVs with a vision-mediated gestural interface. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013), pp. 617–623 (2013)
13. Hansen, J.P., Alapetite, A., MacKenzie, I.S., Møllenbach, E.: The use of gaze to control drones. In: Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA 2014), pp. 27–34 (2014)

14. Kos'myna, N., Tarpin-Bernard, F., Rivet, B.: Bidirectional feedback in motor imagery BCIs: learn to control a drone within 5 minutes. In: CHI 2014 Extended Abstracts on Human Factors in Computing Systems, pp. 479–482 (2014)
15. Davidson, M.J., Dove, L., Weltz, J.: Mental Models and Usability (1999)
16. Rust, K., Malu, M., Anthony, L., Findlater, L.K.: Understanding child-defined gestures and children's mental models for touchscreen tabletop interaction. In: Proceedings of the 2014 Conference on Interaction Design and Children (IDC 2014), pp. 201–204 (2014)
17. Valdes, C., Eastman, D., Grote, C., Thatte, S., Shaer, O., Mazalek, A., Ullmer, B., Konkel, M.K.: Exploring the design space of gestural interaction with active tokens through user-defined gestures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2014), pp. 4107–4116 (2014)
18. Kiesler, S., Goetz, J.: Mental models of robotic assistants. In: CHI 2002 Extended Abstracts on Human Factors in Computing Systems, pp. 576–577 (2002)
19. Powers, A., Kiesler, S.: The advisor robot: tracing people's mental model from a robot's physical attributes. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI 2006), pp. 218–225 (2006)
20. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robot. Auton. Syst. **42**(3–4), 143–166 (2003)
21. Hauptmann, A., McAvinney, P.: Gestures with speech for graphic manipulation. Int. J. Man-Mach. Stud. **38**(2), 231–249 (1993)
22. Sharma, R., Pavlović, V.I., Huang, T.: Toward multimodal human-computer interface. Proc. IEEE **86**(5), 853–869 (1998)
23. Rogalla, O., Ehrenmann, M., Zöllner, R., Becher, R., Dillmann, R.: Using gesture and speech control for commanding a robot assistant. In: Proceedings of the 11th International Workshop on Robot and Human Interactive Communication (RO-MAN 2002), pp. 454–459 (2002)
24. Urban, M., Bajcsy, P.: Fusion of voice, gesture, and human-computer controls for remotely operated robot. In: Proceedings of the 7th International Conference on Information Fusion (FUSION 2005), pp. 1644–1651 (2005)
25. Jones, G., Berthouze, N., Bielski, R., Julier, S.: Towards a situated, multimodal interface for multiple UAV control. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2010), pp. 1739–1744 (2010)
26. Quigley, M., Goodrich, M.A., Beard, R.W.: Semi-autonomous human-UAV interfaces for fixed-wing mini-UAVs. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pp. 2457–2462 (2004)
27. Lichtenstern, M., Frassl, M., Perun, B., Angermann, M.: A prototyping environment for interaction between a human and a robotic multi-agent system. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012), pp. 185–186 (2012)
28. Burke, M., Lasenby, J.: Pantomimic gestures for human–robot interaction. IEEE Trans. Robot. **31**(5), 1225–1237 (2015)
29. Cauchard, J.R., E, J.L., Zhai, K.Y., Landay, J.A.: Drone & me: an exploration into natural human-drone interaction. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2015), pp. 361–365 (2015)
30. Craik, K.: The Nature of Explanation. Cambridge University Press, Cambridge (1943)
31. Norman, D.A.: The Design of Everyday Things. Basic Books, Inc., New York (2002)