On the Ability to Provide Demonstrations on a UAS: Observing 90 Untrained Participants Abusing a Flying Robot

Mitchell Scott
School of MME
Washington State University
mitchell.scott@email.wsu.edu

Bei Peng
School of EECS
Washington State University
bei.peng@wsu.edu

Madeline Chili
Dept. of CS
Elon University
mchili@elon.edu

Tanay Nigam
School of EECS
Washington State University
tanay.nigam@wsu.edu

Francis Pascual
Mathematics and Statistics
Washington State University
jave@wsu.edu

Cynthia Matuszek
Dept. of CSEE
University of Maryland, Baltimore County
cmat@umbc.edu

Matthew E. Taylor
School of EECS
Washington State University
taylorm@eecs.wsu.edu

Abstract

This paper presents an exploratory study where participants piloted a commercial UAS (unmanned aerial system) through an obstacle course. The goal was to determine how varying the instructions given to participants affected their performance. Preliminary data suggests future studies to perform, as well as guidelines for human-robot interaction, and some best practices for learning from demonstration studies.

Introduction

As robots become increasingly common, it becomes critical to allow humans to teach robots new skills – the set of possible skills a robot may need cannot be pre-programmed at design time and must be conveyed without requiring a consumer to write code. We are particularly interested in learning from demonstration (Chernova and Thomaz 2014), where a robot can learn from a human helping a robot to perform a task.

The core hypothesis of this work was that different instructions to the human participant would change how well they performed the task. In particular, we asked participants to fly a robot through an obstacle course. Participants were divided into two groups: one group would be told that the robot was an expensive piece of lab equipment and the other would be told that it was a toy. We expected people to fly more slowly and make fewer errors if they thought it was expensive, while those that thought it was a toy would complete the course faster, at the expense of additional errors. We chose a UAS platform (unmanned aerial system) because we felt it was believable that a flying robot was expensive and that it would be easier to recruit participants.

As detailed in the Results Analysis section, we are able to draw five preliminary conclusions:

- Contrary to our expectation, instructions given to participants had no effect on their performance in terms of speed or number of errors.
- Participants who self-reported as less nervous were significantly faster than those who self-reported as more nervous.
- Participants below the age of 30 were significantly faster than those 30 or over.
- Participants who reported playing over 3 hours of video games per week were significantly faster than those that did not.
- Contrary to our expectation, although we found factors correlated with performance as measured by time to completion, we found no factors correlated with performance as measured by the number of errors or collisions.

We were also quite surprised that after 90 participants and multiple collisions and crashes, the single robot used for the experiments is still able to fly accurately – it is accurate to say that no robots were harmed (much) during these experiments.

These results suggest future experiments to run to better predict the performance of different participants, as well as design principles for human-robot interfaces to maximize demonstration ability.

Background and Related Work

Discovering the best mechanism for controlling robots is of substantial interest to the HRI community, not only for improved operator training for teleoperatation, but also for obtaining improved demonstrations and guidance from human teachers. Learning by demonstration is a powerful tool, allowing training data to be collected from non-experts (Thomaz and Breazeal 2006) and used effectively in a variety of tasks (Chernova and Thomaz 2014). However, without instruction, the information presented by human demonstrators is often inefficiently chosen and ordered; the quality of data collected can be improved by providing demonstrators with clear instructions (Cakmak and Thomaz 2014; Cakmak and Takayama 2014).

Related work in HCI is sometimes applicable to robotics (Kadous, Sheh, and Sammut 2006; Fothergill et al. 2012). There are also parallels that can be drawn with the study of human instruction, particularly regarding scaffolding and ordering of tasks (Van Merriënboer, Kirschner, and Kester 2003). The role of maintaining and sharing models of cognitive state between teacher and student has also been explored in human teaching (Kieras and Bovair 1984),

the study of operator behavior (Boussemart and Cummings 2008), and HRI (Otero et al. 2008; Koenig, Takayama, and Matarić 2010).

The work most closely related to ours is research on the role of instructions in dialog (Foster et al. 2009; Fischer 2011; Cakmak and Thomaz 2012; Cakmak and Takayama 2014) and the interrelationship between user demonstration and type of learning algorithms (Suay, Toris, and Chernova 2012). Some human factors have been found to be correlated to operator performance in UAS design (Mouloua et al. 2001). However, to date, there has otherwise been comparatively little work on the role played by the specific elements of the instructions given to study participants; the best way to provide instructions to demonstrators is a complex question which is only beginning to be seriously explored (Suay, Toris, and Chernova 2012).

Experimental Design

After participants signed the relevant waivers, each underwent a brief training session on how to take off, steer, and land a Parrot AR Drone 2.0. Participants used AR.FreeFlight 2.4, an application developed by the robot manufacturer, on an iPad. Settings limited the maximum possible horizontal speed and the vertical speed was set to the minimum level, allowing the participants to focus on flight in the x-y plane. After training, the participant had three chances to fly the UAS through an obstacle course.

Figure 1 depicts the obstacle course participants are asked to fly. First, the participant takes off from a green square and flies it to the right side of a soft orange pole. Then, she flies it through a hula hoop and around the right of the hula hoop. Finally, the participant must fly around the left of the orange pole before once again landing in the green square. There are two blue flexible poles behind the hula hoop to mark the edge of the course. The participant was told to not pass or hit the blue poles as well as to try to land as much of the the unmanned aerial system (UAS) in the green square as possible. Any questions asked by the participant during the experiment were answered with pre-agreed answers, designed not to bias the participant. Answers to questions (e.g., "How much does the robot cost?") were deferred until after the participant completed the experiment.

Participants were split into two groups with different sets of instructions. One group was told that the UAS was an expensive piece of research equipment and that they should try to avoid damaging it, while the other group was instructed that the UAS was an inexpensive toy robot and we expected some toys would break during the experiment.[1] Both groups were told to fly through an obstacle course as quickly as possible without hitting any of the obstacles.

On each flight through the obstacle course, we recorded the time between takeoff and touchdown and the number of collisions with the obstacles. Collisions included hitting the orange pole, hula hoop, blue pole, and any wall. Other noted errors included the UAS passing the blue pole, if the participant landed less than half of the UAS in the green square,

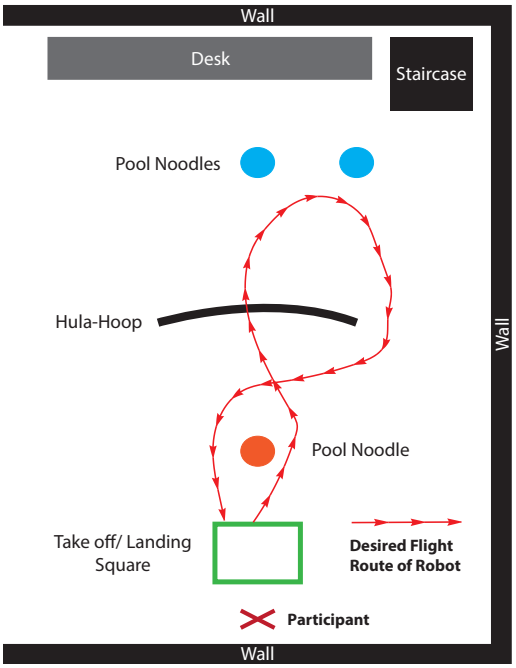

Figure 1: This diagram depicts the obstacle course participants must fly through. The distance from the take-off/landing square to the end of the course is ˜17 feet and the distance between the right side of the hula-hoop and the wall is ˜5 feet.

and if he or she crashed the UAS. A crash is defined as the participant flying into something hard enough to make the UAS's motor power off and fall to the ground, requiring the flight to begin again from the takeoff area.

Once the participants had flown through the obstacle course three times, they were asked to take a brief survey that would help identify features about themselves. Some key questions in this 13 question survey included:

- What is your age?[2]
- Do you agree with the following statement: "Flying the drone made me nervous"?
- How many hours a week do you play video games on average?
- Have you ever flown a drone/UAS before?

To recruit participants, we used several sources to advertise including fliers, an article about the experiment's dates and details on Washington State University's (WSU) news website, an in-person table rented at the WSU student center, and direct emails to a local engineering company. The study included a total of 90 participants: 62 men and 28 women. Each participant was randomly assigned to one of two equal-size groups. After excluding outliers (discussed more in the following section), the participants included 26 women and 49 men. The average age was 36 years old. Participants were compensated by entering into a lottery to win either $50 (two total) or $25 (four total).

[1] Both statements are true – the AR Drone is $300 and is used in the IRL Lab for both research and for classes.

[2] All participants were 18 or older.

Results Analysis

This section presents the results of our user study in four parts. Participants flew through the obstacle course three times but our analysis focuses on the second run. The data from the first trial had a bimodal distribution with a high number of errors and crashes; this first run-through was seen primarily as a practice run that allowed the participants to become more accustomed to the course and robot. By the third run, the deviations in times were noticeably lower as all groups converged towards the mean course time. In most cases, participants improved significantly between the first and second trials. The first trial was considered too volatile while the third trial considered too uniform to detect differences between participants. For these reasons, times from the second trial were considered to carry the most practical significance. The first and third flights of every participant were used to analyze the improvements that participants made over time (see below). One other source of noise that we did not control for is how long participants spent trying to land the robot. Although landing only required the participant to maneuver the robot over the landing square and to press the land button, some participants spent over 30 seconds hovering near the landing square trying to get into the exact center of the square, before landing.

To improve data interpretability, we assessed normality and removed outliers. To check normality, Shapiro-Wilk normality tests were run on the data with a 0.05 significance level. We concluded that the data distributions are not normally distributed but rather right skewed. Furthermore, the normal distribution was an appropriate model for the natural logarithm of the data. That is, the completion times can be modeled lognormal.[3] After log transformation, the data was put into a box and whisker plot. Data points that were 1.5 times the interquartile range above/below the third/first quadrant were generally considered outliers. Potential outliers that were close to this exclusion criteria were also put into a histogram, a probability plot, and/or a studentized deleted residual table to confirm that they were indeed outliers. These outlier tests were run on every data set individually. All hypothesis tests below are based on the transformed data. Comparisons of the means of the log times are equivalent to comparison of the median (untransformed) times.

Instructions Provided: Expensive vs. Toy

The key hypothesis that motivated this work was that giving different sets of instructions to participants would elicit different performances – if participants were told the robot was expensive, we hypothesized they would take longer to complete the course. Surprisingly, there was no correlation between instruction set and performance. The median flight time of both groups were evaluated using a two-sample t-test and was shown to not be statistically significant ($p = 0.5$). One reason could be that participants had different mindsets. Informal observation suggested that some participants aimed to complete the course as quickly as possible without worrying about damaging the robot, while others who were uncomfortable with flying the UAS focused more on not damaging it and worried less about time.[4]

Another observation was that some participants did not pay attention to the instructions: they forgot the correct way to manipulate the interface for flying the UAS after practicing outside before the test. There was not correlation between instructions and perceived nervousness in the post-experiment survey. However, it should be noted that some participants gave explicit feedback to our instructions with saying that it was good to know that the UAS was a toy or they would try to avoid damaging it since it was expensive.

[3] When analyzing the second trial of all participants, the minimum completion time was 14.6 seconds and the average was 31.4 seconds, but the maximum time was 58.1 seconds, showing a skewed distribution.

[4] It is possible that some participants worried about damaging the robot and the first time they hit something, their unease increased, causing them to hit additional obstacles.

Table 1: Summary of Statistical Results

Test	Condition 1 (# participants)	Time (sec.)	Condition 2 (# participants)	Time (sec.)	Stat. Sig.
Instructions	Expensive (35)	31.5	Toy (42)	30.7	No
Age	<30 years (27)	24.3	≥ 30 years (48)	35.1	Yes
Nervousness	Nervous (26)	35.9	Calm (43)	28.4	Yes
Video Games	<3 hours/week (57)	33.4	>3 hours/week (19)	24.3	Yes

Nervous vs. Not Nervous

Question six on the post-experiment survey asked: "Do you agree with the following statement: 'Flying the drone made me nervous.'" and the participants were asked to select from "strongly agree," "agree," "undecided," "disagree," and "strongly disagree." We hypothesized that if a participant was nervous, it should affect their performance. As we expected, the median time of flying the UAS increased significantly (7.5 seconds, confirmed by a two-sample t-test with $p < 0.01$) between the participants who were not nervous (28.4 ± 2.7 seconds) and those identified as nervous (35.9 ± 4.0 seconds) during the experiment. We observed that participants who visually appeared nervous tended to: 1) move the UAS forward a little and then hover it for a long time, 2) forget the correct way to control the UAS, and 3) vocalize louder (and more often) when hitting the obstacles. Furthermore, two-sample t-test shows that the participants who identified as nervous had a significantly ($p < 0.01$) larger average drop in time from the second run to the third compared to the non-nervous participants. This can be due to the fact that larger improvements could be made, while the participants' reduced nervousness made their performance more similar to those who were not nervous initially.

People who had previous UAS flight experience were less likely to be nervous than those who did not. 16.7% of participants who had previous experience flying UASs identified as nervous (the first two responses) while 40.3% of people with no experience identified as nervous (the final two responses). A binomial hypothesis test shows that the number difference between these two groups was statistically significant ($p < 0.05$).

We note that the relationship between course completion time and self-identified nervousness is correlational – future experiments will test whether performing poorly increases participants' nervousness, if nervousness causes poor performance, of if additional factors remain to be identified.

Participant Age

27 participants were below the age of 30 and 48 participants were at or above the age of 30 (again, after excluding outliers). Participants at or over the age of 30 flew 10.8 seconds slower than the younger population (35.1 ± 3.0 seconds vs. 24.3 ± 2.3 seconds). A two-sample t-test shows that the median time difference between these two groups was statistically significant ($p \ll 0.01$). Informal observations suggest that older participants were more deliberate in trying to avoid making mistakes.

Video Game Usage

Participants who played more than three hours of video games per week flew 9 seconds faster than those who played three hours or less (24.3 ± 2.6 seconds versus 33.4 ± 2.7 seconds). A two-sample t-test shows that the median time difference was statistically significant ($p \ll 0.01$) between the two groups. One possible explanation could be that participants who played more video games treated the course as a game rather than a task, achieving higher performance through more excitement and a relaxed attitude. If true, this would motivate our future design of more comfortable and intuitive robot interfaces for helping people provide high quality demonstrations.

Table 1 summarizes our results. The number of crashes (i.e., the robot hits something hard enough it stops flying and must be reset) did not correlate with any other factors, including nervousness, age, and video game usage. We also note that, despite a large number of collisions with stationary objects, we were able to use a single AR Drone for the more than two hours of total flight.

Discussion and Future Work

While this study was primarily exploratory, one concrete outcome of this project is a large data set describing 90 participants flying a robot, which we will release once we publish results in a archival venue. This section discusses take-home messages and future research.

This study focused on self-reported stress. Future work could look at trying to increase or decrease the level of stress by having the experimenters make disapproving vocalizations when the robot crashes, or saying calming phrases when the participant makes a mistake. We will also attempt to empirically measure stress (e.g., heart rate or galvanic skin response) to see if such measures correlate with performance. If they do, the robot could adjust its performance automatically, such as slowing down the maximum speed when the user is very nervous. Similarly, in non-training settings where a human and robot are collaborating on a shared task, the robot may wish to change its behavior depending on the user's stress level (e.g., move slower, take over more of the shared task, or become more autonomous).

In addition to focusing on "errors" made during flight, future work will use a motion tracking system to accurately record demonstrations. This data can then be used by multiple learning from demonstration algorithms in order to determine which users' demonstrations are most useful and what factors can influence the demonstration quality.

Future questions to investigate include:

- How predictable are the results? Can we use measured or

self-reported variables to predict the quality of a future robot demonstration from a user?

- Will increased practice time reduce stress and improve demonstrated performance? If so, can we predict how much practice a user should have with a platform before data is used for demonstration learning?
- If participants practice on a virtual robot using a realistic simulation, what will the impact be on user stress and demonstration quality when the user transitions to a physical platform?
- How does the design of the robot or interface change a user's stress level, and does this impact demonstration quality?

Acknowledgements

This research has taken place in the Intelligent Robot Learning (IRL) Lab, Washington State University. IRL research is support in part by grants from AFRL FA8750-14-1-0069, AFRL FA8750-14-1-0070, NSF IIS-1149917, NSF IIS-1319412, USDA 2014-67021-22174, and a Google Research Award. Madeline Chili was supported by the National Science Foundation Research Experiences for Undergraduates Program under Grant No. 1460917.

References

Boussemart, Y., and Cummings, M. 2008. Behavioral recognition and prediction of an operator supervising multiple heterogeneous unmanned vehicles. *Humans operating unmanned systems*.

Cakmak, M., and Takayama, L. 2014. Teaching people how to teach robots: The effect of instructional materials and dialog design. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 431–438.

Cakmak, M., and Thomaz, A. L. 2012. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 17–24.

Cakmak, M., and Thomaz, A. L. 2014. Eliciting good teaching from humans for machine learners. *Artificial Intelligence* 217:198–215.

Chernova, S., and Thomaz, A. L. 2014. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(3):1–121.

Fischer, K. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine* 32(4):31–38.

Foster, M. E.; Giuliani, M.; Isard, A.; Matheson, C.; Oberlander, J.; and Knoll, A. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the IJCAI conference*, 1818–1823.

Fothergill, S.; Mentis, H.; Kohli, P.; and Nowozin, S. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1737–1746.

Kadous, M. W.; Sheh, R. K.-M.; and Sammut, C. 2006. Effective user interface design for rescue robotics. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, HRI '06, 250–257.

Kieras, D. E., and Bovair, S. 1984. The role of a mental model in learning to operate a device. *Cognitive science* 8(3):255–273.

Koenig, N.; Takayama, L.; and Matarić, M. 2010. Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks* 23(8):1104–1112.

Mouloua, M.; Gilson, R.; Kring, J.; and Hancock, P. 2001. Workload, situation awareness, and teaming issues for uav/ucav operations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 45, 162–165. SAGE Publications.

Otero, N.; Alissandrakis, A.; Dautenhahn, K.; Nehaniv, C.; Syrdal, D. S.; and Koay, K. L. 2008. Human to robot demonstrations of routine home tasks: exploring the role of the robot's feedback. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 177–184.

Suay, H. B.; Toris, R.; and Chernova, S. 2012. A practical comparison of three robot learning from demonstration algorithm. *International Journal of Social Robotics* 4(4):319–330.

Thomaz, A. L., and Breazeal, C. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *AAAI*, volume 6, 1000–1005.

Van Merriënboer, J. J.; Kirschner, P. A.; and Kester, L. 2003. Taking the load off a learner's mind: Instructional design for complex learning. *Educational psychologist* 38(1):5–13.