More Than a Number: A Multi-dimensional Framework for Automatically Assessing Human Teleoperation Skill

Emily Jensen, Bradley Hayes, Sriram Sankaranarayanan
University of Colorado Boulder
Contact: emily.jensen@colorado.edu

University of Colorado Boulder

Research Motivation: Defining Skill

Safe teleoperation is a critical skill in domains such as:
- Robotic surgery
- Space exploration
- Construction and inspection

How do we know if someone is skillful at teleoperation?
- Most work defines manual rubrics
- Evaluation can be inconsistent over time or between judges

Skill is usually defined as a single value
- Performance metrics like accuracy or Boolean completion
- But, performance can be less skillful for different reasons, such as completing too slow or backtracking

This yields the following **research question**:
How can we define different aspects of teleoperation skill in a way that can be automatically assessed by a computer?

Our Contribution (tl;dr)

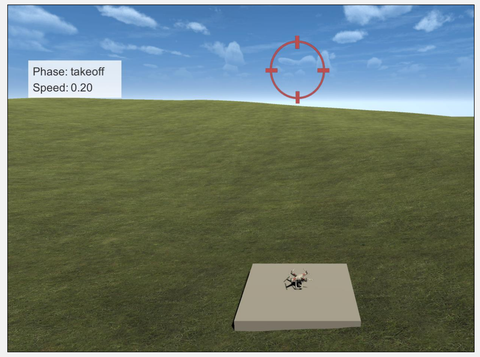
We argue the notion of skill needs nuance beyond a single measurement

We propose four possible dimensions to measure skill (though there may be more)

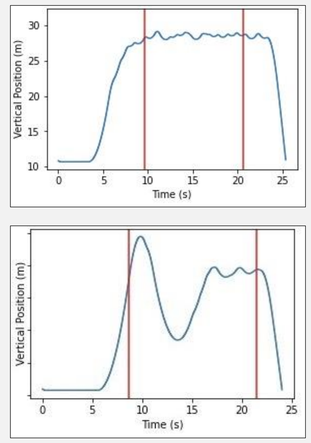
We provide an initial proof of concept to a UAV teleoperation task, using primitive tasks and formal specifications

UAV Teleoperation Task

Proof of concept pilot study
5 participants each competed 2 trials
Operate quadrotor using Xbox controller in Unity
Three task segments
- Take off to reach target
- Hover in target for 5 consecutive seconds
- Land at starting position



Unity task environment

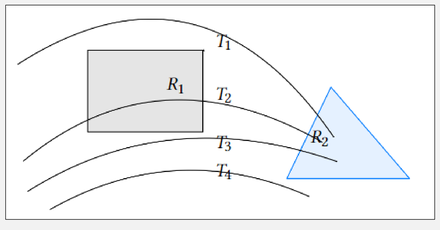


Example trajectories

Approach

Define the trial as a **sequence of primitive tasks**
Take off vertically and stop in floating target
Hover stationary in the target for 5 consecutive seconds
Land vertically and gently on landing pad
Other tasks could include orienting or lateral movements

Define **skill as a measure of performance** for a primitive task
Performance can be measured along **different dimensions**:
Efficiency: minimizing use of time or resources
Resilience: completing the task in adverse environments
Readiness: performing in different personal conditions
Robustness: adherence to task specifications
distinguishes "clearly" vs. "barely" correct
Different dimensions may be needed for other tasks



Calculate skill along efficiency and robustness dimensions
Time efficiency: slower completion is less efficient
Control efficiency: total variation distance of inputs
Robustness: defined specifications for primitive tasks

$$(v_y \geq 0)\ \text{UNTIL}_{[0,T]}\ \left(|y - H| \leq \epsilon\ \wedge\ |v_y| \leq \delta\right)$$

Takeoff: within T seconds, reach height H with velocity 0
Similar formulation for landing

$$\text{EVENTUALLY}\left(\text{Always}_{[0,T]}\ \left(y \in [y_{\min}, y_{\max}]\ \wedge\ |v_y| \leq \epsilon\right)\right)$$

Hover: stay in vertical range for T seconds with velocity 0

Results

- No takeoff or landing segments met specifications
- Relative differences in control efficiency may be more obvious for more complex tasks
- There may be tradeoff between different skill dimensions
- System designers can decide which dimensions are important

Participant	Trial	Takeoff RO	Takeoff TE	Takeoff CE	Hover RO	Hover TE	Hover CE	Land RO	Land TE	Land CE
1	1	-4.99	-5.90	-1.79e-06	0.25	-40.32	**-1.92e-06**	-0.82	-5.38	-5.74e-06
	2	-2.54	-6.11	-1.74e-06	*0.09*	-5.94	-2.91e-06	*-2.19*	*-6.08*	-2.60e-06
2	1	*-7.14*	-6.34	-1.62e-06	0.44	*-44.72*	-4.82e-06	-1.55	-3.53	-6.00e-06
	2	-3.72	-7.60	-2.62e-06	0.12	-15.66	-4.89e-06	-1.59	-4.42	-6.58e-06
3	1	-1.66	*-12.59*	-2.51e-06	0.38	-22.40	-9.72e-06	-1.31	-3.23	-3.03e-06
	2	-7.12	-8.70	**-0.85e-06**	0.25	-7.74	-6.62e-06	**-0.65**	**-2.58**	**-2.07e-06**
4	1	-1.63	-7.95	-1.97e-06	0.71	**-5.22**	-6.84e-06	-1.46	-3.85	*-6.96e-06*
	2	-0.51	-10.33	-1.97e-06	0.93	-7.10	-6.81e-06	-0.13	-4.44	-4.46e-06
5	1	**-0.00**	-9.60	*-5.67e-06*	**1.08**	-6.06	*-11.55e-06*	-1.85	-4.69	-5.65e-06
	2	-4.84	**-5.42**	-2.56e-06	0.95	-13.92	-8.15e-06	-1.32	-2.73	-6.01e-06

Note: RO = Robustness, TE = Time Efficiency, CE = Control Efficiency
best in column = **bold**, worst in column = *italics*

Future Work

Apply framework to more complex, concurrent tasks
Automatically generate training examples to target weak skills
Complete larger, generalizable user study
- Distribution of skill values
- Measure learning curves
- Correlate with user performance ratings