Recognizing User Proficiency in Piloting Small Unmanned Aerial Vehicles (sUAV)

Siya Kunde, Evan Palmer, and Brittany Duncan

***Abstract*—User proficiency in manual operation of autonomous systems is crucial to the performance of these systems because users are often the final barrier in detecting and correcting abnormal behavior in autonomy. This paper presents a new approach to identifying user proficiency in piloting small unmanned aerial vehicles (sUAVs) by first extracting meaningful features and then using a clustering method to generate ground truth. Pilot performance has been broadly explored in the field of aviation, but not for operation of sUAVs, and both are inherently different due expectations for training, location of user, and subsequent change in user's point-of-view. We propose a novel, hybrid approach to evaluate UAV pilot performance: combining human-rater data and computational methods that incorporate performance metrics to tune and homogenize the process (of applying controls) and the product (UAV flight path) of piloting sUAVs. The results reveal a spectrum of user skills that designers of these systems need to account for and the ways that users at different skill levels can be expected to respond, informing future autonomy design. In a 20 participant study, users were asked to fly a sUAV along 8 different flight paths of varying difficulty while the flight trajectory and user control inputs were recorded. We utilized unsupervised learning techniques to group pilots into proficiency groups and analyzed the clusters with respect to the features built. We also identified possible factors based on groupings to target training of these users. We validated our approach using new set of data from 12 participants.**

***Index Terms*—Human Factors and Human-in-the-Loop; Aerial Systems: Applications; Human-Robot Teaming**

I. Introduction

SMALL unmanned aerial vehicles (UAVs) are available in all shapes and sizes to users from different backgrounds—researchers, specialists, hobbyists and novices alike. The systems can be purchased and flown without any requirements for formal manual flight training as they come equipped with easy-to-use assistive modes like altitude hold, headless, one-touch take-off-and-landing. Additionally, the Federal Aviation Administration (FAA) only relies on a knowledge-based assessment to issue certificates for commercial operation.

While this is a great opportunity for wider adoption of UAVs, these systems can fail in many ways and under varying contexts—some of which could be avoided with proper user input and recovery. A study by [1] found that user response times increased when the users were asked to correct the failures occurring during flight by taking over manually. Additionally, accidents and crashes can result from poor manual operation skills itself.

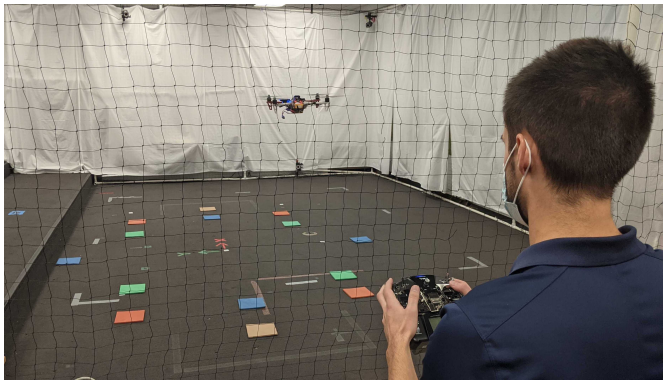


Fig. 1. Participant completing a task using the DJI Flamewheel sUAV and Futaba remote controller.

The performance of human-robot teams could be improved by building adaptable autonomous systems that delegate manual operations responsibilities via calibration to user proficiency levels [2]. Recognising user proficiency is a critical first step in enabling adaptable autonomous systems. Traditionally, in general aviation literature (elaborated in section II-B), work related to user proficiency has been investigated from an expert-novice binary classification perspective. However, [3] suggests that expert classification be approached in a naturalistic way by studying tasks and activities that are sufficiently challenging—so that real expertise can be elicited—and that can be performed on an ad hoc basis at the time of operation.

Our work seeks to address this problem by answering four foundational questions: *R1: Can we identify features that can meaningfully quantify pilot flight performance?*, *R2: Can we use these features to elicit natural proficiency groups within data?*, *R3: Can we provide meaningful labels for each group?*, and *R4: Can we utilize the produced labels to identify proficiency levels of additional pilots?*. Answering *R1* will give us the features that can be used for clustering. Answering *R2* will give insights into the proficiency groups resulting from manual flight skills within the data and the features that are important for each group to be successful in the task at hand. Answering *R3* is important as clustering labels do not typically have a meaning associated with them and we would like to understand the proficiency level of users in a group. Answering *R4* validates our method so that we can extend the findings in this paper to future work.

Manuscript received: September, 9, 2021; Revised December, 4, 2021; Accepted December, 27, 2021.

This paper was recommended for publication by Editor Jee-Hwan Ryu upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by grants from National Science Foundation (IIS-1638099, IIS-1750750 and IIS-1925368).)

Siya Kunde, Evan Palmer, and Brittany Duncan are with Department of Computer Science and Engineering, University of Nebraska-Lincoln, USA skunde, epalmer and bduncan@cse.unl.edu

Digital Object Identifier (DOI): see top of this page.

Leveraging methods from the aviation domain, we explore the four foundational questions using a 20 participant dataset where users were asked to complete 8 tasks of varying difficulty levels (as shown in Fig. 1). Our approach to answer *R1* is through a literature review presented in the next section. We utilize inputs from both human rater classification and computational methods, which we believe go hand-in-hand in exploratory pattern recognition work like this, to answer *R2* and *R3*. Lastly, we validate our approach by classifying a new set of user data from 12 participants (performing the same 8 tasks) to assess *R4*.

The study results produced 3 proficiency groups for the data which were algorithmically assigned meaningful labels with 2 indicating highest proficiency and 0 indicating lowest. Features like simple error distances, more complex time-series distance measures and summary statistics computed over UAV flight trajectory, and frequency domain features computed over user provided control inputs to affect the UAV flight, were all found to be important for various tasks. Finally, we were able to effectively validate our approach on new data using models with non-linear decision boundaries like Naive Bayes and Decision Trees.

II. Related Work

Our research regarding user proficiency assessment for human-UAV interaction is novel with prior work focusing on UAV team assessment and evaluation of proficiency in a UAV simulation environment. Individual performance assessment using flight data has been well explored in the general aviation domain and will be discussed in this section in addition to the mentioned UAV-based research.

A. Performance Analysis with UAVs

User performance analysis with UAVs has been addressed in literature from a couple of different perspectives: factors contributing toward UAV team operation performance [4] and analysis of planning and monitoring performance [5], [6], [7]. In a study, [4] identified and analyzed factors contributing towards team performance like planning, decision making, and situation assessment. In a simulation-based study, user profiles [7], [6] and representative simulation profiles [5] were extracted by using clustering techniques with simulation metrics like Score, Cooperation, and Aggressiveness to assess planning and monitoring skills of users in a multi-UAV simulation environment.

While the above research has been conducted for team settings and in simulation, our research tests individual user performance while manually operating a visual line of sight (VLOS) UAV.

B. Overview of Performance Analysis in Aviation

User performance analysis can help compare how users at different proficiency levels perform tasks and has been traditionally explored from a expert-novice binary classification perspective [8], [9], [10], [11]. Nittala [8] found speed and heading to be the most important features in predicting pilot skill level, but the ground truth in this study was based upon pilot flight hours. When asked to perform tasks in a flight simulator, expert pilots performed better than novices on vertical and longitudinal control, but not lateral control [9]. Xiong [10] found that experts performed better than novices in landing task stability which was evaluated in terms of roll, pitch, yaw, and glide rate. Apart from assessing control performance, a study on pilot ability to anticipate consequences of actions by Doan [11] found experts to be more accurate than novices, especially for trials that involved multiple, meaningfully related control movements.

Performance analysis has also been conducted to determine the impact of different factors on pilot proficiency when operating an aircraft [12], [13], [14]. Environmental factors like exposure to higher carbon dioxide levels [12] and hypoxic hypoxia [14], can negatively impact pilot performance. Recency of flight practice has been found to be a significantly stronger predictor of flight performance compared to time since initial flight training [13].

The work discussed in this section has explored pilot performance analysis in general aviation domain, and we seek to elicit the proficiency groups for users flying sUAVs.

C. Metrics for Flight Performance Analysis

In this section we highlight the various metrics that have been used in literature to quantify operator performance.

1) Distance-Based Similarity Metrics for Flight Path Analysis: Objective measures for trajectory analysis have been used in literature to determine the error in a pilot's flight path relative to some optimal flight path or optimal set of parameters. This can be accomplished by computing arithmetic mean error [15], [16], root mean square error [17], and standard deviation of error [15], [17]. While these metrics are useful in quantifying pilot performance by assessing flight paths [15], [16], [17] and controls [18], they do not account for differences in lengths between the user and target time-series data. Time-series similarity measures such as Dynamic Time Warping (DTW) and Fréchet distance have been applied in domains such as path planning for autonomous robots [19], voice recognition [20], hand-writing processing [21], and protein structure alignment [22] respectively. In this study we used DTW and Fréchet distance as features to enable a fair comparison between trajectories of unequal lengths and computation of trajectory error after completion of a task.

2) Frequency-Domain Measures for Control Input Analysis: Studies in aviation [23], [15], [24], [17] and automobile [25] domains have used frequency-domain metrics to enable analysis of an operator's performance when interacting with manual controls. Johnson [26] established the effectiveness of these metrics by demonstrating that they were capable of distinguishing a high performing pilot from a low performing pilot. We integrate these metrics in our evaluation to assess sUAV pilot flight performance.

III. Data Collection

The dataset (*user* data and the *autonomous* data) (described in Table I and Table II) was collected at the University of

Nebraska-Lincoln's (UNL) NIMBUS laboratory under an exemption filed with UNL IRB since no identifiable information was collected from the user. The DJI Flamewheel 450 sUAV was used along with Futaba controller in *Stabilize* mode. The user control inputs were captured as RC_IN data from the *mavros* Robot Operating System (ROS) package. The transmission frequency capabilities (10 Hz) of the telemetry used on the UAV limited the recording of the RC data, but the sUAV pose data was recorded using the Vicon motion capture system at 200 Hz.

A. User Dataset

The size of the user dataset (with 20 participants) is relatively small because we were limited by the available participants who were trained in flying sUAVs. The amount of prior flight experience varied across users with average flight experience of 62.4 hours (std. dev. = 98.44), with minimum reported as 1 hour and maximum of 400 hours. Flight frequency differed as well where users reported to practice yearly (2 users), monthly (7), weekly (8), daily (2) and once (1) respectively.

In contrast, the average flight experience in 6 months prior to the study was 6.27 hours (std. dev. = 7.74), with minimum reported as 0 hours and maximum of 30 hours. Users reported flight frequency in last 6 months to be monthly (7 users), weekly (8), once (2) and never (3) respectively.

Takeoff, stabilized hover and landing are basic components of any task a UAV pilot may perform. Additionally the operators may have to fly the UAV where they are either flying forward/backward by using the pitch control, left/right by using the roll control, turning the UAV around by using the yaw control, or use any combination of these to complete their task. Since we are presenting an approach to assess user proficiency in flying UAVs, we selected tasks like hover at a location, precision landing, square (with a yaw component in the middle of the task), figure 8 (to combine roll and pitch controls), alpha (to combine roll, pitch and throttle controls). Tasks also related to literature presented in section II-B (individual flight controls, glide path, multiple control movements).

The operator was positioned outside the netted flight cage area (visualized in Fig. 2). Markers were placed on the floor of the flight area (shown in Fig. 1) to indicate starting, intermediate, and ending waypoints, and the flight path was visually demonstrated to the operators by the experimenter prior to flying.

Users were asked to complete 8 tasks of varying difficulty through tracing the requested flight path by visiting designated waypoints in succession. The tasks differed in difficulty depending on the complexity of the trajectory the user was asked to follow and by requesting the user to fly nose-in (NI) (i.e. sUAV facing the user, reversed controls, anticipated to be higher in difficulty) instead of nose-out (NO) (i.e. facing away, traditional controls).

The flight tasks (visualized in Fig. 3) were:

1) Hover (NO): Take-off from (0, 0, 0) to (0, 0, 1), hover for 30 seconds, and land at (0, 0, 0)
2) Hover (NI): Take-off from (0, 0, 0) to (0, 0, 1), hover for 30 seconds, and land at (0, 0, 0)
3) Land (NO, nearby): Take-off from (0, 0, 0) to (0, 0, 1), come to a nearby X marker and land in center
4) Land (NO, far): Take-off from (0, 0, 0) to (0, 0, 1), go to a far X marker and land in center
5) Square (NO, turn, NI): Take-off from (-0.9, -0.9, 0) to (-0.9, -0.9, 1), complete a square with side length 1.5m (NO), come back to (-0.9, -0.9, 1), turn 180 degrees mid-air, complete a square with side length 1.5m (NI), return to (-0.9, -0.9, 1), and land
6) Eight (NO): Take-off from (-2, 0, 0) to (-2, 0, 1), complete a figure-8, return to (-2, 0, 1), and land
7) Eight (NI): Take-off from (-2, 0, 0), to (-2, 0, 1), complete a figure 8, return to (-2, 0, 1), and land
8) Alpha (NO): Take-off from (-1.5, 0.9, 0) to (-1.5, 0.9, 0.5), go right-diagonally upwards to (1.5, -0.9, 1.5), go left to (1.5, 0.9, 1.5) by maintaining the altitude at 1.5m, go to right-diagonally downwards to (-1.5, -0.9, 0.5), and land at (-1.5, -0.9, 0)

As an example, the best paths corresponding to some users are presented in Fig. 2 (selected by rater). These bear a resemblance to the paths the users were asked to follow and can be contrasted with the autonomous flight paths described in the next section and visualized in Fig. 3.

One participant's flight data was corrupted for task 6 (Eight NO) and hence we have 19 data points for that task. In total we had 159 data points, each with sUAV trajectory information provided by Vicon motion capture and control inputs provided by the user.

B. Autonomous Dataset

This dataset was collected as a baseline representing how the sUAV flies in autonomous mode without user assistance for the same trajectories (using Freyja [27]). A sUAV was autonomously flown in the flight paths that the users were asked to follow to complete the tasks of the study (visualized in Figure 3 with the same 8 unique flight paths).

IV. Method

The primary purpose of this work is to analyze user performance in manually flying sUAVs. Here we define a way to measure and detect performance of the sUAV pilots as a first step in creating an adaptable system. In the future, this can be integrated into a system used for adapting the robot task to suit user proficiency level to improve performance of the human-autonomy team. Below we outline the steps used to elicit user flight proficiency with UAVs.

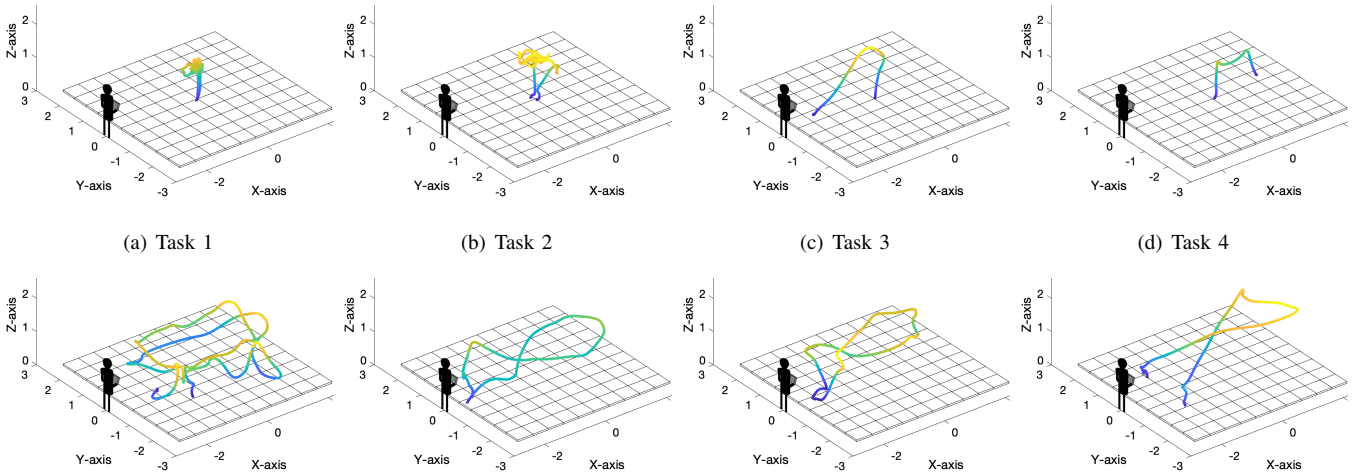
A. Ranking Trajectories

A rater ranked the visualizations of user flight trajectories, taking into account the definition of the task assigned to the participant and based on Shape criteria defined as *"Shape: The pilot has created a well-formed trajectory that matches the path described in the task definition."*

This criteria was chosen since it can be inferred to some extent by looking at visualizations. The visualization of the UAV

TABLE I
DESCRIPTION OF PARTICIPANT TASK PATTERNS

Designated Tasks	Hover		Land		Square	Eight		Alpha
	NO	NI	nearby	far	NO, turn, NI	NO	NI	NO
No. of Trajectories	20	20	20	20	20	19	20	20
Avg. Length (coordinates)	9722	8457	3152	3242	7244	5362	4046	5264
Avg. Duration (seconds)	48.892	42.44	15.754	16.228	36.59	26.829	20.464	26.41



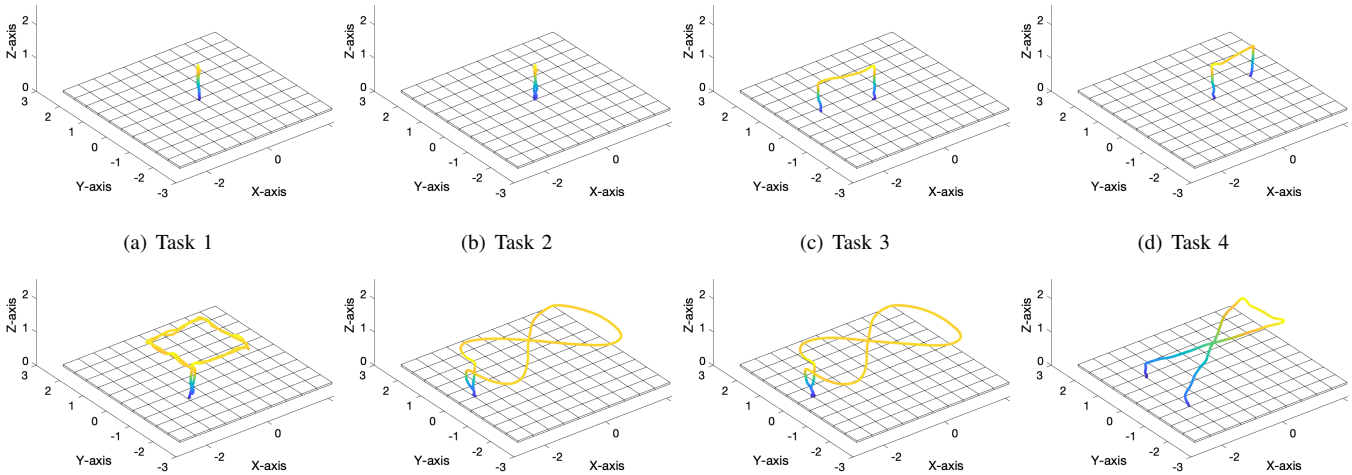
(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4

(e) Task 5 (f) Task 6 (g) Task 7 (h) Task 8

Fig. 2. When considering the shape of the trajectory, the rater ranked the following trajectories highest for Hover(NO): 2(a), Hover(NI): 2(b), Land(nearby): 2(c), Land(far): 2(d), Square(NO,turn,NI): 2(e), Eight(NO): 2(f), Eight(NI): 2(g) and Alpha: 2(h).

TABLE II
DESCRIPTION OF AUTONOMOUS TASK PATTERNS

Designated Tasks	Hover		Land		Square	Eight		Alpha
	NO	NI	nearby	far	NO, turn, NI	NO	NI	NO
No. of Trajectories	1	1	1	1	1	1	1	1
Avg. Length (coordinates)	7641	7634	2842	2849	13576	5624	5612	6866
Avg. Duration (seconds)	38.253	38.153	14.206	14.261	67.873	28.111	28.069	34.25



(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4

(e) Task 5 (f) Task 6 (g) Task 7 (h) Task 8

Fig. 3. Hover(NO): 3(a), Hover(NI): 3(b), Land(nearby): 3(c), Land(far): 3(d), Square(NO,turn,NI): 3(e), Eight(NO): 3(f), Eight(NI): 3(g) and Alpha: 3(h) created by the sUAV autonomously.

flight path created through autonomous flight was provided for reference as the "best" trajectory in the rankings and the user trajectories were presented in random order. The rater was first provided training practice in ranking trajectories with circle and a pattern and then asked to rank the users' trajectories on the 8 tasks of the study.

The ranking gave us a general idea of which trajectories were the best or worst and, in further steps, this would allow us to confirm the clustering outcomes for the data.

Additionally, the rater was asked to visually segment the trajectories so that we could extract the hover parts for task 1 and 2 (remove takeoff and landing segments), and the landing segments of the tasks 3 and 4. The subsequent analyses were completed by considering the relevant hover parts of task 1 and 2, full trajectories for tasks 3 to 8, and computing additional landing errors for tasks 3 and 4.

B. Feature Construction

We constructed a set of features from the flight data, which consisted of pilot control inputs and flight trajectory data, by which pilot performance would be evaluated. The features were identified according to previous literature discussed in section II-C. The reason for applying such features is because of their demonstrated success in alternative domains of research. The features were computed using (Crane [28], a high-performance computational platform by Holland Computing Center at UNL). We computed the following set of features for each trajectory, for each task:

1) DTW distance similarity measures (using library by [29]) of the flight paths in X (DTW_X), Y (DTW_Y), and Z (DTW_Z) axes considered individually; XY (2D points) together (DTW_{XY}); and XYZ (3D points) together (DTW_{XYZ}).
2) Fréchet distance (FR) measures (using library by [30]) of the flight paths.
3) Mean and standard deviation of the first derivative of the flight paths in X (V_X), Y (V_Y), and Z (V_Z) axes considered individually; XY (2D points) together (V_{XY}); and XYZ (3D points) together (V_{XYZ}).
4) Mean and standard deviation of the power spectral density of the control inputs (roll (P_R), pitch (P_P), yaw (P_Y), and throttle (P_T)).
5) For the landing tasks, we computed X (δ_X), Y (δ_Y), and XY (δ_{XY}) landing errors.

A total of 24 features were computed for tasks 1, 2, 5, 6, 7, and 8, (outlined in 1-4 above) and 27 features for the 2 landing tasks 3 and 4 (outlined in 1-5 above).

The chosen features are capable of representing user proficiency. The distance-based measures compare the user trajectory to a reference trajectory (in our case, the autonomously flown "best" trajectory), where a lower distance value indicates better performance by staying close to the expected path, while a larger value indicates worse performance as the user may have strayed from the expected path. When considering UAS operation within the scope of the completed task and the size of the flight area, a lower mean velocity and lower mean control inputs are indicative of a better controlled flight. Furthermore, a lower standard deviation of velocity measurements obtained from a particular task indicates that the pilot applied a more consistent control strategy during the respective task.

C. Detecting Natural Groups within the Data

In this step we tested several traditional clustering algorithms: KMeans, Spectral, and Agglomerative [31]. We used leave-one-out cross validation to conduct experiments across methods to arrive at a final algorithm to be used and the number of clusters detected in data.

We analyzed the generalizability of the models using two different validation metrics: *Dunn Index* and validation accuracy for ground truth data available to us in terms of task completion (1:yes, 0:no). Dunn Index provides a single score which summarizes the information about the compactness and the separation of clusters. Task completion classification accuracy, is an appropriate validation ground truth because it was not part of the features used to train the clustering model to make a prediction. As a last step in model selection process, both metrics were considered to choose the final model.

D. Assigning Meaningful Proficiency Labels to Clusters

For each task, data from all participants was clustered using the selected algorithm and number of clusters, and cluster labels were obtained. These labels are typically randomly assigned, and do not carry any meaning. To assign meaningful proficiency labels to the clusters, like 0 for lowest proficiency, we used the following method. We took the rankings produced by the rater (where 1 is best and N is worst ranked) for each task, computed the reciprocal rank and the root mean square (RMS) of the reciprocal rank over the clusters produced. A sorted ascending order of the values provided us the proficiency groups, such that the cluster with the higher RMS reciprocal rank was considered to be the cluster of users with highest proficiency, and the one with the lowest value was considered to be the cluster of users with lowest proficiency. This metric was used because it provided more weight for being at the top or bottom of a ranking.

V. Results And Discussion

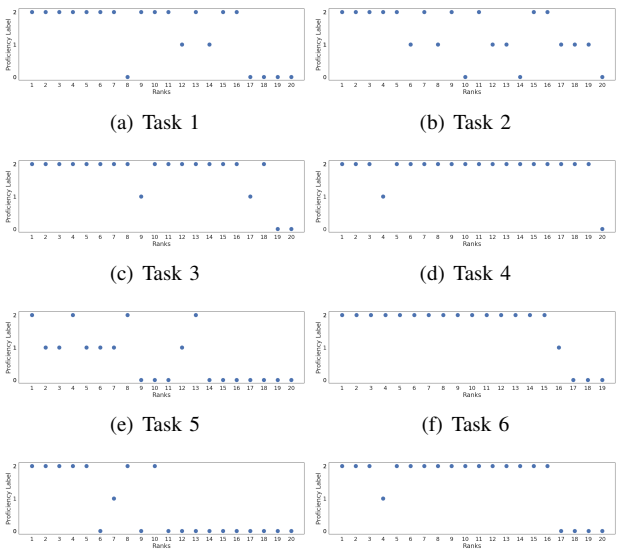
In this section, we describe and discuss the results by proficiency group and cluster label.

A. Detecting User Proficiency Groups

Table III shows the validation results for all the tested models and validation metrics (described in Section IV-C) averaged across all folds of the leave-one-out cross validation. Agglomerative (k=2,3) method produced high *Dunn Index* values, and we chose k=3 to optimize for task completion accuracy.

Next, the selected model was used to detect the proficiency groups and the methodology described in section IV-D was followed to produce meaningful proficiency labels. Hence, we had clusters corresponding to experts (label 2), intermediates (label 1) and novices (label 0). The user data is visualized

in Fig. 4 as proficiency label vs. ranks provided by the rater (higher ranked user is represented by a lower number with best rank as 1). A quick look confirms that *experts* with proficiency label 2 are generally located higher in the ranking order, followed by *intermediates* (label 1) and then *novices* (label 0). As a final step in this analysis, we used pair-wise ANOVA between proficiency groups to get features that produced significant results ($p <= 0.016$ with Bonferroni correction). The results of this step provide a high level overview of the clusters and is presented in Table IV.



(a) Task 1 (b) Task 2 (c) Task 3 (d) Task 4 (e) Task 5 (f) Task 6 (g) Task 7 (h) Task 8

Fig. 4. Visualization of proficiency labels produced by our clustering method (y axis) versus ranks provided by rater (x axis) for each data point for each task. User assignment to rankings and clusters changed based on the task. Proficiency was variable and not always in the order by the rater.

B. Assigning meaningful proficiency labels to clusters

The user proficiency labels (2 indicating highest proficiency and 0 indicating lowest) assigned to each cluster (and hence the user trajectories in that cluster) were averaged across all 8 tasks to understand the overall proficiency of the user (shown in Figure 5). The figure shows how users with higher mean proficiency have lower standard deviation values, indicating a more consistent performance. This is supported in literature where [32] found pilots with more training and practice (and hence higher expected proficiency) exhibited less variance in flight performance compared to pilots with less practice and training.

VI. Recommendations

Based on our investigation and the findings presented, we present recommendations for future work examining user proficiency. These relate to tasks to further discriminate user groups and user proficiency recommendations.

A. Task Discrimination

Overall, easier tasks that could be completed by all (except one) users, like landing, were not very discriminative. Tasks

TABLE III
Leave-one-out cross-validation accuracy for traditional clustering methods for number of clusters = 2,3,4,5.

Method	Dunn Index				Task Completion Acc.			
	No. of Clusters							
	2	3	4	5	2	3	4	5
Agglomerative	**0.54**	**0.52**	0.47	0.50	0.85	**0.90**	0.90	0.90
KMeans	0.45	**0.51**	0.46	0.46	0.84	**0.91**	0.90	0.89
Spectral	**0.36**	0.30	0.30	0.34	0.87	**0.91**	0.89	0.89

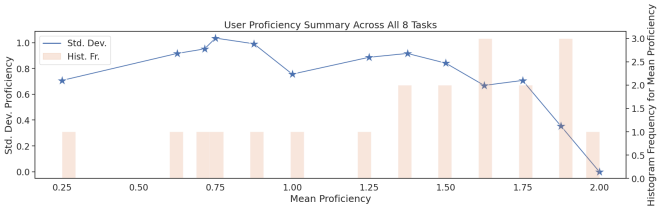


Fig. 5. User Proficiency Summary Across All 8 Tasks

that could be completed by majority of the users like hover (NO), eight (NO), and alpha (NO) seem to be able to detect well who did or did not complete a task and discriminate even further to some extent. The hover (NI) task was next in the level of difficulty was ideal for discriminating user profiles. Tasks square (NO, turn, NI) and eight (NI) proved to be very difficult for users to complete at all but the model was still able to discriminate user proficiency profiles. In future work, we recommend creating more intermediate tasks that are long enough to collect sufficient data and to refine the current classifications.

B. User Proficiency

Users with higher proficiency were clearly able to complete tasks that involved gradual altitude change, like task 8, by better controlling the mean and standard deviation Z velocity. Overall they had lower velocity metric values for easier tasks like hover, and landing, but the higher values for more difficult tasks like square and eight. A combination of strategies was observed where only $STD(V_X)$ was lower during the eight (NO) task and all metrics of Z velocity were lower during the alpha task, while all others were higher values. The high proficiency group had lower values for all metrics computed for assessing landing precision in both landing tasks. They also had lower values for other distance metrics for all tasks except DTW_Z for hover (NI) and eight (NO). For the hover tasks and eight (NO) task, higher mean and standard deviation values of control power spectral density were observed for the high proficiency group, while for task land (nearby) lowest proficiency group had the highest values, and intermediate group for the square (NO, turn, NI). For other tasks, either no trends were found to be significant or different groups had higher values. Based on these observations, we recommend working with users to improve their flight velocity, control inputs, and trajectory following to minimize distance.

VII. Inferring Proficiency of New Pilots

In this step, we validate the output of our approach (proficiency labels produced for each flight path flown by users)

TABLE IV

ORDERED USER PROFICIENCY LABELS (2 HIGHEST AND 0 LOWEST) TO SHOW RELATIONSHIP AMONG CLUSTERS FOR FEATURES (AS DEFINED IN SECTION IV-B). STATISTICALLY SIGNIFICANT DIFFERENCES (SIGNIFICANCE LEVEL AT 0.016 WITH BONFERRONI CORRECTION) ARE INDICATED WITH A FOR (0 AND 1), B FOR (0 AND 2), AND C FOR (1 AND 2). A BOLD PROFICIENCY LABEL INDICATES CLUSTER WITH SINGLE USER.

Task	**Proficiency Ordering**	**Features Exhibiting Trend**
Hover (NO)	$2 < 1 < 0$	FRbc, MEAN(V_X)b, STD(V_X)b, MEAN(V_Y)ab, STD(V_Y)b, MEAN(V_Z)b, STD(V_Z)b, MEAN(V_{XY})b, STD(V_{XY})b, MEAN(V_{XYZ})b, STD(V_{XYZ})b
	$0 < 1 < 2$	MEAN(P_P)bc, MEAN(P_R)b, MEAN(P_T)b, STD(P_P)b, STD(P_R)b, STD(P_T)b, STD(P_Y)b
	$2 < 0 < 1$	DTW_Xac, DTW_{XY}ac, DTW_{XYZ}ac
	$0 < 2 < 1$	MEAN(P_Y)b
Hover (NI)	$1 < 0 < 2$	DTW_Zc, MEAN(P_P)b, MEAN(P_R)bc, MEAN(P_T)bc, MEAN(P_Y)bc, STD(P_P)bc, STD(P_R)bc, STD(P_T)bc, STD(P_Y)bc
	$2 < 1 < 0$	MEAN(V_X)b, MEAN(V_Y)ab, MEAN(V_Z)ab, MEAN(V_{XY})ab, MEAN(V_{XYZ})ab
	$1 < 2 < 0$	STD(V_Y)ab, STD(V_Z)ab, STD(V_{XY})ab, STD(V_{XYZ})ab
Land (Nearby)	$1 < 2 < 0$	MEAN(P_P)ac, MEAN(P_R)ac, MEAN(P_Y)ac, STD(P_P)ac, STD(P_R)ac, STD(P_T)ac, STD(P_Y)ac, DTW_Xb, DTW_Yb, DTW_{XY}ab, MEAN(V_Y)b, STD(V_Y)b, δ_Yb
	$2 < 1 < 0$	FRb, DTW_{XYZ}b, MEAN(V_Z)b, MEAN(V_{XYZ})b, δ_Xb, δ_{XY}b
Land (Far)	$2 < \mathbf{0} < \mathbf{1}$	MEAN(V_X)c, STD(V_X)c, MEAN(V_Y)bc, MEAN(V_Z)c, MEAN(V_{XY})bc, STD(V_{XY})c, MEAN(V_{XYZ})c
	$\mathbf{0} < 2 < \mathbf{1}$	STD(V_Y)c, STD(V_Z)c, STD(V_{XYZ})c
	$2 < \mathbf{1} < \mathbf{0}$	FRb, DTW_Xb, δ_Xb, δ_{XY}b
	$\mathbf{1} < 2 < \mathbf{0}$	DTW_Yb, DTW_{XY}b, DTW_{XYZ}b, δ_Yb
Square	$0 < 1 < 2$	MEAN(V_X)bc, STD(V_X)bc, STD(V_Y)b, MEAN(V_Z)bc, STD(V_Z)b, MEAN(V_{XY})b, STD(V_{XY})b, MEAN(V_{XYZ})b, STD(V_{XYZ})bc
(NO,turn,NI)	$1 < 2 < 0$	DTW_Yab
	$0 < 2 < 1$	STD(P_P)a, STD(P_R)a, STD(P_T)a, STD(P_Y)a
Eight (NO)	$2 < \mathbf{1} < 0$	FRbc, DTW_Xbc, DTW_{XY}bc, DTW_{XYZ}bc
	$2 < 0 < \mathbf{1}$	DTW_Ybc, STD(V_X)ac
	$\mathbf{1} < 0 < 2$	MEAN(P_P)bc, MEAN(P_Y)c, STD(P_P)bc, STD(P_R)bc, STD(P_Y)bc
	$0 < 2 < \mathbf{1}$	DTW_Za, MEAN(V_X)c, STD(V_Y)a, MEAN(V_Z)c, STD(V_Z)c, MEAN(V_{XY})c, STD(V_{XY})ac, MEAN(V_{XYZ})c, STD(V_{XYZ})ac
Eight (NI)	$2 < \mathbf{1} < 0$	FRb, DTW_Xbc
	$2 < 0 < \mathbf{1}$	DTW_Zc, DTW_{XY}bc, DTW_{XYZ}bc, MEAN(P_Y)c
	$0 < 2 < \mathbf{1}$	MEAN(V_X)ac, STD(V_X)ac, MEAN(V_Y)a, STD(V_Y)ac, MEAN(V_Z)ac, STD(V_Z)a, MEAN(V_{XY})ac, STD(V_{XY})ac, MEAN(V_{XYZ})ac, STD(V_{XYZ})ac, STD(P_P)b, STD(P_R)b, STD(P_T)b, STD(P_Y)b
Alpha (NO)	$2 < 0 < \mathbf{1}$	MEAN(V_Z)c, STD(V_Z)ac, MEAN(V_{XYZ})c, STD(V_{XYZ})ac
	$0 < 2 < \mathbf{1}$	STD(V_X)ac, MEAN(V_Y)ac, STD(V_Y)ac, MEAN(V_{XY})ac, STD(V_{XY})ac
	$\mathbf{1} < 2 < 0$	FRb, DTW_Xab, DTW_{XY}ab
	$2 < \mathbf{1} < 0$	DTW_Yb, DTW_Zb, DTW_{XYZ}ab

TABLE V

TASK COMPLETION (TC) AND CRASH DATA FOR DIFFERENT USER PROFICIENCY LABELS (2 HIGHEST AND 0 LOWEST)

Proficiency Label	**T1**		**T2**		**T3**		**T4**		**T5**		**T6**		**T7**		**T8**	
	TC	Crash	TC	Crash	TC	Crash	TC	Crash	TC	Crash	TC	Crash	TC	Crash	TC	Crash
0	1/5	4/5	1/3	0/3	1/2	0/2	0/1	1/1	0/10	7/10	0/3	1/3	0/12	7/12	0/4	2/4
1	2/2	0/2	1/7	3/7	2/2	0/2	1/1	0/1	2/6	3/6	0/1	0/1	0/1	1/1	1/1	0/1
2	13/13	0/13	9/10	0/10	16/16	0/16	18/18	0/18	1/4	1/4	14/15	0/15	3/7	1/7	15/15	0/15

TABLE VI

F1 SCORES FOR USER PROFICIENCY CLASSIFICATION USING MODELS: NAIVE BAYES (NB), DECISION TREE (DT), LOGISTIC REGRESSION (LR), RANDOM FOREST (RF), AND K NEAREST NEIGHBORS (KNN)

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**
NB	**1.00**	**1.00**	**1.00**	0.05	**1.00**	**0.28**	0.67	0.43
DT	0.79	0.09	0.72	**0.41**	0.72	0.25	**0.76**	**0.75**
LR	0.36	0.00	0.50	0.05	0.58	0.28	0.70	0.44
RF	0.36	0.00	0.54	0.05	0.58	0.28	0.68	0.43
KNN	0.36	0.00	0.57	0.05	0.66	0.28	0.70	0.44

by classifying data from new users into the proficiency groups detected for the previous dataset. For this, data was collected from 12 new participants for the same 8 tasks. A rater was provided a interactive visualization of the user trajectories to inspect and asked to assign either *Expert*, *Intermediate*, or *Novice* labels (since previously 3 groups were detected) to each flight path, which were then used to assess the classification accuracy. These values were assessed for agreement (based on the agreement formula by Wobbrock [33]) with ratings from two other raters collected in similar fashion. The agreement scores for each task, averaged across all participants were 0.7 for Hover(NO), 0.92 for HOVER(NI), 0.92 for Land (nearby), 0.77 for Land (far), 0.81 for Square (NO, turn NI), 0.92 for Eight(NO), 0.92 for Eight (NI), and 0.77 for Alpha(NO).

Table VI shows that out of all the classification methods, NaiveBayes (NB) and DecisionTree (DT) classifiers performed the best. NB was able to recover the proficiency labeling generated by the clustering for tasks 1, 2, 3, 5 and 6, on the unseen data, and with highest possible value of the F1 score for four tasks. DT scored highest for the tasks 7 and 8 (difficult for most users to perform), and task 4. No method performed well for task 4 and we hypothesize that this is because majority of

the users were assigned to the same cluster with single users in the other two clusters.

Based on these outcomes, we observe that linear decision boundaries are inadequate to classify proficiency.

VIII. Conclusion

Identifying user proficiency in manual handling of sUAVs is an important aspect of the design of adaptable autonomous system. It's crucial to the performance of such systems to be able to detect proficiency groups and assign new users to them preferably in real time before or during flights. In this work we focused on achieving this goal by identifying and computing relevant features which were then used along with clustering techniques to determine proficiency groups of users. Additionally the groups were given meaningful proficiency labels. The proficiency profiles identified in this work are based solely on pilots' flying capabilities and were validated with data from new users. Future work will focus on extending the analysis to understand a temporal evolution of proficiency profiles, and users' non-flight related skills.

Acknowledgment

This work was supported by grants from National Science Foundation (IIS-1638099, IIS-1750750 and IIS-1925368).

References

[1] S. Kunde, S. Elbaum, and B. A. Duncan, "Characterizing user responses to failures in aerial autonomous systems," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1587–1594, 2020.

[2] M. Fdez-Carmona, C. Urdiales, and F. Sandoval, "Reactive adapted assistance for wheelchair navigation based on a standard skill profile," in *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*. IEEE, 2014, pp. 2152–2157.

[3] T. Farrington-Darby and J. R. Wilson, "The nature of expertise: A review," *Applied ergonomics*, vol. 37, no. 1, pp. 17–32, 2006.

[4] N. J. Cooke, H. K. Pedersen, C. Olena, J. C. Gorman, and A. Dee, *20. Acquiring Team-Level Command and Control Skill for UAV Operation*, ser. Advances in Human Performance and Cognitive Engineering Research. Emerald Group Publishing Limited, Jan 2006, vol. 7.

[5] V. Rodríguez-Fernández, H. D. Menéndez, and D. Camacho, "Analysing temporal performance profiles of uav operators using time series clustering," *Expert Systems with Applications*, vol. 70, 2017.

[6] ——, "Analyzing planning and monitoring skills of users in a multi-uav simulation environment," in *Advances in Artificial Intelligence*, J. M. Puerta, J. A. Gámez, B. Dorronsoro, E. Barrenechea, A. Troncoso, B. Baruque, and M. Galar, Eds. Cham: Springer International Publishing, 2015, pp. 255–264.

[7] ——, "Automatic profile generation for uav operators using a simulation-based training environment," *Progress in Artificial Intelligence*, vol. 5, no. 1, pp. 37–46, Feb 2016.

[8] S. K. Nittala, C. P. Elkin, J. M. Kiker, R. Meyer, J. Curro, A. K. Reiter, K. S. Xu, and V. K. Devabhaktuni, "Pilot skill level and workload prediction for sliding-scale autonomy," in *International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018.

[9] A. H. Bellenkes, C. D. Wickens, and A. F. Kramer, "Visual scanning and pilot expertise: the role of attentional flexibility and mental model development." *Aviation, space, and environmental medicine*, vol. 68, no. 7, pp. 569–579, 1997.

[10] W. Xiong, Y. Wang, Q. Zhou, Z. Liu, and X. Zhang, "The research of eye movement behavior of expert and novice in flight simulation of landing," in *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 2016, pp. 485–493.

[11] S. M. Doane, Y. W. Sohn, and M. T. Jodlowski, "Pilot ability to anticipate the consequences of flight actions as a function of expertise," *Human factors*, vol. 46, no. 1, pp. 92–103, 2004.

[12] J. G. Allen, P. MacNaughton, J. G. Cedeno-Laurent, X. Cao, S. Flanigan, J. Vallarino, F. Rueda, D. Donnelly-McLay, and J. D. Spengler, "Airplane pilot flight performance on 21 maneuvers in a flight simulator under varying carbon dioxide concentrations," *Journal of Exposure Science & Environmental Epidemiology*, vol. 29, no. 4, 2019.

[13] A. Haslbeck and H.-J. Hoermann, "Flying the needles: Flight deck automation erodes fine-motor flying skills among airline pilots," *Human Factors*, vol. 58, no. 4, pp. 533–545, 2016.

[14] F. Bouak, O. Vartanian, K. Hofer, and B. Cheung, "Acute mild hypoxic hypoxia effects on cognitive and simulated aircraft pilot performance," *Aerospace medicine and human performance*, vol. 89, no. 6, 2018.

[15] M. Ebbatson, D. Harris, J. Huddlestone, and R. Sears, "Combining control input with flight path data to evaluate pilot performance in transport aircraft," *Aviation, space, and environmental medicine*, vol. 79, no. 11, pp. 1061–1064, 2008.

[16] ——, "The relationship between manual handling performance and recent flying experience in air transport pilots," *Ergonomics*, vol. 53, no. 2, pp. 268–277, 2010.

[17] E. Schubert, B. Appel, and G. Hüttig, "Assessment of manual flying skills by combining aircraft parameters with pilot control inputs," in *17th International Symposium on Aviation Psychology*, 2013, p. 177.

[18] A. Haslbeck, H.-J. Hoermann, and P. Gontar, "Stirring the pot: Comparing stick input patterns and flight-path control strategies in airline pilots," *The International Journal of Aerospace Psychology*, vol. 28, no. 1-2, pp. 15–30, 2018.

[19] T. Dang, F. Mascarich, S. Khattak, C. Papachristos, and K. Alexis, "Graph-based path planning for autonomous robotic exploration in subterranean environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[20] L. Muda, B. KM, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *Journal of Computing*, vol. 2, no. 3, 2010.

[21] E. Sriraghavendra, K. Karthik, and C. Bhattacharyya, "Fréchet distance based approach for searching online handwritten documents," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 1. IEEE, 2007, pp. 461–465.

[22] T. Wylie and B. Zhu, "Protein chain pair simplification under the discrete fréchet distance," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1372–1383, 2013.

[23] P. A. Hebbar and A. A. Pashilkar, "Pilot performance evaluation of simulated flight approach and landing manoeuvres using quantitative assessment tools," *Sādhanā*, vol. 42, no. 3, pp. 405–415, Mar 2017.

[24] M. Ebbatson, J. Huddlestone, D. Harris, and R. Sears, "The application of frequency analysis based performance measures as an adjunct to flight path derived measures of pilot performance," *Human Factors and Aerospace Safety*, vol. 6, no. 4, pp. 383–394, 2006.

[25] D. V. McGehee, J. D. Lee, M. Rizzo, J. Dawson, and K. Bateman, "Quantitative analysis of steering adaptation on a high performance fixed-base driving simulator," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 7, no. 3, pp. 181–196, 2004.

[26] N. Johnson, E. Rantanen, and D. Talleur, "Time series based objective pilot performance measures," *International Journal of Applied Aviation Studies (IJAAS)*, 2004.

[27] A. Shankar, S. Elbaum, and C. Detweiler, "Freyja: A full multirotor system for agile & precise outdoor flights," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[28] "Hcc. holland computing center — nebraska." [Online]. Available: https://hcc.unl.edu/

[29] P. Rouanet, "Dynamic time warping python module," 2014. [Online]. Available: https://github.com/pierre-rouanet/dtw

[30] T. Eiter and H. Mannila, "Computing discrete fréchet distance," Christian Doppler Laboratory for Expert Systems, TU Vienna, Austria, Technical Report CD-TR 94/64, 1994.

[31] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[32] A. Haslbeck, P. Kirchner, E. Schubert, and K. Bengler, "A flight simulator study to evaluate manual flying skills of airline pilots," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 58, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2014, pp. 11–15.

[33] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers, "Maximizing the guessability of symbolic input," in *CHI'05 extended abstracts on Human Factors in Computing Systems*, 2005, pp. 1869–1872.