Pilot Workload Rating Predictions Using Image Data and Recurrent Neural Networks

Keiko Nagami
Aerospace Engineer
Ames Research Center
Moffett Field, CA, USA

Carlos Malpica
Aerospace Engineer
Ames Research Center
Moffett Field, CA, USA

Mac Schwager
Associate Professor
Stanford University
Stanford, CA, USA

ABSTRACT

In this work, we augmented existing methods for estimating pilot workload ratings with deep neural networks trained using data from simulated flight tests in the Vertical Motion Simulator (VMS). We used an existing method, Spare Capacity Operations Estimator (SCOPE), along with a recurrent neural network and conducted comparison studies between the two methods individually, and when used together. We found that using both methods together can improve the result over using either approach alone. In our first test case, we achieved an improved linear correlation coefficient of 0.409 over that of SCOPE alone at 0.352 on the training dataset. Through cross validation, we also found that the results may be dependent on the split of training vs. validation data, and that further investigation should be conducted to understand what additional inputs to the neural network model should be made.

INTRODUCTION

Pilot workload ratings play a key role in validating the control system design of an aircraft. However, this metric is difficult to predict, and can be a subjective measurement, prone to variability. Typically, a workload rating is determined by surveying pilots who test a control system via specified Mission Task Elements (MTEs). Understanding how to predict the workload rating of a vehicle and its control system can inform engineers on how to improve designs before conducting flight test experiments. Requiring a set of pilots to perform flight tests while iterating in the design stage can restrict the amount of tuning that can be done before testing. Having a model that can gauge workload ratings will benefit this design process. Current work in this field involves complex models that can reach estimates with as high as 93% correlation (Ref. 1) to pilot workload ratings. However, there are several factors that could inform a pilot's workload rating that are not contained in these models. For this reason, we propose a method that uses deep neural networks that have the capability of maintaining a larger set of inputs that could influence a pilot's workload rating. Specifically, we will use a learning-based approach which leverages the success of existing methods, while using deep neural networks to account for unmodeled information from data.

Pilot workload is a metric that considers how much effort a pilot must exert to reach a desired flight performance. This metric is particularly difficult to quantify, as there are several factors that could influence the workload, and because effort

Figure 1. Image of Vertical Motion Simulator cab

can be a subjective measure. The assessment of the handling qualities of an aircraft considers the combination of both pilot workload and performance. Presently, a common approach to evaluate handling qualities is conducted by meeting specifications standardized in ADS-33 (Ref. 2). These metrics allow engineers to iterate through the control design process and optimize toward Level 1 handling qualities. In practice however, there are several other factors that are not expressed through these standards when pilots evaluate a control system. For this reason, test pilots validate control systems through Mission Task Elements (MTEs), in simulated and real flight tests, and report workload and handling qualities ratings through scales like Bedford ratings (Ref. 3) and Cooper-

Presented at the VFS Aeromechanics for Advanced Vertical Flight Technical Meeting, San Jose, CA, Jan 25-27, 2022.

Harper (Ref. 4) ratings, respectively. As such, workload and handling qualities ratings prediction models can be difficult to formulate, as there are multiple factors that may influence a pilot's rating.

Handling qualities provide valuable insight in the design of aircraft and their control systems and consider the flight experience of both pilots and passengers. While the ability to achieve desired task performance is an important factor, the effort required to achieve this level of performance should not be overlooked. For this reason, we study methods to improve prediction methods of pilot workload ratings of aircraft. In this work we will be using data taken from simulated flights in the Vertical Motion Simulator (VMS) at Ames Research Center shown in Figure 1 to produce pilot workload rating predictions.

RELATED WORK

Previous works on developing models to quantify pilot workload have been presented by Roscoe and Wilkinson (Ref. 5), and Bachelder (Ref. 1). Specifically, the method presented in Reference 1, Spare Capacity OPerations Estimator (SCOPE), has been applied to several applications to estimate workload (Ref. 6), and extended to estimating handling qualities ratings (Ref. 7, 8), modeling pilot behavioral objectives (Ref. 9) and control models (Ref. 10). However, it is unclear if the inputs and parameters of these models are sufficient to characterize the human-vehicle interaction. As an example, some works have shown that a pilot's role as an active pilot compared to a system supervisor contributes to a pilot's ability to determine the vehicle altitude (Ref. 11, 12), and some works use information about optical flow as a variable in the workload metric (Ref. 6). Other models determine that workload is linearly dependent on the phase margin of the system (Ref. 9). As such, it is challenging to determine which factors have the largest impact on pilot workload and handling qualities, and how control systems can be improved according to these metrics.

While existing approaches to determining pilot cost functions (Ref. 9), pilot control models (Ref. 13, 14, 10, 15), and workload estimates (Ref. 1) provide relatively good approximations, learning-based methods could potentially improve upon these models using collected data. Learning-based methods that apply neural networks to existing methods have been used in several robotics applications (Ref. 16, 17, 18) and could be used to augment current methods for determining metrics that are difficult to model analytically such as pilot workload rating estimates. These learning methods are beneficial in that they leverage models that are already known to well approximate the system they are applied to. This is in contrast to other learning-based methods that start 'from scratch' and have longer training times, and require far more samples to produce an adequate solution. In our approach, we augment existing methods from the literature on estimating pilot workload ratings with a neural network to produce an improved solution.

Ultimately, we aim to develop a model that predicts pilot workload ratings based on the following inputs: visual information, inertial measurements, and pilot inputs from data collected in Vertical Motion Simulator (VMS) experiments. Our goal is to leverage the flexibility of learning-based methods while taking advantage of good models that approximate pilot workload ratings. We want to utilize the correlations that previous works have shown (Ref. 1) and take a learning-based approach that will incorporate previous work by first computing the pilot workload estimate approximated by these existing methods, and then learning a residual value to obtain an improved estimate. We use this approach to determine if additional inputs and a neural network will aid in formulating a model for workload metrics.

VMS DATA

Data was collected via the Vertical Motion Simulator (VMS) at NASA Ames Research Center over a series of four weeks (Ref. 19). The VMS simulates several aspects of real flight for test pilots. The cab of the VMS exerts accelerations based on the simulated flight of the vehicle. The pilot is also able to view a simulated scene of a virtual environment from the windows of the cab, and visual cues on the dashboard give the pilot information on the aircraft's state. The pilot then interacts with the environment through control inputs that

Table 1. Performance Standards Tested for Vertical Maneuver MTEs

Description	Desired (ADS-33 / UAM)	Adequate (ADS-33 / UAM)
Maintain the longitudinal and lateral position within X ft of a point on the ground	$\pm$3 ft / $\pm$1.5 ft	$\pm$6 ft / $\pm$3.9 ft
Maintain start/finish altitude within X ft	$\pm$3 ft / $\pm$1 ft	$\pm$6 ft / $\pm$2.6 ft
Maintain heading within X deg	$\pm$5 deg / $\pm$5 deg	$\pm$10 deg / $\pm$10 deg
Complete the maneuver within X sec	13 sec / 30 sec	18 sec / 40 sec

replicate those of an aircraft. Tests were conducted for various maneuvers and task performance standards.

For this paper, we focus on a Vertical Maneuver MTE, which is described as follows: The pilot begins in a stabilized hover, then proceeds to execute a vertical ascent to a specified altitude. At the top of this ascent, the pilot stabilizes the aircraft, then subsequently descends back to the initial hover position. Two variations of this Vertical MTE were tested: One designated for Urban Air Mobility (UAM), and another based on ADS-33 (Ref. 2). Descriptions of their respective performance standards are listed in Table 1. The UAM tests were conducted both with and without turbulence while the ADS-33 tests were conducted solely without turbulence.

Additionally, four vehicle configurations were considered, three of which use variable rotor speed control (with constant pitch), and one which uses collective control (with constant rotor speed). The three variable rotor speed control configurations vary in their heave disturbance rejection and control response specifications. In Figure 2, we show the mean and standard deviations of Cooper-Harper Handling Qualities Ratings from the VMS data, sorted by performance standard, vehicle, and turbulence level. Additional information on the VMS experiments and data collected is documented by Withrow-Maser et al. in Reference 20.

APPROACH

Our approach is to use a combination of existing methods from prior works and learning-based methods to develop a method to predict pilot workload. The existing approach used in this work is from Reference 1, where the authors quantify a Bedford rating estimate with the following equation,

$$B_{est} = C\mu^{\gamma} + D, \tag{1}$$

where B_{est} is the SCOPE Bedford rating estimate, μ is the workload stimulus, and $C, D,$ and γ are constants specific to the MTE. Here, the workload metric μ is computed as follows:

$$\mu = \sigma_{\dot{y}_e}\, \sigma_{\dot{\delta}}, \tag{2}$$

which is consistent with the Hover MTE application in Ref. 8. Here, $\sigma_{\dot{y}_e}$ is the standard deviation of the position error rate between the aircraft and its target, and $\sigma_{\dot{\delta}}$ is the standard deviation over each run's time series data of the stick control rate. Once the Bedford rating estimate is computed using SCOPE, this value is then used with the output of a neural network that is trained on the ground truth Bedford rating from data to obtain an improved estimate as follows,

$$\hat{B}_{est} = B_{est} + b = C\mu^{\gamma} + D + b, \tag{3}$$

where $\hat{B}_{est}$ is our Bedford rating estimate, and b is the learned output of a neural network. A visualization of the approach presented in this paper is shown in Figure 3.

By focusing solely on the UAM Vertical Maneuver MTE, we can use the data to produce the necessary SCOPE parameters specific to this MTE, and to train the neural network. The data input to SCOPE includes the time series data of the pilot's collective stick input, and the vehicle altitude. The data input to the neural network is a sequence of images of the environment that would have been seen through the front window of the cab during the maneuver. This allows us to see if using a neural network with image inputs helps to capture information about pilot workload that isn't captured by the SCOPE inputs.

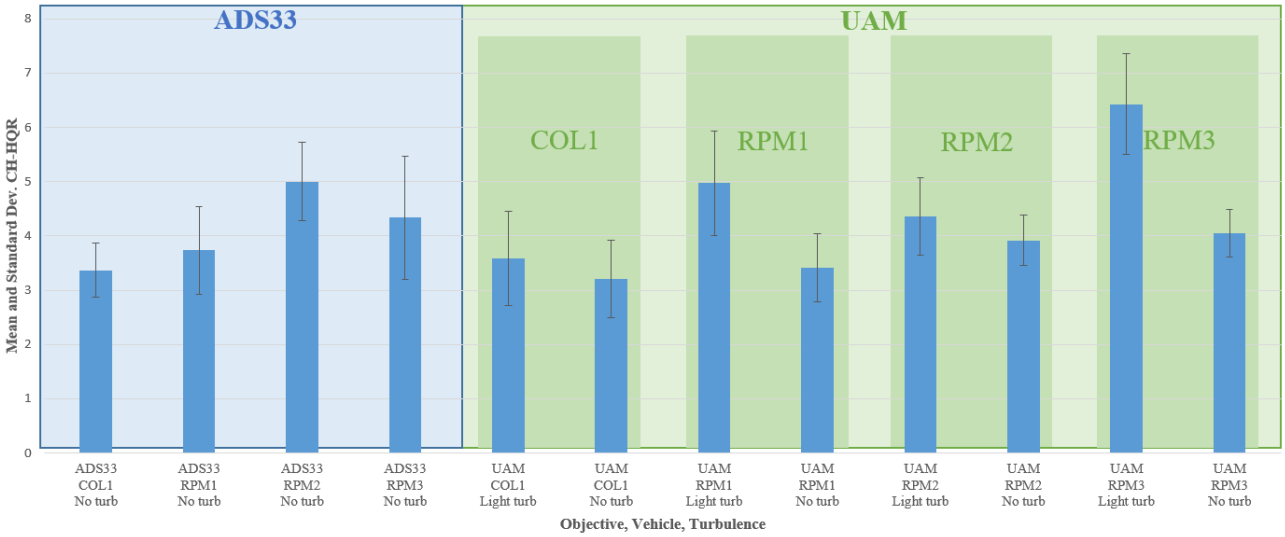

Figure 2. Mean and Standard Deviation of Cooper-Harper Handling Qualities Ratings Vertical Maneuver MTE

Because the current VMS dataset solely contains performance metrics and Cooper-Harper Handling Qualities ratings, we use a mapping defined by Table 2 from Ref. 7 and 8 to convert to Bedford Workload Ratings.

Table 2. Mapping of Performance and Bedford Rating to Cooper Harper Handling Qualities Ratings (Ref. 7, 8).

Performance	Bedford Rating	Cooper Harper Rating
Desired	$1 \leq B < 5$	$1 \leq HQ < 5$
Adequate	$B < 6$	$5 \leq HQ < 6$
Adequate	$6 \leq B < 7$	$6 \leq HQ < 7$
Inadequate	$B < 8$	$7 \leq HQ < 8$
Inadequate	$8 \leq B < 9$	$8 \leq HQ < 9$
Inadequate	$9 \leq B < 10$	$9 \leq HQ < 10$
Inadequate	$B = 10$	$HQ = 10$

SCOPE Approach

The SCOPE approach detailed in Reference 1 depends on the availability of data of several runs of the same MTE. In this work, we implement a form of SCOPE which requires as input time series information of the collective stick input, and vehicle altitude. From this information, we compute a finite difference to obtain the rate of change of this information and use these as detailed in the SCOPE approach.

Because some parameters involved in the SCOPE approach are specific to an MTE, only Vertical Maneuver UAM MTEs are considered, both in cases with turbulence and in cases without. We note that the use of data from a variety of vehicle control law configurations, and turbulence conditions may inject additional noise into our data, as testing showed that some control laws are able to minimize horizontal drift better than others. The parameters that can be adjusted in the SCOPE approach are listed in Table 3, and the corresponding values used in this work are also listed.

Table 3. SCOPE Vertical MTE Parameters

Parameter	Value
ω_μ	2 rad/sec
γ	0.07
C	32.34
D	−28.34

The approach detailed in Reference 1 uses Equation 1 to compute an estimated Bedford rating and is referred to as SCOPE. As shown in Figure 3, there are a few steps that must be taken to get to Equation 1. Firstly, it is necessary to have the stick input rates and position error rates of the aircraft throughout the maneuver. These signals are first filtered by a low-pass filter defined at a specified frequency ω_μ. In our approach, we use $\omega_\mu = 2\ rad/sec$, as this value was used in previous work applied to a Hover MTE (Ref. 8). The resulting signal after the low-pass filter is then a filtered signal defined by the subscripted f. Next, the standard deviation of these signals across time are computed and multiplied with one another to produce the workload stimulus μ. In some works, SCOPE is applied at each time step of the flight, computing a standard deviation across a sliding window (Ref. 21, 22). However, here we only compute the standard deviation for the whole trajectory. In this paper, we tested two values of $\gamma = 0.07$ and $\gamma = 0.6$ (chosen based on previous works) and found that a value of $\gamma = 0.07$ produced higher correlation values. Last of the tunable parameters are constants C and D. These are found by implementing least squares to find the C and D parameters that best fit the estimates to the known Bedford ratings. In References 7 and 8, the authors perform outlier rejection on the data, and use a sample Pearson

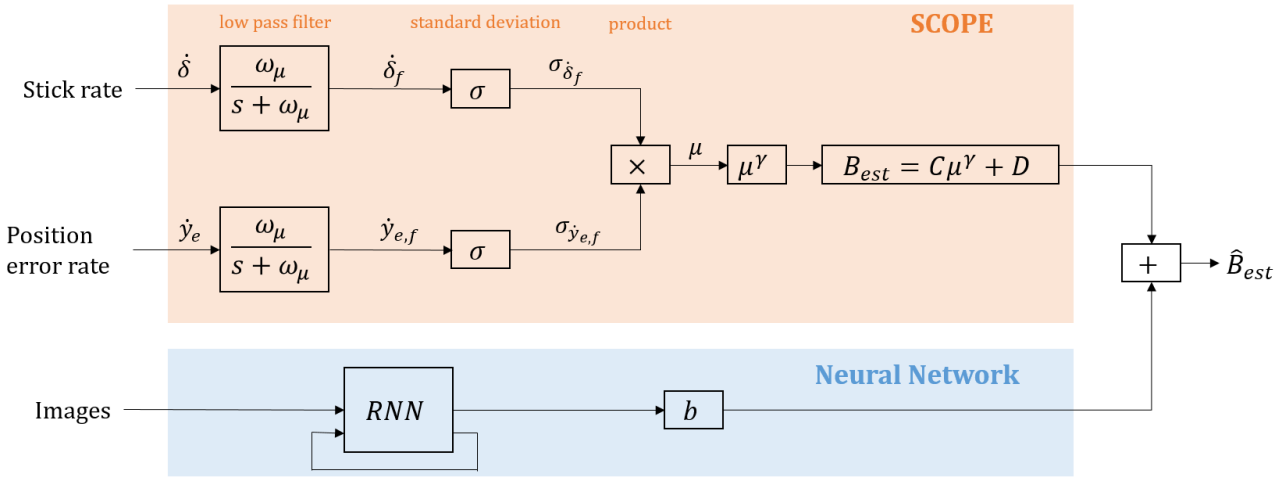

Figure 3. A visualization of our approach, using SCOPE and a Recurrent Neural Network (RNN) to produce an estimate of pilot workload based on stick rate inputs, position error rates, and an image sequence.

correlation coefficient, R, to measure how well the estimates match the data. This coefficient is calculated by

$$R = \frac{\sum_{i=1}^{n}(B_i - \bar{B})(B_{est,i} - \bar{B}_{est})}{\sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}\sqrt{\sum_{i=1}^{n}\left(B_{est,i} - \bar{B}_{est}\right)^2}} \quad (4)$$

where B_i and $B_{est,i}$ are the sample points of the Bedford ratings and the corresponding Bedford estimates, $\bar{B}$ and $\bar{B}_{est}$ are the averages of the Bedford ratings and Bedford estimates, and n is the number of samples. While in Equation 4 we use B_{est} to indicate the SCOPE Bedford estimate, the equation could be applied to compute the correlation coefficient for the SCOPE with Neural Network estimate by replacing this variable with $\hat{B}_{est}$. This Pearson correlation coefficient ranges in values between -1 and 1, where -1 indicates a negative linear correlation, 0 no correlation, and 1 a linear correlation. In Figure 4, we show how the Pearson correlation coefficient changes as the amount of data rejected from an outlier rejection method increases.

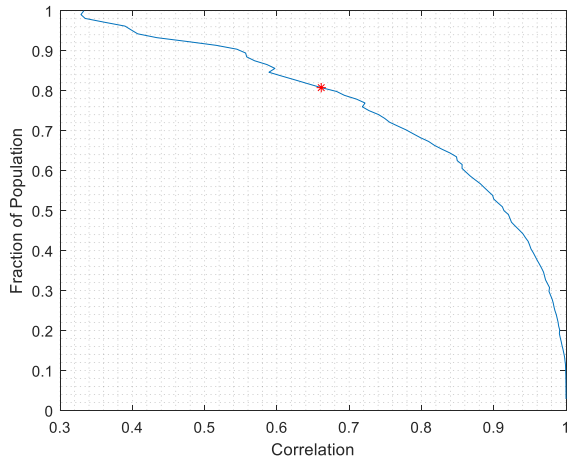

Figure 4. Correlation vs. Fraction of Population

The parameters of C and D were computed at the point marked in Figure 4, where the fraction of the population was 0.8077 (corresponding to 84 samples), and the Pearson correlation coefficient was 0.66.

This correlation coefficient was lower than expected but could be attributed to the variety of vehicle control system configurations tested, as well as the presence of turbulence in only a subset of the data points. We choose to use this varied dataset to keep as many data points as possible to create a single model to estimate workload ratings for the UAM Vertical Maneuver MTE.

RNN Approach

The first step in constructing the proposed learning-based approach is to ensure that the image data is suited to be used as input to the neural network. To do this, we use existing video logs of the VMS flight tests taken at 30 frames per second, convert these to still images, and crop the relevant views. Samples of such images (nonconsecutive) are shown in Figure 5. Here we can also see the visual cues seen by the pilot, where the goal pose can be found by aligning the small black box in the center of the larger white box at both the top and bottom of the maneuver.

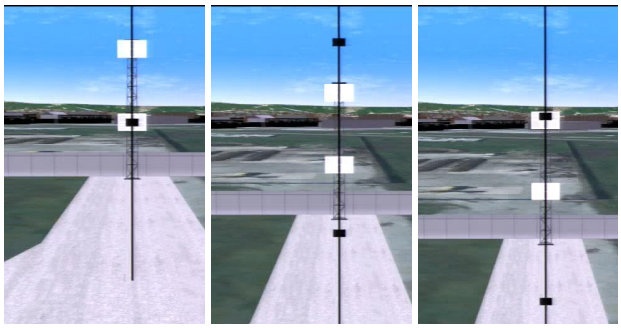

Figure 5. Sample Images for Neural Network input

A lower sampling rate of 15 frames per second was used instead of the original 30 frames per second to reduce the number of still image inputs. To further reduce the size of the input images, we first grayscale the images before feeding through the network. From our dataset, we consider runs of the Vertical Maneuver MTE UAM performance standard for all vehicle configurations and both turbulence levels. Additionally, we consider only runs in which there is desired performance, and Cooper Harper ratings range from one to five, to reflect a one-to-one mapping between Cooper-Harper and Bedford Rating scales in Table 3. The set of runs with these conditions contains 73 total samples. This value is lower than the number of samples used to compute the SCOPE parameters because video feed was not recorded for all MTE runs.

Once the data is prepared for training, it is separated into a training set and a validation set. In our implementation, we use 47 datapoints for training, and 26 datapoints for validation. This allows us to test whether the neural network portion of our model performs well on data unseen during training. Training is conducted in a supervised fashion, where a series of images and raw inputs are used as input. We use SCOPE (Ref. 1) to generate a preliminary workload estimate. The output of the neural network is added to this value, and a Huber loss computed on the resulting sum:

$$loss = \begin{cases} 0.5\left(B_i - \hat{B}_i\right)^2, & if\ |B_i - \hat{B}_i| < 1 \\ |B_i - \hat{B}_i| - 0.5, & otherwise \end{cases} \quad (5)$$

In our loss function, i is the sample index, B is the ground truth Bedford rating from the VMS data, and $\hat{B}_{est}$ is the output from our combined SCOPE and Neural Network method. To minimize this equation, the Adam optimizer (Ref. 23) is used on the weights of the neural network.

Due to the sequential nature of our input data, a Recurrent Neural Network in the form of a Long-Short Term Memory (LSTM) block is used to track the time varying information of the data across each maneuver. Figure 6 shows that the Neural Network is comprised of a Convolutional Neural Network (CNN) that processes each individual image input, that then passes the output to a max pool layer and LSTM block. In training, each image in the sequence is passed through the network, and only the final output value is trained on the target Bedford rating from data. This network architecture is similar to that of Ref. 24 which also requires collecting time varying information in training.

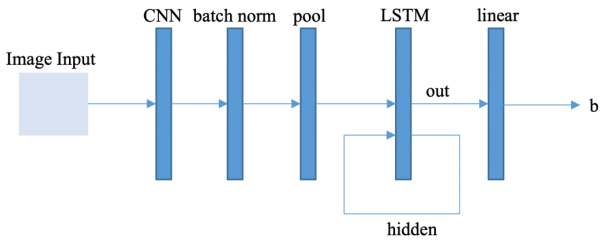

Figure 6. Neural Network Architecture

In training we used a decaying learning rate, where three learning rates of $\alpha \in [0.01, 0.001, 0.0001]$ are used, starting with the largest value. Additionally, we chose to divide the training data into batches of 10 datapoints and run for 100 epochs per learning rate. In this work, we used a single output channel from the CNN with a Rectified Linear Unit (ReLU) activation function, and a max pool function with kernel size 4.

Ultimately, to test whether our method generates predictions with good linear correlation to the true Bedford Ratings, we generated correlation plots between SCOPE, pilot Bedford ratings, and our method. Correlation metrics allowed us to determine whether our method outperforms SCOPE alone.

RESULTS

After training the neural network as described in the previous section, the resulting training loss and validation loss decreased as shown in Figure 7. The goal was to see if our network was learning generalizable information from the image sequence training data. For this reason, we tested the loss and correlation of our method on two different datasets: a training dataset and a validation dataset. Only the training dataset was used during the training process, while the validation data was unseen until test time.

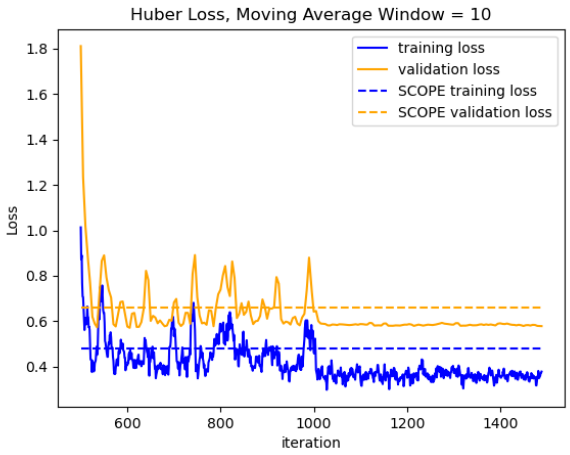

Figure 7. Training and validation loss of a neural network with SCOPE

Here we show the last few hundred iterations of the training and validation loss, along with the value of the loss obtained using only SCOPE for each subset of the data. We see that the loss metric of using SCOPE alone is higher for both the training and validation subsets of the data. The full loss plot for all iterations can be seen in Figure 13 in the Appendix section. With this trained network, we also tested whether using this neural network with SCOPE would produce a higher correlation coefficient to the data than SCOPE alone. Seeing a higher correlation coefficient with the validation dataset would tell us if the network was overfitting to the training data instead of generalizing the relationship between image sequences and pilot workload ratings. In Figure 8 we show the correlation with the validation data of our method vs. using SCOPE alone.

As shown by the correlation coefficients in Figure 8, we found that our method was able to produce a higher correlation coefficient on the validation data with R = 0.148 compared to SCOPE alone, which produced a correlation coefficient of 0.125. This increase, along with the reduced loss value for both training and validation suggest that the image inputs and neural network help to improve the Bedford workload estimate of SCOPE. In order to further explore this approach, we perform cross validation and consider training a neural network without SCOPE.

Cross Validation

In this section, we verify whether the results presented in the preceding section were dependent on the data separation between training and validation. While our results showed that using image inputs to an RNN with SCOPE produced higher correlation coefficients and lower loss values both for the training data and validation data, we want to test whether this would continue to be the case for other groupings of data. For this reason, we ran the same test on a case where a different split of training and validation data was chosen. We

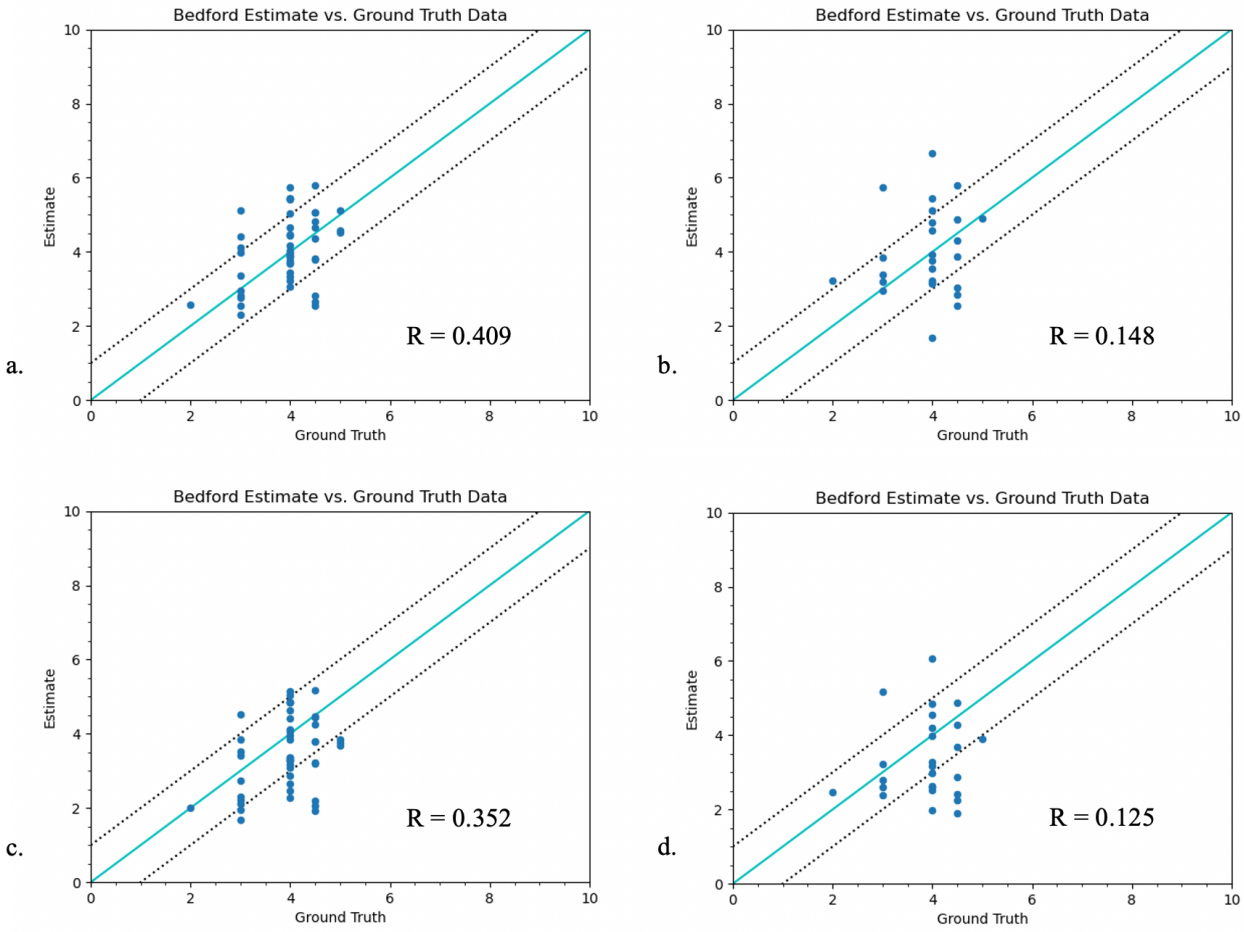

Figure 8. Correlation plots and coefficients (-1 < R < 1) between estimate methods for training and validation datasets. a) SCOPE with NN on training data, b) SCOPE with NN on validation data, c) SCOPE alone on training data, d) SCOPE alone on validation data.

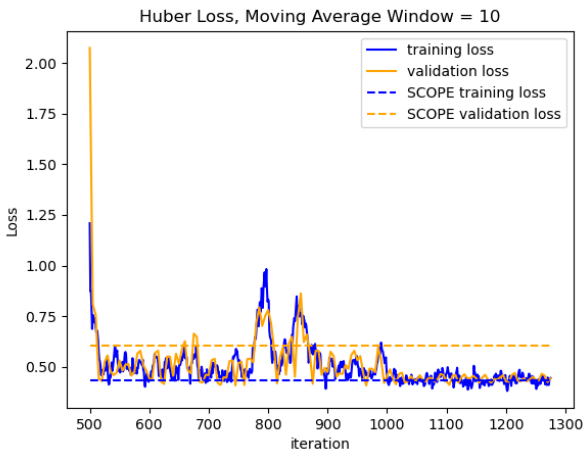

Figure 9. Training and validation loss of neural network with SCOPE with Data Split 2

will refer to this split of data as Data Split 2, and the first split of data as Data Split 1. Each of these data splits have the same amount of data points in training and validation. The loss plot for the last few iterations of this case is shown in Figure 9, while the full plot of all iterations is shown in Figure 14 in the Appendix section.

From changing how the data was split, we can see that while the training loss converges to a value very similar to that of SCOPE, the validation loss did produce a lower loss than SCOPE alone. Again, we also present the linear correlation between the ground truth labels and the estimated Bedford ratings in Figure 10 for this case.

From Figure 10, we see a slight improvement in the training correlation coefficient and slight reduction in the validation correlation coefficient. This test suggests that the initial performance of the approach to using SCOPE with a neural network showing improvements in both linear correlation coefficients and loss may be dependent on how the data was split, and that a closer look at the types of data in each training and validation split should be made in future work.

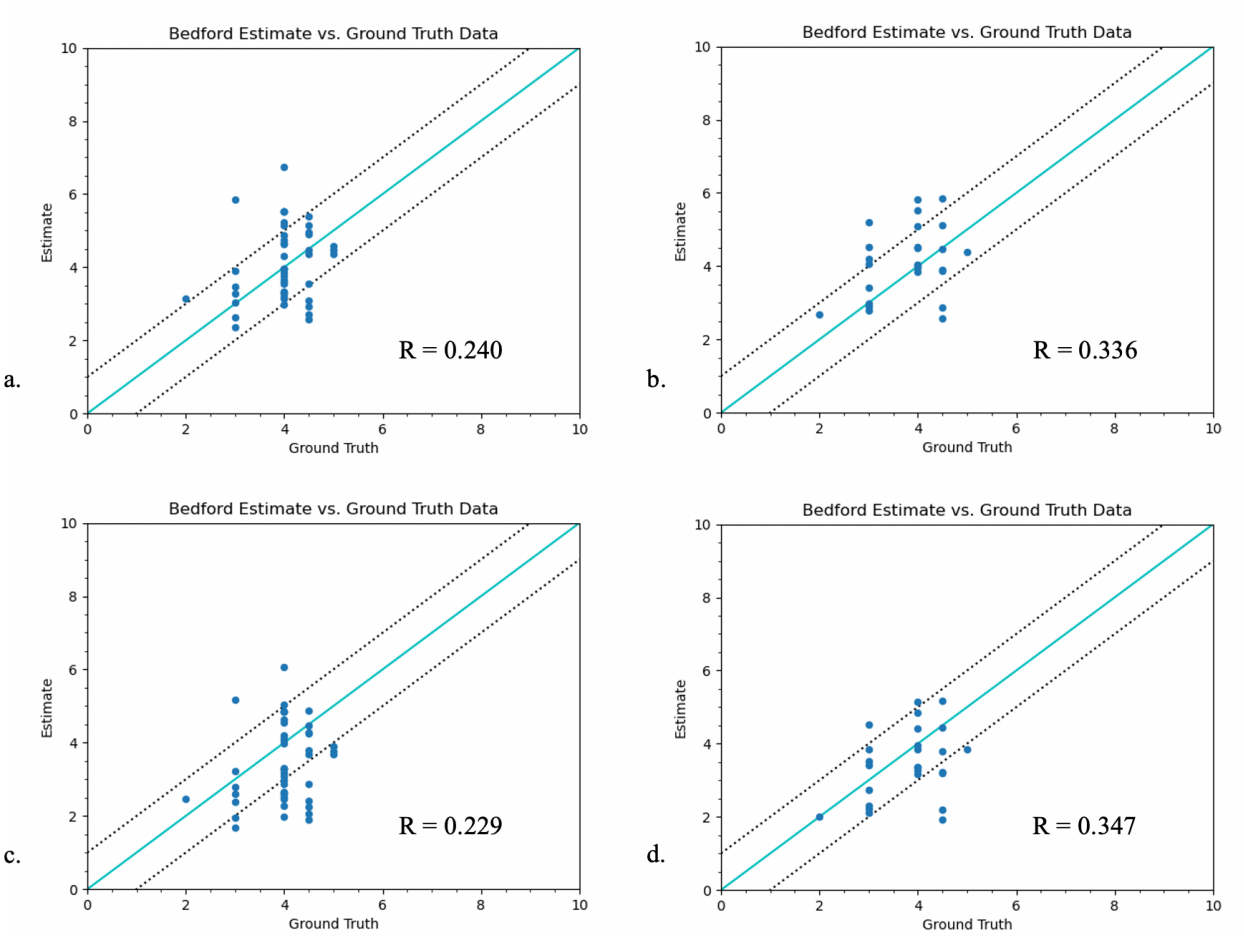

Figure 10. Correlation plots and coefficients (-1 < R < 1) between estimate methods for training and validation datasets. a) SCOPE with NN on training data from Data Split 2, b) SCOPE with NN on validation data from Data Split 2, c) SCOPE alone on training data from Data Split 2, d) SCOPE alone on validation data from Data Split 2.

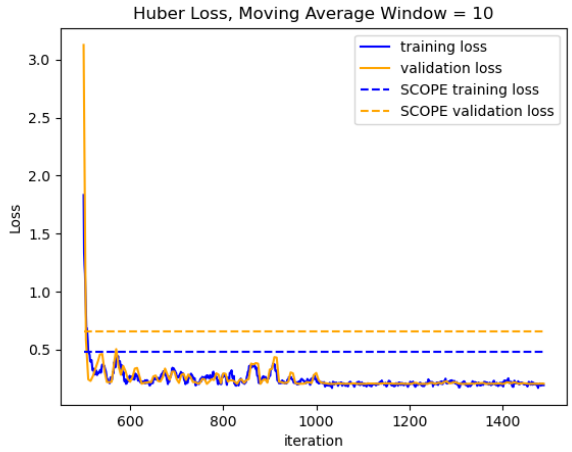

Figure 11. Training and validation loss of neural network alone

Effect of SCOPE

Another variation we consider is the effect of using the SCOPE approach from Reference 1. The motivation for using a learning-based approach with SCOPE was based on the small amount of data available for training. The idea was to use the image data to capture unmodeled dependencies of workload ratings to the visual scene of the environment. To see if this approach was necessary, we also compared our method of using SCOPE with a neural network to a neural network alone.

In Figure 11, we can see that the training and validation loss are both lower than the SCOPE approach, which may suggest that using a neural network alone might produce improved results compared to using a combination of both a Neural Network and SCOPE. However, a further look at the linear correlation coefficient plots shown in Figure 12 allows us to see that this neural network trained without SCOPE learns to output a constant value of 5 for each input from our data. A plot of the loss for all iterations is shown in the Appendix section in Figure 15.

While lower training and validation losses are generally considered to be good, we can see here that it is important to also continue to consider the linear correlation coefficients,

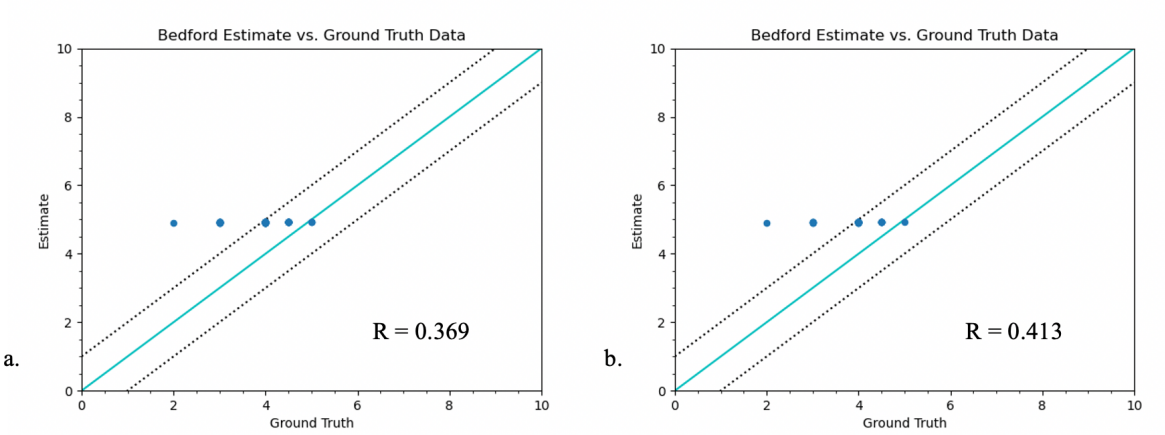

Figure 12. Correlation plots and coefficients (-1 < R < 1) between estimate methods for training and validation datasets. a) NN on training data from Data Split 1, b) NN on validation data from Data Split 1.

and generally the outputs of the network. A model which outputs the same value no matter the input is not particularly helpful in determining the relationship between the experience of the pilot during flight and their workload ratings. We can also observe that using a neural network alone produces a lower correlation coefficient than using a neural network with SCOPE, suggesting that using SCOPE with a Neural Network does aid in the model's performance.

A summary of the results from our tests are shown in Table 4. Initial results in the first original test case show performance improvements compared to using SCOPE alone in both loss metrics and linear correlation metrics. However, a cross validation showed that this was not necessarily the case across different splits of training and validation data, and that further investigation is necessary. Lastly, we see a benefit in using SCOPE with a Neural Network, where we see that the network alone learns to directly output a constant value regardless of the input.

CONCLUSIONS

In this paper we present a method to estimate pilot workload ratings by augmenting an existing method with a recurrent neural network. We showed that using SCOPE with a recurrent neural network with the architecture shown in Figure 6 with inputs of image sequences helped to improve the correlation metric after training for 100 epochs per learning rate with a data batch size of 10. Further, we implemented cross validation and found that there is need for further investigation in the relationship between how the data is split and the performance of this approach. Lastly, we showed that our method in using SCOPE with a Neural Network produced a more informative model than using a Neural Network alone.

We also found that while the use of a neural network improved correlation coefficients in the original case, the overall correlation coefficients themselves were relatively low for SCOPE overall when all data points in the dataset were used. This generally lower correlation coefficient could be due to the diversity of the data that was collected, amongst several pilots for four different vehicle control law configurations in two different turbulence modes. This diversity in data may also contribute to the result found in the cross validation, where additional inputs to the estimation model may be required to differentiate these different attributes in the data (i.e. turbulence level, vehicle configuration, etc.).

Table 4. Correlation Coefficients of SCOPE with trained RNN, and SCOPE alone for training and validation datasets.

Test Case	Dataset Train / Val Split	Method	Training Set	Validation Set
Original	Dataset Split 1	SCOPE w/ RNN	$R = 0.409$	$R = 0.148$
		SCOPE	$R = 0.352$	$R = 0.125$
Cross Validation	Dataset Split 2	SCOPE w/ RNN	$R = 0.240$	$R = 0.336$
		SCOPE	$R = 0.229$	$R = 0.347$
Neural Network Alone	Dataset Split 1	RNN	$R = 0.369$	$R = 0.413$
		SCOPE	$R = 0.352$	$R = 0.125$

FUTURE WORK

In future work, other inputs could be tested using a Neural network, to determine if there are other metrics that might help to further increase the correlation between estimated pilot workload ratings and the data. Some inputs that could be used are a known turbulence level or vehicle configuration. Another interesting relation to explore is the pilot induced oscillation, and how this may play into handling qualities and workload ratings.

Further, the downside to using a neural network to define the relationship between the image inputs and workload estimates is that relevant contribution of the image inputs to the improved workload estimates are not explicitly known. Future work in investigating the explainability of the convolutional neural network could be tested, to determine what parts of the images are most relevant to the neural network.

Author contact: Keiko Nagami keiko.nagami@nasa.gov
Carlos Malpica carlos.a.malpica@nasa.gov
Mac Schwager schwager@stanford.edu

APPENDIX

In this section we include figures of the loss plots for all iterations from each of the test cases that were run in the Results section. The first few iterations are omitted in Figures 7, 9, and 11 in the Results section for clarity.

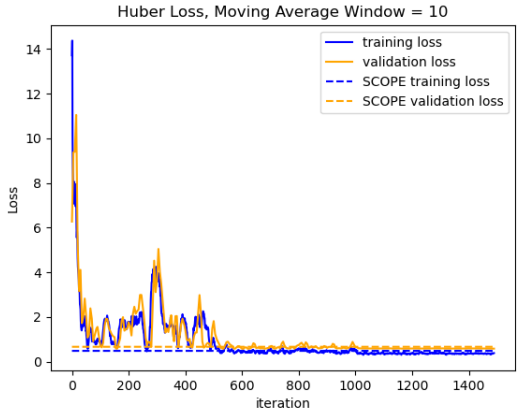

Figure 13. Full training loss plot for all iterations of first test case.

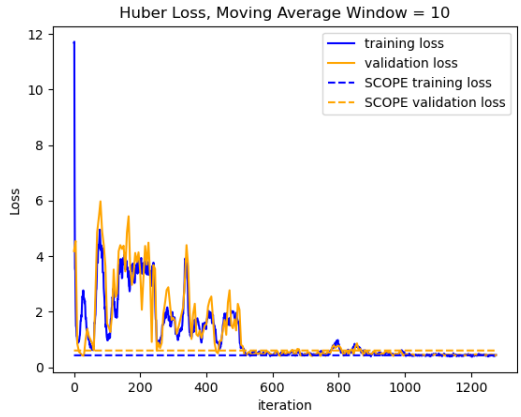

Figure 14. Full training loss plot for all iterations of cross validation test case.

Figure 15. Full training loss plot for all iterations of neural network only test case.

ACKNOWLEDGMENTS

The authors give thanks to Dr. Edward Bachelder for improving our understanding of SCOPE, to Ethan Romander and Kristen Kallstrom for their help in the network training process, and to Stefan Schuet, Allen Ruan, Jeremy Aires, and Shannah Withrow-Maser for VMS data collection and analysis.

Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Advanced Supercomputing (NAS) Division at Ames Research Center.

REFERENCES

1. Bachelder, E., "SCOPE-Pilot Workload Estimation Using Control Response: Theoretical Development and Practical Demonstration," AIAA Scitech 2020 Forum, Orlando, FL, Jan. 6-10, 2020.

2. Anon., "Handling Qualities Requirements for Military Rotorcraft, ADS-33E-PRF," U.S. Army Aviation and Missile Command, Mar, 2000.

3. Roscoe, A. H., and Ellis, G. A., "A subjective rating scale assessing pilot workload in flight. A decade of practical use. Royal Aerospace Establishment," Technical Report 90019, Farnborough. UK: Royal Aerospace Establishment, 1990.

4. Cooper, G. E. and Robert P. Harper, J., "The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities," Tech. Rep. NASA TN D-5153, National Aeronautics and Space Administration, April 1969.

5. Roscoe, M., and Wilkinson, C., "DIMSS-JSHIP's M&S Process for Ship/Helicopter Testing & Training," AIAA Modeling and Simulation Technologies Conference and Exhibit, Aug. 2002.

6. Bachelder, E., Berger, T., Godfroy-Cooper, M., Aponso, B., "Pilot Workload and Performance Assessment for a Coaxial-Compound Helicopter and Tiltrotor During Aggressive Approach," Vertical Flight Society's 77th Annual Forum Proceedings, Virtual, May 2021.

7. Bachelder, E., Lusardi, J., Aponso, B., Godfroy-Cooper, M., "Estimating Handling Qualities Ratings from Slalom Flight Data: A Psychophysical Perspective," Vertical Flight Society's 76th Annual Forum Proceedings, Virtual, Oct. 2020.

8. Bachelder, E., Lusardi, J., Aponso, B., "Estimating Handling Qualities Ratings from Hover Flight Data Using SCOPE," AIAA Scitech 2021 Forum, Virtual, Jan. 2021.

9. Bachelder, E., and Aponso, A., "A Theoretical Framework Unifying Handling Qualities, Workload, Stabilize, and Control," Vertical Flight Society's 77th Annual Forum Proceedings, Virtual, May 2021.

10. Bachelder, E., and Aponso, B., "Novel Techniques for Characterizing and Testing Aircraft Handling Qualities," Vertical Flight Society's Rotorcraft Handling Qualities Technical Meeting, Huntsville, AL, Feb. 18-19, 2020.

11. Godfroy-Cooper, M. Sarrazin, J. C., Bachelder, E., Miller, J. D., and Denquin, F., "Influence of Optical and Gravito-inertial Cues to Height Perception During Supervisory Control," Vertical Flight Society's 76th Annual Forum Proceedings, Virtual, Oct. 2020.

12. Godfroy-Cooper, M. Sarrazin, J. C., Bachelder, E., Miller, J. D., Denquin, F., Bardy, B., "Visual-Gravitoinertial Interactions for Altitude Perception During Manual and Supervisory Control," Vertical Flight Society's 77th Annual Forum Proceedings, Virtual, May 2021.

13. McRuer, D. T., and Jex H. R., "A Review of Quasi-Linear Pilot Models," *IEEE Transactions on Human Factors in Electronics*, Vol. HFE-8, (3), Sept. 1967, pp. 231-249. DOI: 10.1109/THFE.1967.234304.

14. Hess, R. A., and Chan, K. K., "Preview control pilot model for near-earth maneuvering helicopter flight" *Journal of Guidance, Control, and Dynamics,* Vol. 11, (2), 1988, pp. 146-152.

15. Zaal, P., and Sweet, B., "Identification of Time-Varying Pilot Control Behavior in Multi-Axis Control Tasks," AIAA Modeling and Simulation Technologies Conference, Minneapolis, MN, Aug. 2012.

16. Shi, G., Hönig, W., Yue, Y., and Chung, S. J., "Neural-swarm: Decentralized Close-Proximity Multirotor Control Using Learned Interactions," 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020.

17. Zeng, A., Song, S., Lee, J., Rodriguez, A., and Funkhouser, T., "Tossingbot: Learning to Throw Arbitrary Objects with Residual Physics," *IEEE Transactions on Robotics*, Vol. 36, (4), 2020, pp. 1307-1319.

18. Ajay, A., Wu, J., Fazeli, N., Bauza, M., Kaelbling, L. P., Tenenbaum, J. B., and Rodriguez, A., "Augmenting Physical Simulators with Stochastic Neural Networks: Case Study of Planar Pushing and Bouncing," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2018.

19. Aponso, B. L., Tran, D. T., Schroeder, J. A., and Beard, S. D., "Rotorcraft Research at the NASA Vertical Motion Simulator," Paper AIAA 2009-6056, AIAA Atmospheric Flight Mechanics Conference, Chicago, IL, August 10–13, 2009.

20. Withrow-Maser, S., Aires, J., Malpica, C., and Schuet, S., "Handling Qualities Evaluations of Urban Air Mobility (UAM) eVTOL Disturbance Rejection and Control Response Criteria Using the Vertical Motion Simulator," VFS Aeromechanics for Advanced Vertical Flight Technical Meeting, San Jose, CA, Jan 25-27, 2022.

21. Bachelder, E., and Godfroy-Cooper, M., "Pilot Workload Estimation: Synthesis of Spectral Requirements Analysis and Weber's Law," AIAA Scitech 2019 Forum, 2019.

22. Bachelder, E., Godfroy-Cooper, M., and Bimal, A., "Interplay Between Optic Flow, Pilot Workload and Control Response During Aggressive Approach to Hover Maneuvers for Three Vertical Lift Vehicle

Models," AHS International 75th Annual Forum, Philadelphia, PA, USA. 2019.

23. Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).

24. Wang, S., et al. "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017.