The Stability of Human Supervisory Control Operator Behavioral Models Using Hidden Markov Models

Haibei Zhu and Mary L. Cummings

Abstract— **Human supervisory control (HSC) is a widely used knowledge-based control scheme, in which human operators are in charge of planning and making high-level decisions for systems with embedded autonomy. With the variability of operators' behaviors in such systems, the stability of an operator modeling technique, i.e., that a modeling approach produces similar results across repeated applications, is critical to the extensibility and utility of such a model. Using an unmanned vehicle simulation testbed where such vehicles can be hacked, we compared two operator behavioral models from two different experiments using a hidden Markov modeling (HMM) approach. The resulting HMM models revealed operators' dominant strategies when conducting hacking detection tasks. The similarity between these two models was measured via multiple aspects, including model structure, state distribution, divergence distance, and co-emission probability distance. The similarity measure results demonstrate the stability of modeling human operators in HSC scenarios using HMM models. These results indicate that even when operators perform differently on specific tasks, such an approach can reliably detect whether strategies change across different experiments.**

I. INTRODUCTION

Human supervisory control (HSC) [1] has been widely utilized in various human-automation collaboration scenarios, including remote surveillance, search, rescue applications [2]. In HSC scenarios, human operators interact with autonomous systems to receive and send high-level decision commands, such as path management for single-operator multi-drone control systems [3], [4].

Many challenges and potential problems exist in HSC scenarios [5], including how to achieve optimal operator performance and maintain effective problem-solving strategies [6]. While improving operators' strategies is critical, they are not directly observable as there are only intermittent interactions between operators and HSC control interfaces. Thus, operator behavioral models are needed to investigate and improve operators' control strategies.

Understanding that the descriptive and predictive quality of an operator behavioral model is directly related to the quantity and quality of the training data [7], the variability of operators' behavior patterns and the limited amount of collected actions in an HSC setting makes it difficult to interpret results from repeated scenarios. Thus, the stability of a behavioral modeling technique is critical for investigating operators' strategies in HSC scenarios, especially for engineers trying to find critical points of intervention. In this paper, we demonstrate the stability of the hidden Markov modeling approach in HSC scenarios by utilizing multiple similarity measurements to compare HMM models developed from known HSC scenarios.

This paper is organized as follows. Section II provides the background of HMMs and several HMM model similarity measurement metrics. Section III describes the human-in-the-loop experiment design for two different experiments using the same interface, while Section IV presents the resulting operator behavior HMM models. The HMM similarity measurement results are illustrated in Section V. Section VI concludes this paper with a detailed discussion and potential future research directions.

II. BACKGROUND

Many operator behavioral modeling techniques with different mathematical representations have been proposed [8]–[10]. Hidden Markov models have been utilized to investigate operators' strategies in many HSC scenarios since they focus on modeling the interactions and cooperation between operators and digital systems [11], [12]. An HMM has a two-layer structure, including a hidden state layer and an observation layer [13]. Thus, a weighted cluster of operators' actions can be considered as an abstract behavioral/functional group, which can be represented by a hidden state. Similarly, an HMM observation can represent an observable interaction between operators and interfaces of autonomous systems.

The HMM structure can be defined as a tuple [14]:

$$\lambda = \{S, O, A, B\}.$$

In such a notation, $S = \{S_1, S_2, ..., S_N\}$ represents N different hidden states, while $O = \{O_1, O_2, ..., O_M\}$ represents M different observations. Transition and emission probabilities of an HMM model connect its hidden states and observations. $A = \{a_{ij}\}$ is a $N \times N$ transition probability matrix, in which $a_{ij} = P\{S_j^{t+1} | S_i^t\}$, $i, j = 1, 2, ..., N$, and $B = \{b_{ik}\}$ is a $N \times M$ emission probability matrix, in which $b_{ik} = P\{O_k | S_i\}$, $i = 1, 2, ..., N$, $k = 1, 2, ..., M$, with both $a_{ij}, b_{ik} \geq 0$. In an HMM model, the current system state transfers among hidden states sequentially based on a transition probability matrix. As mentioned above, each hidden state in an HMM model can be considered as a combination of weighted observations, which are represented by the emission probability matrix.

The robustness of an operator behavioral modeling technique highly depends on the degree of variance of operators' performance and the quantity of operators' action data [15]. In order to present the stability of a modeling technique,

This work was supported by the ONR under agreements number N00014-17-1-2012 and N00014-17-1-2504.

The authors are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA (haibei.zhu, m.cummings)@duke.edu

measurements are necessary to show to similarity level between models. Many HMM similarity measurement methods have been proposed for various applications, including the investigation of HMM development, HMM classification, and sensitivity tests on HMM model parameters [16]–[18].

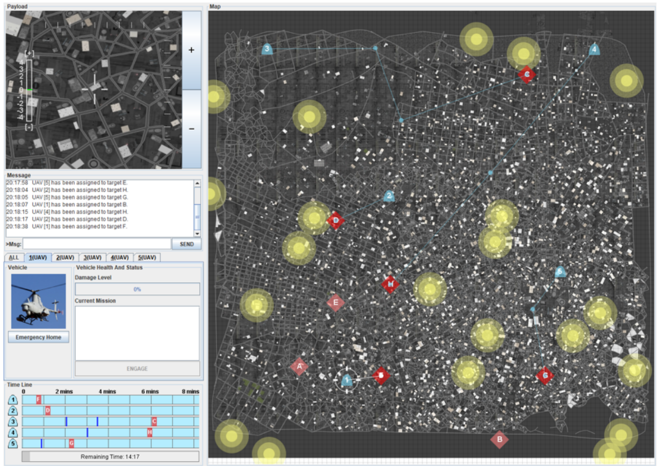
In addition to the direct comparison of state interpretations and model structures between HMM models, two HMM model similarity measurement methods have been utilized to quantitatively investigate the similarity between HMM models, including the divergence distance [19] and co-emission probability distance measure [18]. Specifically, the divergence measure focuses more on the probability aspect of different HMM models applied to the same given data sequences, while the co-emission measure emphasizes on the quantitative distance between HMM model vectors in a high dimensional space. While there has been no such application of similarity metrics to models that represent human behavioral states, we propose that similar similarity measurement results can demonstrate the stability of modeling operators' strategies across HSC scenarios using HMM models.

We define HMM stability as a model's ability to produce similar results across repeated applications with different operators. Because of significant human variability due to individual differences, operator models are difficult to generalize across repeated applications of the same HSC interface. For example, when a drone operator supervises his or her system, task-based models of an operator are likely to differ widely between two or more people. Moreover, such models can vary widely when a person operates a drone one day and then again on a separate day. If a model fails to generalize when applied to different people or even under slightly different conditions, such a model will not be useful outside very narrow circumstances. To this end, we explored how stable the HMM approach is in an HSC context when applied to different sets of people.

III. EXPERIMENT - DATA GENERATION

Two human-subject experiments were conducted using the Security-Aware Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU-SA) experiment platform to collect operators' action data for developing operator behavioral models. The RESCHU-SA is a simulation-based experiment platform for a single-operator with multi-UAV supervisory control scenarios [20]. It provides multi-tasking scenarios, including both UAV navigational and imagery analysis tasks. This platform can also simulate UAV GPS spoofing attacks, in which a hacker system generates counterfeit GPS signal to deceive UAV operators and navigate hacked UAVs to unexpected destinations, along with real or false system notifications.

The interface of the RESCHU-SA platform is shown in Figure 1. Operators mainly focus on two components during the experiment, including the map area and the payload camera view. The map area, which occupies the majority of the interface on the right, displays the surveillance area with real-time information of all UAVs, targets, and hazards areas. Operators manage and navigate all UAVs on the map, avoiding all hazard areas. The underlying UAV-target assignment algorithm is intentionally suboptimal, requiring the operator to pay attention and fix suboptimal assignments. Suspected GPS spoofing attacks are signaled to operators through a pop-up window. Once an operator receives a notification of a potential hacking attack, the operator then investigates by checking the UAV payload camera view and matches it against the position of the UAV on the map.


20:17:58 UAV [5] has been assigned to target E.
20:18:04 UAV [2] has been assigned to target H.
20:18:05 UAV [5] has been assigned to target G.
20:18:07 UAV [1] has been assigned to target B.
20:18:15 UAV [4] has been assigned to target H.
20:18:17 UAV [2] has been assigned to target D.
20:18:38 UAV [1] has been assigned to target F.

Damage Level: 0%

Remaining Time: 14:17

Fig. 1. RESCHU-SA experiment platform interface

In general, three different high-level tasks are assigned to operators using the RESCHU-SA interface, including 1) the navigation task - to ensure all UAVs avoid with hazards areas, 2) the imagery task - to perform reconnaissance tasks in the payload camera view when UAVs reach assigned targets, and 3) the hacking detection task - to determine whether UAVs are under GPS spoofing attacks. Specifically, in hacking detections, we assume that false alarms and detection failures exist in the autonomous detection system [21], [22], and the experiment platform provides both situations for operators. Utilizing human visual advantages, we have proposed that humans can assist an autonomous system in determining if it has been hacked by investigating the potential differences between the location interpreted from the camera view and the GPS reported location on the map [23]. Two similar RESCHU-SA experiments were conducted to collect operators' action data for developing operator behavioral models.

A. The First Experiment

The first experiment was conducted to determine 1) whether human operators could assist autonomous detection systems in determining UAV hacking events, 2) what general hacking detection strategies were exhibited by operators, and 3) whether different objective task load levels could affect operators' performance and strategies. The full experiment details and results can be found in other publications [23], [24]. In this experiment, 36 participants performed two test scenarios at a high and a low level of tasking, counterbalanced across participants. Participants' performances were not statistically different between the two different task load levels.

B. The Second Experiment

A second experiment was conducted to evaluate the utility of HMM-based operator behavioral models for reducing operator inefficiencies in hacking detection tasks and to investigate the potential use of such models in providing additional assistance for operators in hacking detection tasks. While there were multiple conditions in this experiment, this paper will examine just the data set that replicated the conditions of the first experiment. In this portion of the experiment, 45 participants, completely different from the first experiment, completed the same scenario with the same experimental settings.

Between these two experiments, participants' performance in the three major tasks mentioned in the previous section was directly compared using t-tests with a significance level of $alpha = 0.05$. The statistical analyses show a marginally significant difference in the hacking detection success rates between the two experiments ($p = 0.092 > 0.05$), and a statistical difference between the success rates in imagery task identification ($p = 0.027 < 0.05$) and the UAV damage levels caused by encountering hazards areas ($p < 0.001$) present statistical differences. Given the same experimental settings and different groups of participants in these experiments, we considered that the differences in imagery and navigation task performance were likely caused by the variance of participants' subjective strategies. Recall that there was no difference in performance when the same participants repeated the first experiment.

Thus, in the following sections, we focus on developing operator behavioral models of the hacking detection task from these two experiments and comparing models to investigate the similarity between operators' strategies and the stability of this HMM modeling technique in HSC scenarios.

IV. OPERATOR BEHAVIORAL MODELS

Participants' actions, including keystrokes and mouse actions, were collected in both experiments and categorized as one of ten observations, as shown in Table I. With a focus on studying aggregate operators' strategies in hacking detection tasks, we developed two hacking detection HMM models based on detection tasks data sequences.

TABLE I

HMM OBSERVATIONS FROM THE RESCHU-SA INTERFACE

1 Add waypoint	2 Move waypoint	3 Delete waypoint
4 Move endpoint	5 Switch target	6 Engage task
7 Monitor UAV	8 Perceive hacking	9 Detection decision
10 Adjust zoom level		

Both models were trained using the multi-sequence Baum-Welch algorithm, which is an unsupervised HMM training method [25]. HMM model training results were selected based on the number of hidden states using the Bayesian Information Criterion (BIC), which balances the increase of the model complexity by penalizing the number of free model parameters [26]. Thus, the resulted models are expected to achieve both high model likelihood fitting on training data and reasonable structures.

A. The First Experiment Operator Hacking Detection Behavioral Model

The first operator hacking detection behavioral model was developed using detection observation sequences from both scenarios of the first experiment since they were statistically not different. As shown in Figure 2, the resulting model is a 6-state HMM model with a "Start" and an "End" state. In all hacking events, suspected hacking notifications are provided by the system, and an operator must clear a screen indicating that he or she has seen this alert, which indicates the start status of a detection event. Similarly, operators' final decisions as to whether the UAV was actually hacked indicate the end status of a detection event. Thus, the state transitions in Figure 2 beginning from the "Start" state and ending at the "End" state illustrate operators' behavioral paths for hacking events.

Each state can be described as a weighted cluster of the ten observations, and such a weighted combination can be visualized as emission probabilities using a histogram, shown in Figure 3. Different states were interpreted based on their different dominant observations. The state interpretations in Figure 2 are 1) the "Start" state, in which operators received hacking notification; 2) "Monitor UAV", in which operators predominantly monitored the camera view of the selected UAV; 3) "Adjust target (manage UAVs)", in which operators reassigned a UAV's target because the pairing was suboptimal; 4) "Engage task", in which operators engaged in imagery counting tasks; 5) "Adjust waypoint (navigate UAVs)", in which operators changed the moving trajectory of the selected UAV to make it pass over something of interest on the ground which would be salient in the payload camera; and 6) the "End" state, in which operators reported their final decisions on hacking events. The occurrence frequency and percentage of all hidden states are also calculated by applying the HMM model to the original training data sequences using the Viterbi algorithm [27].

According to the state transition probabilities, two major operational paths representing operators' dominant behavioral flows in hacking detections can be observed. The first path starts at the "Start", passes through the "adjust waypoint" and ends at the "End" state. This path demonstrates a possible detection strategy of manipulating UAVs' trajectories to proactively investigate the potential difference between the UAV camera view and its GPS reported location during a hacking event using human geo-location. The second path also starts at the "Start", passes through "adjust target" and "monitor UAV", then ends at the "End" state. This path indicates a reactive strategy where people waited for UAVs to approach interesting points. They also typically completed the secondary task of managing UAVs' assigned targets while waiting.

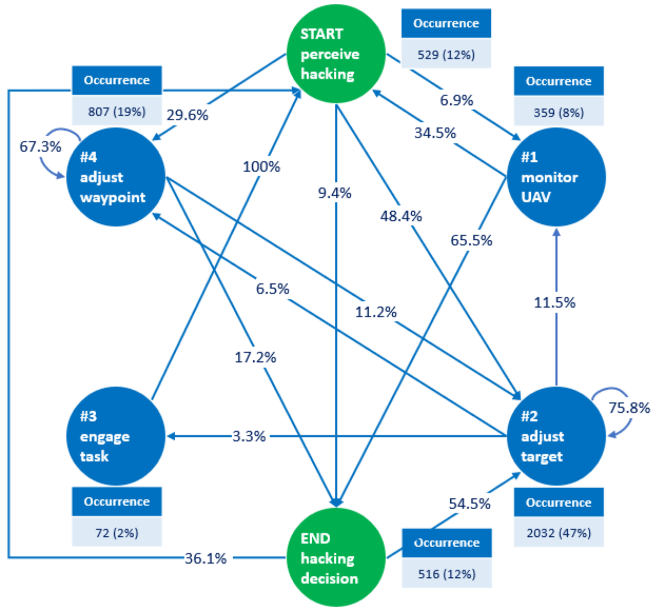


Fig. 2. The first experiment operator hacking detection behavioral model

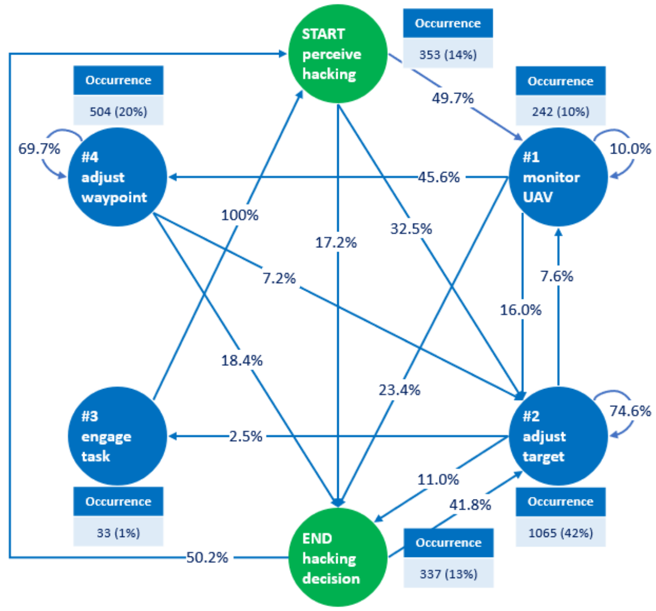


Fig. 4. The second experiment operator hacking detection behavioral model

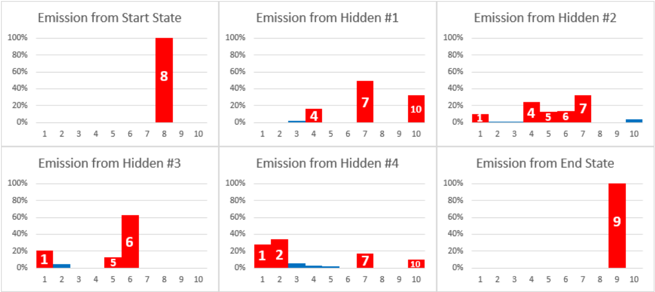


Fig. 3. The emission probabilities of the first HMM behavioral model

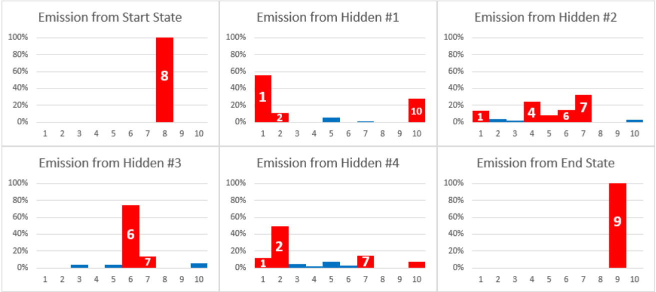


Fig. 5. The emission probabilities of the second HMM behavioral model

B. The Second Experiment Operator Hacking Detection Behavioral Model

The second operator hacking detection behavioral model was developed using detection observation sequences from the second experiment as discussed previously. The resulting second behavioral model is also a 6-state HMM model, as shown in Figure 4, with similar "Start" and "End" states. The interpretation of each state was also determined based on its emission probabilities, shown in Figure 5 using histograms. Although the emission probabilities of each hidden state in the first model (Figure 2) are slightly different than the corresponding hidden state in the second model (Figure 4), both models share the same overall structure.

The second model presents different major operational paths as compared to the first model. The dominant strategy in this model starts from the "Start", passes through "monitor UAV" and "adjust waypoint", then ends at the "End" state. Other slight differences in state transitions, especially the transitions related to the "monitor UAV" state, can also be observed. Participants in the first experiment behaved somewhat differently than participants in the second experiment, even with the same experimental conditions. As a consequence, the first model clustered slightly different observations with different weights for each hidden state as compared to the second model. Thus, variations in participants' actions in different experiments caused slight differences between these two HMM behavioral models.

However, while it seems there are only slight differences in these two model structures, we need more objective methods that allow us to compare the similarity between two different models. While the basic inferential statistics tell us the performance was different between these two experiments, it is not clear whether the difference was due to divergent strategies or differences in individual variability. Thus, having a more objective comparison of similarity in strategies can help to elucidate this difference. To further quantitatively compare these models, specific similarity measurements were applied to both models in the following section.

V. SIMILARITY MEASUREMENTS BETWEEN OPERATOR DETECTION MODELS

Inspection of the visualizations between these two HMM detection behavioral models with their emission histograms is seen in Figures 2 and 4, which provides an initial, albeit subjective measure of similarity. One objective interpretation, however, is in the sheer number of states that emerge from the Bayesian Information Criterion (BIC) model selection process. In both cases, a 6-state HMM model structure was suggested for both detection models by the BIC analysis. While this metric provides some confidence in model similarity and stability, additional analyses are needed to account for both the state and transition probability similarities. The next sections will provide such analyses.

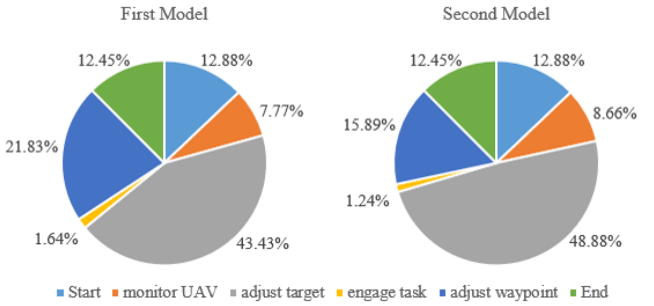


Fig. 6. State occurrence distributions of applying both detection HMM models respectively on the combined data set from both experiments

A. Hidden State Occurrence Distribution Comparison

To further quantitatively measure the similarity between HMM models, the Viterbi algorithm was utilized to generate hidden state sequences across the combined data set, which includes the data from the first experiment and the data from the first scenario of the second experiment. Then the two HMM models mentioned in the previous sections were applied to this combined data set to determine the likelihoods of state occurrence, as shown in Figure 6.

These distributions provide a more quantitative comparison as opposed to a simple visual inspection, where it can be seen that between the two experiments is the $\sim 5-6\%$ different in the waypoint and target adjustments. In addition, comparing to the hidden state occurrence percentage shown in Figure 2 and 4, Figure 6 presents similar distributions with $\sim 5-6\%$ differences in percentage for both models respectively. Considering variations in different operators' hacking detection strategies under the same experimental conditions, these slight differences between detection behavioral models seem reasonable and acceptable. Thus, the similarity of the hidden state occurrence distributions between both HMM models demonstrates the stability of using HMM models in similar HSC scenarios.

B. Divergence Distance Measure

To measure the similarity between the HMM models from a probabilistic aspect, the divergence distance measure was developed based on the concept of divergence and cross-entropy [19]. The divergence distance between two HMM models is defined as:

$$D_{KL}(\lambda_1 \| \lambda_2) = \frac{1}{M}(\log(P(O|\lambda_1)) - \log(P(O|\lambda_2))) \qquad (1)$$

Here λ_1 is the first HMM model developed from the first experiment, while λ_2 is the second HMM model developed from the second experiment. In addition, $O = \{O_1, O_2, ..., O_M\}$ represents the observation sequences, and M is the total number of observations. Hence, $\log(P(O|\lambda))$ is the log-likelihood value of an HMM model fitting on a given data set O.

TABLE II
THE DIVERGENCE DISTANCE MEASURE BETWEEN TWO HMM MODELS

1st experiment model vs 2nd experiment model				
data set	M	1st experiment model	2nd experiment model	$D_{KL}(\lambda_1 \| \lambda_2)$
		$\log(P(O_M \| \lambda_1))$	$\log(P(O_M \| \lambda_2))$	
1st experiment data	4315	-9.6116×10^3	-1.0676×10^4	0.2466
2nd experiment data	2534	-5.8314×10^3	-6.3351×10^3	0.1988
both experiment data	6849	-1.5443×10^4	-1.7011×10^4	0.2289
1st experiment high task load model vs low task load model				
data set	M	high taskload model	low taskload model	$D_{KL}(\lambda_h \| \lambda_l)$
		$\log(P(O_M \| \lambda_h))$	$\log(P(O_M \| \lambda_l))$	
high taskload data	2537	-5.5460×10^3	-5.9369×10^3	0.1541
low taskload data	1778	-3.9639×10^3	-4.0468×10^3	0.0466
both taskload data	4315	-9.5099×10^3	-9.9837×10^3	0.1098

To increase the confidence of the similarity measurement, different combinations of the original data sets were utilized for the divergence distance measurement between both detection behavior HMM models. To this end, the first experiment data and the second experiment data were tested using the divergence measurement respectively, then the combination of the experimental data was tested separately.

In order to provide a more objective view of the similarity and model stability, a baseline for the divergence measure is needed. Given no significant differences in participants' performances in the first experiment between their two scenarios and the fact that the same group of participants executed both, we believed that the behavior patterns in these two subsets of different task load scenarios should be the most consistent. Thus, the HMM behavioral models developed from these two subsets share the same model structure of a 6-state HMM shown in Figure 2. Similarly, these two scenario behavioral models, labeled as high task load model λ_h and low task load model λ_l, were tested on different combinations of the subsets.

Understanding that the divergence value will converge with over 1200 observations for HMM comparisons of similar model complexity [19], we are confident in the divergence value calculations since all combinations of data sets contain over 1500 observations. Table II lists all values for the calculation of the equation (1), including the number of observations, M, the log-likelihood value of the first HMM model applied on a given data set O, $\log(P(O|\lambda_1))$ or $\log(P(O|\lambda_h))$, and the log-likelihood of the second HMM model applied on the same data set, $\log(P(O|\lambda_2))$ or $\log(P(O|\lambda_l))$.

The difference between these two HMM behavioral models is smallest for the two scenarios from the first experiment (0.1098) when compared to the difference between the models from the two different experiments (0.2289). This is what we would expect since the same people performed the two scenarios in the first experiment, whereas different people performed the same mission in the second experiment. Another interesting result from Table II is examining the second experiment data across the first and second experiment models, where the divergence value from the second experiment data was the least among these three values. So, the variance of the underlying patterns in the second experiment data is slightly less than the first experiment data, which means the first experiment participants had less stable general strategies. Also, the comparison between two scenario models shows that the difference was the least for the low task load scenario data, which contains less variance than the high task load scenario data. Understanding that smaller divergence values indicate higher consistency, our hypothesis that the first experiment would have a higher degree of similarity was correct, although this analysis indicates that this experiment also produced more variable models.

C. Co-emission Probability Distance Measure

Another similarity measurement of the co-emission probability distance that focuses on the geometrical distance aspect was also used to investigate the similarity between detection models and to compare results with the divergence distance metric. The co-emission probability of two models also presents the generalizability of the models on across given data sets [18]. The co-emission probability of two HMM models is defined as:

$$A(\lambda_1,\lambda_2) = \sum_{O_M \in O} P_{\lambda_1}(O_M) P_{\lambda_2}(O_M) \tag{2}$$

Similarly, λ_1 represents the first detection HMM model and λ_2 represents the second detection HMM model. O_M is a sub-sequence of all observation sequences O. Thus, $P_\lambda(O_M)$ represents the probability of an HMM model fitting on a given data sequence O_M.

Using the co-emission probability, the similarity between two HMM models, λ_1 and λ_2, is defined as:

$$S(\lambda_1,\lambda_2) = A(\lambda_1,\lambda_2)/\sqrt{A(\lambda_1,\lambda_1)A(\lambda_2,\lambda_2)} \tag{3}$$

Such a similarity measurement follows the calculation of cosine similarity, which is represented using a dot product and magnitudes of two vectors:

$$similarity = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}} \tag{4}$$

TABLE III

THE CO-EMISSION PROBABILITY DISTANCE MEASURE BETWEEN TWO HMM MODELS

1st experiment model vs 2nd experiment model				
Data set	$A(\lambda_1,\lambda_2)$	$A(\lambda_1,\lambda_1)$	$A(\lambda_2,\lambda_2)$	$S(\lambda_1,\lambda_2)$
1st experiment data	1.2812×10^{-4}	8.5447×10^{-4}	1.9418×10^{-5}	0.9946
2nd experiment data	1.6038×10^{-4}	1.0613×10^{-3}	2.4409×10^{-5}	0.9965
both experiment data	2.8850×10^{-4}	1.9157×10^{-3}	4.3827×10^{-5}	0.9957
1st experiment high task load model vs low task load model				
Data set	$A(\lambda_h,\lambda_l)$	$A(\lambda_h,\lambda_h)$	$A(\lambda_l,\lambda_l)$	$S(\lambda_h,\lambda_l)$
high taskload data	2.6331×10^{-4}	1.8615×10^{-3}	3.8577×10^{-5}	0.9826
low taskload data	1.1222×10^{-4}	7.7597×10^{-4}	1.7389×10^{-5}	0.9661
both taskload data	3.7553×10^{-4}	2.6375×10^{-3}	5.5966×10^{-5}	0.9774

Similarly, the co-emission probability distance measurement was tested on the same combinations as with the divergence distance metric. A baseline was also created using the two scenario behavioral models, λ_h and λ_l as mentioned in the previous section, and two different task load data sets in the first experiment. The similarity distance results between various HMM models, $S(\lambda_1,\lambda_2)$ and $S(\lambda_h,\lambda_l)$, according to the equation (3) are listed in Table III.

Based on the definition of the cosine similarity, a similarity value of 1 means two tested models share the exact same structure, and a similarity value of 0 indicates orthogonality or decorrelation between two models. The high similarity between two HMM behavioral models is indicated by the similarity values, which are all close to 1, shown in Table III. Thus, the co-emission probability measure results support the high similarity level between HMM models. However, different from the divergence distance results, the difference between the two task load scenario models are slightly larger than the difference between the two experiment models in the co-emission probability measure. Thus, given that the co-emission probability represents the generalizability of models, the models for the people that repeated the experiment are less generalizable than the models from the different groups. So, while precisely capturing the underlying behavioral patterns from the training data, the scenario models have limited generalizability across other data sets generated with similar settings.

Like the divergence distance comparison results, the variation between both models is the smallest for the second experiment data among all data sets. Such a similar calculation

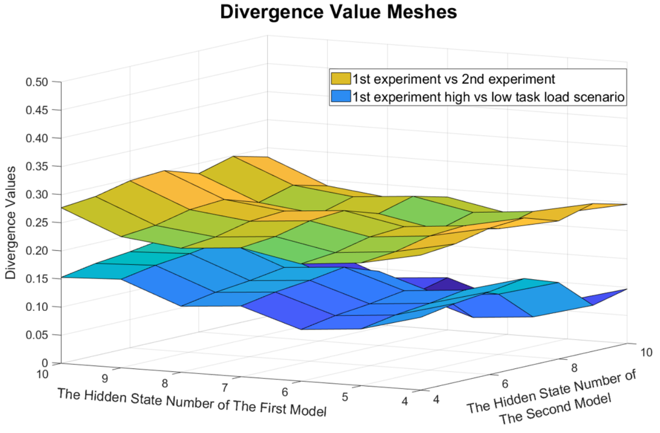


Fig. 7. The divergence values between the different model comparisons with all possible numbers of hidden states

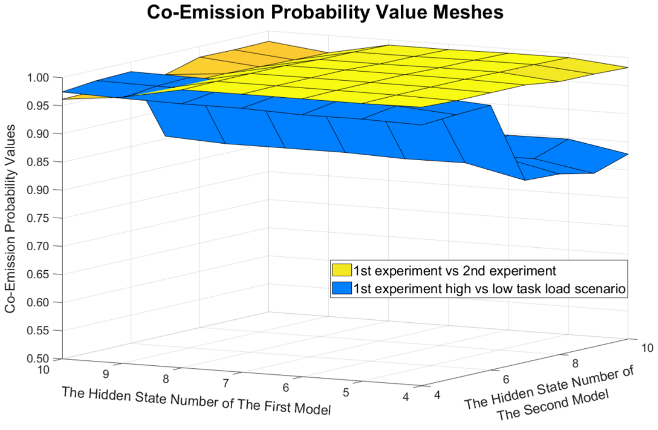


Fig. 8. The co-emission probability distance values between the different model comparisons with all possible numbers of hidden states

result ranking enhances the confidence of the comparison results from both the divergence distance and co-emission probability measure.

D. Mesh Analysis

In order to better understand the results from this analysis, including the degree of difference between various models and the stability of the underlying models, a visualization of the divergence distance was created. First, HMM models were developed for the first and second experiment data set respectively with all possible numbers of hidden states. Given that the “START” and “END” states represent hacking detection initiation and ending events, operator behavioral HMM models developed from such an experimental setting should have at least four hidden states. Given that we categorized participants’ actions into ten observations for model development, the HMM models should have at most ten hidden states. Thus, the range of possible hidden states is from four to ten.

Two divergence value meshes were plotted based on the divergence values, which were calculated on the combined data set from those additional HMM models. The lower reference mesh was plotted based on the two different task load data sets, the high versus low task load scenario for the same participants, in the first experiment. As discussed previously, this mesh represents the highest degree of similarity. The divergence values have an average of 0.1225 with a variance of 0.0012.

The upper mesh represents the comparison between the first and second experiments (Figure 7). These divergence values have an average of 0.2446 with a variance less than 0.0001, which demonstrates that regardless of the number of hidden states, the similarity across the models is consistent. This suggests that such small variation of divergence values illustrates the stability of the HMM models for representing behavioral models. In theory, Figure 7 suggests that a researcher could extract similar abstract information from the underlying combined data of the two experiments regardless of the number of hidden states.

Interestingly, the variation in the lower mesh model in Figure 7 from the two repeated test scenarios using the same people is larger than for the model comparison for different people completing the test scenario. So, while the same group of people exhibited more similar behavioral models, there was more variation in their models than for the different groups of people. One possible reason could be the difference in the amount of data points, in that for the two scenario comparisons for the same group of people, there were 36 people in each scenario and for the comparison of the two experiments, there were 72 and 45 people, respectively.

Two co-emission probability distance meshes, as shown in Figure 8, were also plotted based on the distance values from the comparisons between the two experiment models, shown as the upper mesh, and the two scenario models, shown as the lower mesh, respectively. While the meshes are fairly close, mirroring the results in Table III, there is a noticeable drop on the low task load model side starting from the 7 hidden states on the lower mesh. Such a drop implies that in the model selection process, models with more than 7 hidden states may not fit other data sets as well as for models with less hidden states. This fact also indicates the potential problems with model instability, possibly caused by the limited amount of data for the two test scenario models with the same subjects.

VI. DISCUSSION AND CONCLUSION

In general, the stability of a modeling technique represents the ability of a model to extract similar abstract data representations by developing similar model structures from data sets which are generated based on similar conditions. Specifically, for the development of diagnostic human operator behavioral models in HSC scenarios, the stability of a modeling technique is considered as the model’s ability to capture typical operator behavioral patterns under multiple HSC scenarios with similar conditions. Considering individual differences among human operators that inevitably lead to variations in strategy and performance, it is important to be able to have a modeling technique that can objectively determine when two models are close enough to say they represent similar strategies.

While the stability of the HMM operator modeling technique is presented in this paper, certain limitations still exist, especially in HSC scenarios. One core limitation in the use of the HMM model selection process is that when applying the BIC approach, experimenters' subjective judgments are required in determining and then interpreting the hidden states. Given the divergence value and the co-emission distance mesh analyses mentioned in previous sections, it is clear that the choices of a model from the developer could still slightly affect the final similarity computation. Moreover, the assessment of similarity in terms of what is either a small divergence distance or a large co-emission probability is also subjective and more work is needed to determine how to assess such metrics either relatively or absolutely for operator behavioral model comparison. This is currently the subject of continuing research.

This effort has shown that at least in this one case, HMMs can produce stable models of human operators in HSC scenarios, and also provide a basis to assess degree of similarity when comparing models with the same or different groups of people. What remains to be seen is how we can now leverage similarity metrics of HMM models to effectively use such models either diagnostically or predictively to understand when an operator has moved into a suboptimal or even unsafe state. For example, we will explore whether such similarity metrics can be used to flag when a strategy has become too dissimilar, which could indicate likely performance consequences. Additional future research directions will include whether such metrics can make the HMM model selection process more objective as well as whether such similarity metrics can identify groups of people in terms of knowledge, skills, and abilities which could inform training programs for HSC applications.

ACKNOWLEDGMENT

We gratefully acknowledge the efforts of those who have assisted in developing the RESCHU-SA experiment platform including Andy Wang, who worked as a summer intern in the Duke Humans and Autonomy Lab (HAL), and Duke undergraduate students, Jeffrey Wubbenhorst, Adithya Raghunathan and Yi Yan Tay. We also gratefully acknowledge the assistance of Matthew Seong, who is a UNC undergraduate student, in conducting experiments, and the assistance of Rong Xu, who is a Duke graduate student, in data analysis.

REFERENCES

[1] T. B. Sheridan, *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press, 1992.

[2] G. Pajares, "Overview and current status of remote sensing applications based on unmanned aerial vehicles (uavs)," *Photogrammetric Engineering & Remote Sensing*, vol. 81, no. 4, pp. 281–330, 2015.

[3] M. L. Cummings, S. Bruni, S. Mercier, and P. Mitchell, "Automation architecture for single operator, multiple uav command and control," *The International Command and Control Journal*, vol. 1, no. 2, pp. 1–24, 2007.

[4] M. L. Cummings, "Human supervisory control of swarming networks," in *2nd Annual Swarming: Autonomous Intelligent Networked Systems Conference*, Jun. 2004.

[5] M. L. Cummings, S. Bruni, and P. J. Mitchell, "Human supervisory control challenges in network-centric operations," *Reviews of Human Factors and Ergonomics*, vol. 6, no. 1, pp. 34–78, 2010.

[6] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman, *Engineering psychology & human performance*. Hove, U.K.: Psychology Press, 2015.

[7] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. New York: Springer, 2001, vol. 1, no. 10.

[8] Y. Boussemart, M. L. Cummings, J. L. Fargeas, and N. Roy, "Supervised vs. unsupervised learning for operator state modeling in unmanned vehicle settings," *Journal of Aerospace Computing, Information, and Communication*, vol. 8, no. 3, pp. 71–85, 2011.

[9] T. V. Duong, D. Q. Phung, H. H. Bui, and S. Venkatesh, "Human behavior recognition with generic exponential family duration modeling in the hidden semi-markov model," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 202–207.

[10] V. Grimm and S. F. Railsback, *Individual-based modeling and ecology*. Princeton, NJ: Princeton University Press, 2013.

[11] V. Rodríguez-Fernández, A. Gonzalez-Pardo, and D. Camacho, "Finding behavioral patterns of uav operators using multichannel hidden markov models," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–8.

[12] X. Meng, K. K. Lee, and Y. Xu, "Human driving behavior recognition based on hidden markov models," in *2006 IEEE International Conference on Robotics and Biomimetics*. IEEE, 2006, pp. 274–279.

[13] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[14] L. R. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[15] M. Lewis, J. Wang, and S. Hughes, "Usarsim: Simulation for the study of human-robot interaction," *Journal of Cognitive Engineering and Decision Making*, vol. 1, no. 1, pp. 98–120, 2007.

[16] S. M. E. Sahraeian and B.-J. Yoon, "A novel low-complexity hmm similarity measure," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 87–90, 2011.

[17] C. Bahlmann and H. Burkhardt, "Measuring hmm similarity with the bayes probability of error and its application to online handwriting recognition," in *Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2001, pp. 406–411.

[18] R. B. Lyngso, C. N. Pedersen, and H. Nielsen, "Metrics and similarity measures for hidden markov models," in *Proc. 7th Int. Conf. on Intelligent Syst. for Molecular Biology (ISMB-99)*, 1999, pp. 178–186.

[19] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, Feb. 1985.

[20] B. Donmez, C. Nehme, and M. L. Cummings, "Modeling workload impact in multiple unmanned vehicle supervisory control," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 6, pp. 1180–1190, Nov. 2010.

[21] A. J. Kerns, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, "Unmanned aircraft capture and control via gps spoofing," *Journal of Field Robotics*, vol. 31, no. 4, pp. 617–636, 2014.

[22] A. Broumandan, A. Jafarnia-Jahromi, V. Dehghanian, J. Nielsen, and G. Lachapelle, "Gnss spoofing detection in handheld receivers based on signal spatial correlation," in *Proceedings of the 2012 IEEE/ION Position, Location and Navigation Symposium*. IEEE, Apr. 2012, pp. 479–487.

[23] H. Zhu, M. Elfar, M. Pajic, Z. Wang, and M. L. Cummings, "Human augmentation of uav cyber-attack detection," in *International Conference on Augmented Cognition*. Springer, 2018, pp. 154–167.

[24] H. Zhu, M. L. Cummings, M. Elfar, Z. Wang, and M. Pajic, "Operator strategy model development in uav hacking detection," *IEEE Transactions on Human-Machine Systems*, pp. 1–10, 2019.

[25] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[26] G. Schwarz *et al.*, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[27] G. D. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.