Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing

Angela R. Harrivel[1], Chad L. Stephens[2], Robert J. Milletich[3], Christina M. Heinich[4], Mary Carolyn Last[5], Nicholas J. Napoli[6], Nijo A. Abraham[7], Lawrence J. Prinzel[8], Mark A. Motter[9], *and* Alan T. Pope[10]
NASA Langley Research Center, Hampton, Virginia, 23681

The Commercial Aviation Safety Team found the majority of recent international commercial aviation accidents attributable to loss of control inflight involved flight crew loss of airplane state awareness (ASA), and distraction was involved in all of them. Research on attention-related human performance limiting states (AHPLS) such as channelized attention, diverted attention, startle/surprise, and confirmation bias, has been recommended in a Safety Enhancement (SE) entitled "Training for Attention Management." To accomplish the detection of such cognitive and psychophysiological states, a broad suite of sensors was implemented to simultaneously measure their physiological markers during a high fidelity flight simulation human subject study. Twenty-four pilot participants were asked to wear the sensors while they performed benchmark tasks and motion-based flight scenarios designed to induce AHPLS. Pattern classification was employed to predict the occurrence of AHPLS during flight simulation also designed to induce those states. Classifier training data were collected during performance of the benchmark tasks. Multimodal classification was performed, using pre-processed electroencephalography, galvanic skin response, electrocardiogram, and respiration signals as input features. A combination of one, some or all modalities were used. Extreme gradient boosting, random forest and two support vector machine classifiers were implemented. The best accuracy for each modality-classifier combination is reported. Results using a select set of features and using the full set of available features are presented. Further, results are presented for training one classifier with the combined features and for training multiple classifiers with features from each modality separately. Using the select set of features and combined training, multistate prediction accuracy averaged 0.64 +/- 0.14 across thirteen participants and was significantly higher than that for the separate training case. These results support the goal of demonstrating simultaneous real-time classification of multiple states using multiple sensing modalities in high fidelity flight simulators. This detection is intended to support and inform training methods under development to mitigate the loss of ASA and thus reduce accidents and incidents.

I. Introduction

THE Commercial Aviation Safety Team (CAST) has established a Safety Enhancement entitled "Training for Attention Management,"[1,2] which calls for research and development on the detection of attention-related human performance limiting states (AHPLS). Such cognitive states can cause pilots to lose airplane state awareness, which CAST has specified as a causal factor in commercial aviation accidents and incidents. Analogous

[1] Researcher, Crew Systems & Aviation Operations Branch, Mail Stop 152, AIAA Senior Member
[2] Researcher, Crew Systems & Aviation Operations Branch, Mail Stop 152
[3] Researcher, Comprehensive Digital Transformation, Big Data Analytics and Machine Intelligence Team
[4] Researcher, Comprehensive Digital Transformation, Big Data Analytics and Machine Intelligence Team
[5] Researcher, Crew Systems & Aviation Operations Branch, Analytical Mechanics Associates, Inc., Mail Stop 152
[6] Researcher, National Institute of Aerospace, Systems and Information Engineering, University of Virginia
[7] Aerospace Engineer, Systems Integration and Test Branch, Mail Stop 424, AIAA Member
[8] Researcher, Crew Systems & Aviation Operations Branch, Mail Stop 152
[9] Researcher, Electronic Instrumentation Systems Branch, Mail Stop 488
[10] Researcher, Crew Systems & Aviation Operations Branch, Mail Stop 152

to training to recognize one's own reaction to symptoms of hypoxia or fatigue and respond appropriately, the detection of these states will aid the development of training methods to improve self-monitoring of and response to one's own attentional performance, thus increasing airplane state awareness to avoid accidents. Therefore, the cognitive states of interest in the present study are AHPLS - specifically channelized attention, diverted attention, confirmation bias, and startle/surprise.

The Crew State Monitoring group at the NASA Langley Research Center (LaRC) has implemented an end-to-end system to test the capability of predicting the occurrence of these cognitive states simultaneously in real time using psychophysiological data recorded via multiple sensing modalities. Here, results from a study conducted in a high fidelity motion-based flight simulator are reported which follow up on initial results previously reported from a study conducted in a fixed-base flight simulator. The initial results quantified the ability to discriminate between cognitive states as induced by benchmark tasks.[3] In the current study, the same benchmark tasks were used with new pilot participants to train classifier models which are then used to predict the cognitive state of those participants during flight simulation scenarios. The initial focus is on the states of Channelized Attention and Startle/Surprise.

Multimodal psychophysiological sensing for engagement, fatigue, emotion and workload prediction is emerging,[4-8] but prior work has not fully investigated the classification accuracy for the particular group of AHPLS of interest to CAST based on multimodal sensing in a real-time system using operationally-relevant, realistic flight scenarios. Further, data fusion methods for classifying psychological states from psychophysiological measures have not yet matured to universal acceptance[9] and the generalizability of classifier models across tasks, time and participants is yet to be fully determined. The study aim was to determine the accuracy with which AHPL state predictions can be made using multimodal psychophysiological measures as classifier input features. To answer this question, a variety of classifier models were implemented to predict AHPLS and determine how well those predictions converged with the intention of AHPLS-inducing flight simulation scenarios.

Data for the present report were collected during a human subject study using multiple simultaneous psychophysiological measures as a means of detecting AHPLS during both benchmark tasks and during experimental flight scenarios presented in a high-fidelity motion-based flight simulator. Each sensor's measured time series were processed individually prior to use to derive variables as needed and extract psychophysiological information relevant to AHPLS detection. Benchmark task data were used to train classifier models using all sensing modalities as input features on a per-participant basis. Then, a state prediction from among the states of interest was made for each subsequent small time window during simulated flight. If successful, the methods and prototype system will be ready for hardening toward further validation in future flight simulation experiments and for development as a mobile system to support demonstrations and use outside the research environment. One example of such use is in commercial operational flight training facilities.

II. Data Collection Methods

Twenty-four commercial aviation pilots (two per crew, none female) were asked to perform tasks in a motion-based flight simulator while wearing psychophysiological sensors. All participants consented to take part in the study as approved by the Institutional Review Board of NASA LaRC. The tasks included: resting tasks, benchmark tasks designed to induce AHPLS and low/high workload conditions, and experimental flight scenarios. Data collection was performed in the Research Flight Deck in the Cockpit Motion Facility at LaRC. Variations in attentional demand, startle/surprise and task load throughout the experimental flight scenarios were designed to induce AHPLS.

Psychophysiological sensors were applied to both participants to measure electroencephalography (EEG) signals using an Advanced Brain Monitoring electrode head net,[10] and to measure three-point electrocardiogram (ECG), respiration, and galvanic skin response (GSR) signals via a Mind Media, B.V. Nexus Mark II system. All measures were recorded passively throughout all tasks and simulated flight performance. Functional Near Infrared Spectroscopy (fNIRS) signals were measured for the captain's seat only using the Imagent by ISS, Inc. Pilot visual behavior and physiological eye responses were recorded in the flight simulator at both seats via a Smart Eye, AB. eye tracking system. Simulator flight control inputs were also recorded.

Three self-report questionnaires were administered after each task: the subjective NASA Task Load Index to evaluate each participant's workload,[11] a questionnaire regarding the extent to which each AHPLS was experienced, and the qualitative NASA Situation Awareness Rating Technique (SART)[12] to assess situation awareness. Additionally, at the conclusion of the day-long study, participants were asked: "1. Do you have any feedback on wearing the physio equipment?" and "2. Did any of the equipment hinder your performance?" This served to formally collect first-hand opinions regarding the comfort and intrusiveness of the sensor instrumentation.

All measured time series were recorded using MAPPS (EyesDx, Inc., Coralville, IA), a software suite designed to collect aircraft and simulator state, event markers, video, and pilot psychophysiological and behavioral data. The software time synchronizes all data channels for real-time review, as well as post-hoc analysis. The sampling rate was 256 Hz. EEG was reduced to examine frequency domain components via spectral analysis,[13,14] heart rate variability analysis was conducted on the ECG, and skin conductance level and skin conductance responses were derived from GSR signal to generate classifier input features in addition to the normalized time series measurements. The fNIRS signals were not recorded directly into MAPPS due to limitations with the MAPPS software. These were recorded at 6.25 Hz. Task engagement will be assessed[15] depending on fNIRS signal quality and presented separately. Also, data dropout issues were too pervasive to include eye tracking. Thus, psychophysiological signals included in these initial flight simulation prediction results are EEG, GSR, ECG and Respiration.

A "benchmark task" was used to induce each AHPL state. Use of benchmark tasks was modeled after the methods of Hirshfield, et al.[16] The AHPLS to be predicted and the selected benchmark tasks are listed in Table 1 and were described previously.[3] These tasks are used to induce AHPLS under controlled conditions for 6 minutes each, and were chosen for their high likelihood to induce these experiences in isolation and with the full knowledge of the participant (except for the startle task and the high versus low workload condition). Many of these tasks have been employed in previous task-oriented research.[17-21] Moments of reflexive startle and subsequent surprise due to expectation mismatch lasted only 15 seconds per benchmark state induction, producing class imbalance.

AHPLS	*Task*
Baseline rest	Rest, eyes open, crosshair
Channelized Attention	Tetris
Diverted Attention	Vigilance Task with Math
High Workload complex multi-task	MATB* High Workload
Low Workload complex multi-task	MATB Low Workload
Startle / Surprise	Movie Scene Observation
Confirmation Bias	Prestige number sequences

*Multi-Attribute Task Battery: available at http://matb.larc.nasa.gov/

Table 1. Benchmark Tasks.

Features extracted from signals collected during the benchmark tasks were used for classifier training purposes. Subsequently, and separately, simulated flight scenarios were presented to the pilot participants to strategically induce AHPLS. Signals collected during the simulated flight tasks were used as the test data, from which the same features were extracted, to produce classifier predictions. Accuracy reported here is determined by comparing the state predicted by the classifier model to the state induction intended by the flight scenario design. Aside from the use of one crew's benchmark and simulation data to select classifier tuning parameters, none of the flight simulation data were used to generate the model. Immature classifier models were used to predict state in real time during data collection. To develop mature and accurate classifier models and methods, the results reported here explore the sensing modalities, feature types, and classifier training method combinations which produce the most accurate results.

III. Data Processing Methods

As introduced above, and similar to methods previously reported,[3] measurements made during the benchmark tasks provide required ground truth for use with state classifiers using supervised machine learning techniques. Classifier models were trained to recognize pilot state during the experimental flight scenarios based on patterns of the physiological signals measured during the benchmark tasks. In this way, the benchmark data was used without a preconceived model of expected physiological signal change. Classifier model training data was that which was collected during both baseline resting and benchmark-task-induced states, enabling multi-state classification. The state classifiers then were used to characterize operator state during the times of intended state induction for each time point during the experimental flight scenario. During the flight simulation, channelized attention and startle/surprise were induced twice each during a Line-Oriented Flight Training scenario. Channelized Attention was induced after a Hydraulic System Pressure alert and a Trailing Edge Flap Asymmetry problem. Startle and surprise[22] were induced by unexpected but operationally-realistic flight events on takeoff and landing. This produced 3.6 to 7.3 minutes of channelized attention simulation prediction data, and 25 to 42 seconds of startle and surprise data. Classifier training and state prediction for Channelized Attention and Startle/Surprise were performed for 13 of the 24 participants who had complete data sets.

A. Feature Generation and Down-selection

For each participant and for each sensing modality, normalized time series features were calculated over a 5 second window and updated every second to support near real-time output. To deal with class imbalance, data consisted of 30 seconds of Startle/ Surprise benchmark data and 30 seconds of Channelized Attention benchmark data. The derived features are listed and summed in Table 2. The summary statistics consisted of 15 features: the first four moments, coefficient of variation, seven quantiles (p = 0.25, 0.50, 0.75, 0.025, 0.975, 0.005, 0.995), interquartile range, entropy, and area under the curve. Power spectral density (PSD) features consisted of the PSD estimates for 1 - 40 Hz in 1Hz bins, generating 40 features per channel. For the features generated by wavelet decomposition, we have employed an extension of the work of Von Tsharner[23] wherein a flattened Gaussian basis function is used. However, contrary to the use of an arbitrary cutoff frequency, filter bank design was constrained via an objective function to represent the EEG frequency bands between 1 and 40Hz. Wavelet decomposition features consisted of the wavelet coefficients for a level 7 decomposition using a Daubechies-4 wavelet, with a window size of 256 samples and window increment of 32.[24,25] Respiration rate was calculated using a weighted average of the fundamental frequency per minute. Average slope for GSR was calculated as the first-order difference of the signal. Drop score for GSR was calculated as the count of time instances in which GSR slope drops below -0.25. Programming was performed in MATLAB version 2015a.[26]

Modality	*Feature Types*	*Number of Features*
ECG	• Summary statistics of time series	15
HRV	• Mean and variance	2
EEG	• Summary statistics for time series • PSD frequencies 1-40Hz • Wavelet decomposition	15 x 20 channels 40 x 20 channels 33 x 20 channels
Respiration	• Summary statistics of time series • Respiration rate	15 1
GSR	• Summary statistics of time series • Average slope and drop score	15 2
Total		1810

Table 2. Feature types per sensing modality.

Features were down-selected from 1810 to 635 using 5-fold cross validation of complex tree models using the benchmark training data and predicting channelized attention and startle/surprise. Features were selected based on both counting the number of times a feature was used to create a split in the tree model, and on a mean squared feature importance metric.[27] Six important features for ECG (Variance, Skewness, Kurtosis, Quantile (p = 0.005), Interquartile Range, Entropy[28,29]), five important features for respiration (Respiration rate, Variance, Quantile (p = .975), Quantile (p = 0.025), Quantile (p = 0.25)), and five important features for GSR (Average slope, Drop score, Mean, Skewness, Quantile (p = 0.975)) were identified. Heart rate variability (HRV) mean and variance also were included, derived from the ECG signal. The Pan Tompkins[30] algorithm was used to detect the QRS complex in the ECG signal and return the locations of the R waves. HRV was calculated as reciprocal of the difference between adjacent R wave locations. Feature importance results for predicting Channelized Attention and Startle/ Surprise informed the feature set down-selection. The most important feature was found to be one of various EEG channels,[31] as shown in Table 3. The top seven of these were used for "select" analyses while all were used for "all" analyses. The second-most

Importance Rank	*EEG Channel*	*Occurrences as most important*
1	FP1	9
2	T5	5
3	O2	4
4	Fz	4
5	O1	3
6	F7	3
7	F8	3
8	T6	2
9	C3	2
10	T3	1
11	P4	1

Table 3. EEG channel importance.

important feature was respiration rate. The third was GSR slope change. Feature sets, *all*, and *select* were each used separately to produce results.

B. Classifier Types and Training Methods

Four different machine learning classifiers were investigated as candidate classifier types: (1) random forests (Scikit-Learn*), (2) gradient boosting (XGB†), (3) Nu-Support Vector Machine (ν-SVM*) with radial basis function kernels, and (4) ν-SVM with polynomial (2nd and 3rd) kernels. Random forest is an ensemble model which trains multiple weak decision tree classifiers that are combined into a single, robust model.[32] Random forest decision trees are trained using feature bootstrap aggregation, or bagging. Gradient boosting is an ensemble machine learning technique which fits many classification and regression trees (CART) to the input data.[33] CARTs are trained via the boosting method where multiple weak models are eventually combined into a single, effective model. ν-SVMs are similar to classic SVMs, but use a parameter ν to control the number of support vectors.[34] Programming was performed in MATLAB[24] and Python version 2.7‡.

Due to class imbalance in the benchmark data (the ratio of Channelized Attention samples [approximately 350 samples] to Startle/Surprise samples [approximately 25 samples] being roughly 14:1), a training/testing method based on the EasyEnsemble[35] algorithm was used. Each classifier's tuning parameters, per modality, were optimized for one crew using benchmark training and flight simulation test data from that one crew. Two sets of parameters were determined separately for the captain in the left seat and the first officer in the right seat because event experiences may have differed depending on their different flight tasks. These parameters were used for all the other participants depending on the seat. The parameters examined are given in Table 4.

Classifier Model	*Parameter values*
Random forest	
Estimators	25, 50, 100, 200
Criterion	Gini, entropy
Max features	sqrt(q), log2(q), 1
Gradient boosting	
Estimators	50, 100, 200
Row subsampling	0.75, 1
Max depth	3, 6
Learning rate	0.1, 0.01
L2 norm	0, 0.01
Feature subsampling	0.8, 1
v-SVM radial basis	
Gamma	1, 0.01, 0.001, 0.0001
Nu	0.1, 0.3, 0.5, 0.7, 0.9
v-SVM polynomial	
Gamma	1, 0.01, 0.001, 0.0001
Nu	0.1, 0.3, 0.5, 0.7, 0.9
Polynomial degree	1, 2, 3
q denotes the number of features	

Table 4. Model Hyperparameters.

Binary class (Channelized Attention or Startle/Surprise) predictions were made per participant (using tuning parameters depending on the seat) using each of the four candidate classifier types. Channelized Attention state induction lasted approximately 6 minutes and Startle/Surprise state induction lasted approximately 30 seconds total during flight simulation. During these times, accuracy was determined using a weighted area under the curve (AUC) to account for imbalanced classes. Results were generated using features from one, two, three and all four of the included sensing modalities, in all combinations. This was repeated across four different classifier training cases: two different feature sets (described in section III A) and two different training approaches. The two training approaches are as follows. Classifier models were trained on: (1) sensing modality features independently, generating multiple separate models and multiple predictions to feed one overall prediction, and (2) sensing modality features combined to train one model and produce one prediction. For the separate training method, the

* SKLearn available at: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

† eXtreme Gradient Boosting (XGB) available at: https://xgboost.readthedocs.org/en/latest/

‡ Python Software Foundation. Python Language Reference, version 2.7 available at http://www.python.org

overall prediction was based on first summing the predicted probabilities, and then using the argmax of the class probability vector as the class label. Results were calculated using all features and again using the selected features as described above. Thus, the four cases are: select separate, select combined, all separate, and all combined. The best-performing (highest weighted area under curve, AUC) candidate classifier type for each modality combination was then reported. Results for each case are presented below.

IV. Results and Discussion

Participant responses regarding the comfort and intrusiveness of the sensor instrumentation for the 24 of the 26 participants who responded are tallied according to level of concern in Table 5. The overall lack of concern encourages confidence in the validity of performance results despite the sensors. Major concerns such as GSR finger sensor location and wire management may be readily addressed in future studies. Further investment in sensor obtrusiveness reduction should improve crew acceptance toward implementation in operational contexts.

Concerns	*Question 1*	*Question 2*	*feedback*
None	11	12	• "Didn't notice it much after a while" • "Considering all of the things it was measuring, it was pretty comfortable, even toward the end of the day" • "Virtually invisible. After a few minutes, it's hard to even know it's there"
Minimal	10	9	• "Awkward and uncomfortable, but manageable" • A few comments concerning the fNIRS being uncomfortable • Gear didn't necessarily hinder flight performance, but still remained "a little distracting" for some
Major	3	3	• "Would have been better to have finger sensors on stick hand" • "Restricted my motion on several occasions as the wires would seem to be binding against the harness or the chair or something else in the cockpit"

Table 5. Categorization of Participant Feedback on Psychophysiological Gear.

Subjective self-report of situational awareness (SART),[12] reported in Table 6, showed the flight training scenario challenged the situational awareness of the participants as intended, with an average SART score of 44 which fell between the low workload and high workload SART ratings. Subjective self-report of workload (TLX)[11] showed that the workload was high, with an average TLX score of 62 which fell nearest to the TLX scores of the channelized attention and high workload benchmark tasks.

AHPL task	*NASA TLX*	*SART*
Rest	25	132
Diverted Attention	59	28
Channelized Attention	62	34
Low Workload	49	58
High Workload	64	29
Confirmation Bias	58	-6
Startle / Surprise	27	81
Flight Scenario	62	44

Table 6. Subjective self-report of workload (TLX) and situational awareness (SART).

Not surprisingly, the state prediction accuracy results show a dependency on machine learning classifier type, sensing modalities, feature sets, and training methods. Signal processing and data quality are also expected to affect outcomes, but were not parameterized here. Optimization is also expected to vary across participants. The maximum, minimum, average and standard deviation across the 13 participants is given in Tables 7 and 8.

Using *all* features with separate training per modality, using three modalities yielded the result with the overall best accuracy for this study (0.95 with EEG, Respiration and GSR). Using *all* features with combined training, using two modalities yielded the result with the second-best accuracy for this study (0.93 with ECG and GSR). However, in general, combined training resulted in mean AUC results both greater than and less than those with separate training. For separate training, comparing results between using all and selected features showed no significant differences.

	All Separate				*All Combined*		
number of modalities	*1*	*2*	*3*	*4*	*2*	*3*	*4*
average AUC	0.630	0.618	0.587	0.534	0.643	0.613	0.495
std. deviation	0.119	0.135	0.153	0.151	0.128	0.110	0.155
max AUC	0.845	0.876	0.953	0.823	0.934	0.871	0.735
min AUC	0.509	0.455	0.393	0.268	0.500	0.499	0.117
	times the modality appeared in the best combination				*times the modality appeared in the best combination*		
R	2	8	10	13	5	11	13
EEG	4	2	6	13	5	7	13
ECG	5	9	12	13	9	10	13
GSR	0	7	11	13	7	11	13
	times the modality appeared in a combination with accuracy above 0.70				*times the modality appeared in a combination with accuracy above 0.70*		
R	1	3	2	2	1	2	1
EEG	2	0	2	2	1	2	1
ECG	1	2	2	2	4	3	1
GSR	0	3	3	2	2	2	1

Table 7. Results using all features for the separate and combined training cases. *Classifier models were trained on: (1) sensing modality features independently, generating multiple separate models and multiple predictions to feed one overall prediction, and (2) sensing modality features combined to train one model and produce one prediction. Results were calculated using all features.*

Using *select* features in the case of three modalities, combined training was significantly better than training separate classifiers for each modality ($0.64 \pm 0.14 > 0.57 \pm 0.09$, $p<0.025$ (single tail, paired)). Using *select* features for two modalities, combined training tended to be better than separate. While no single sensing modality appeared significantly more often in high-accuracy combinations, EEG was among the *select* features which yielded all the best (maximum) AUC results regardless of the number of modalities used.

Results are encouraging and informative for work toward a translational real-time, multimodal, multi-state system under development. We note that both the SVM with a polynomial kernel and XGBoost have provided good results. However, further testing is needed before recommendations can be made regarding which classifier types are best for such a system. Combining modalities to train one classifier generally led to slightly better classifier performance than training separate modality-specific classifiers. Training with modalities combined allows synergistic or mutual information to be used. On the other hand, using modalities separately offers a more robust solution for operations, for example if signal is lost for one of the modalities, the other classifiers would still be active.

V. Limitations and Future Work

Further analysis will be performed to validate state induction by the simulations with physical behavioral and flight technical performance data, and subject matter expert opinion. Cognitive state predictions during the flight simulation scenarios may be further validated (beyond a comparison to the intended state induction) by the additional convergence of behavioral markers and missed manipulation checks.

	Select Separate				*Select Combined*		
number of modalities	*1*	*2*	*3*	*4*	*2*	*3*	*4*
average AUC	0.637	0.609	0.567*	0.495	0.649	0.637*	0.499
std. deviation	0.110	0.119	0.090	0.110	0.098	0.138	0.191
max AUC	0.832	0.832	0.808	0.772	0.853	0.860	0.837
min AUC	0.501	0.495	0.463	0.351	0.500	0.490	0.040
							* $p < 0.025$
	times the modality appeared in the best combination				*times the modality appeared in the best combination*		
R	0	8	8	13	8	10	13
EEG	7	6	7	13	6	8	13
ECG	3	8	12	13	7	10	13
GSR	1	4	12	13	5	11	13
	times the modality appeared in a combination with accuracy above 0.70				*times the modality appeared in a combination with accuracy above 0.70*		
R	0	2	0	1	0	2	2
EEG	3	4	2	1	1	3	2
ECG	0	0	2	1	1	3	2
GSR	1	2	2	1	2	4	2

Table 8. Results using select features for the separate and combined training cases. *Classifier models were trained on: (1) sensing modality features independently, generating multiple separate models and multiple predictions to feed one overall prediction, and (2) sensing modality features combined to train one model and produce one prediction. Results were calculated using select features.*

Classification accuracy is reported here for participant-dependent classifier models. In future efforts, both participant-dependent and participant-independent classification will be used to investigate the performance of model individualization methods[36] vs. more easily-implemented participant-independent methods. In this way, the benchmark and flight simulation data from many participants could be used to create generic models with hyperparameters optimized on benchmark-to-flight generalizability. Individualized, participant-dependent models may offer greater accuracy, especially if individually tuned, while participant-independent models obviate user-specific classifier training time. The need for less time would offer a significant advantage for operational use, such as training pilots to recognize AHPLS and better manage their own attention. Future efforts will also incorporate eye tracking and fNIRS data as additional, informative features, and will make predictions for High and Low Workload[37,38] and Diverted Attention in addition to Channelized Attention and Startle/Surprise.

Experiments and techniques to assess and improve generalizability across days[39] are yet to be explored, and also of importance for operational use. Better understanding of the natural overlap of and switching between AHPLS during flight and during simulated flight would inform future applications of a translatable system envisioned for use during flight instruction. Such applications will require cross-task generalizability for use across various scenario events and AHPLS induction periods. Benchmark task development and refinement toward more flight-like tasks and toward an improved and better-understood Confirmation Bias task is also indicated. Notably, the SART score for the benchmark Confirmation Bias task was negative. Also, increasing the amount of training data available for fleeting Startle/Surprise events may allow for separate prediction of these two responses. For further improvement of accuracy in operational contexts, employing adaptive on-line machine learning techniques is of

interest to reduce or eliminate individual classifier training time while maintaining useful positive predictive power.[9,40-42] Finally, data quality issues need to be addressed, including the implementation of data quality metrics to inform end users and guard against false positive or nuisance predictions. The eye tracker sometimes lost sight of the gaze, leading to dropped eye tracking data. Certain tasks led to a higher chance of this data loss, indicating that data quality metrics may potentially be used as informative features of their own.

VI. Conclusion

The results of this work are useful for determining the value of simultaneous multimodal psychophysiological measures and the value each sensing modality brings to classifier accuracy. Sensor instrumentation may then be chosen by weighing their value against the cost of using them in operational training contexts. Costs include pilot acceptance, obtrusiveness, comfort and privacy considerations, time spent training the classifier or applying sensors, and potential distraction from primary tasking – that of safe flight, real or simulated. However, such costs may not be appropriately weighed against the value of psychophysiological sensing until that value is adequately assessed and understood. This work begins to determine that value, and the projected efficacy of a crew state monitoring system and its potential future impact on the avoidance, detection, mitigation, and recovery from safety-critical human crew error. Real time state prediction information can be fed to the pilot themselves, their flight instructors, or automated intelligence in the cockpit to improve pilot flight performance, assist in the avoidance of errors during flight path monitoring, and optimize human-automation interaction.

Acknowledgments

This research is supported under NASA's Airspace Operations and Safety Program's Technologies for Aircraft State Awareness Sub-Project. We thank Kevin Shelton for fNIRS headgear fabrication, Charles Liles and Lin Chen for real time machine learning algorithm implementation, Harry Verstynen, Daniel Kiggins and Ray Comstock for contributions to flight scenario development, Joshua Reed and the Simulation Development and Analysis Branch for flight simulator support, Kyle Ellis and Stephanie Nicholas for participant and IRB protocol collaboration, and Trey Arthur for benchmark task software development.

References

[1]Commercial Aviation Safety Team, "Airplane State Awareness Joint Safety Analysis Team Interim Report," URL: http://www.skybrary.aero/index.php/Commercial_Aviation_Safety_Team_%28CAST%29_Reports [cited 3 March 2015].

[2]Commercial Aviation Safety Team, "SE211: Airplane State Awareness - Training for Attention Management (R-D)," URL: http://www.skybrary.aero/index.php/SE211:_Airplane_State_Awareness_-_Training_for_Attention_Management_(R-D) [cited 3 March 2015].

[3]Harrivel, A., Liles, C., Stephens, C., Ellis, K., Prinzel, L., Pope, A. Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. in American Institute of Aeronautics and Astronautics, SciTech. 2016: San Diego, California.

[4]Fairclough, S., and Gilleade, K., "Capturing user engagement via psychophysiology: measures and mechanisms for biocybernetic adaptation," International Journal of Autonomous and Adaptive Communications Systems, Vol. 6, No. 1, 2013, pp. 63–79.

[5]Thomas, L., Gast, C., Grube, R., Craig, K. Fatigue detection in commercial flight operations: Results using physiological measures. Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics. 2015: Las Vegas, Nevada.

[6]Verma, G. K., Tiwary, U. S., "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," NeuroImage, Vol. 102, 2014, pp. 162–172.

[7]Wilhelm, F., and Grossman, P., "Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment," Biological Psychology, Vol. 84, 2010, pp. 552–569.

[8]Wilson, G. F. and Russell C. A., "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," Hum. Factors, Vol. 45, No. 4, 2003, pp. 635-643.

[9]Novak, D., Mihelj, M., Munih, M., "A survey of methods for data fusion and system adaption using autonomic nervous system responses in physiological computing," Interacting with Computers, Vol. 24, 2012, pp. 154-172.

[10]Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., Craven, P. L., "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks," Aviation, Space and Environmental Medicine, Vol. 78, No. 5, 2007, Section II.

[11]Hart, S. G. and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload, North Holland Press, Amsterdam, 1988.

[12]Taylor, R. M., "Situation awareness rating technique (SART): the development of a tool for aircrew systems design," Situational Awareness in Aerospace Operations, Neuilly sur-Seine, France, NATO-AGARD-CP-478, 1990.

[13]Li, F., "Improving Engagement Assessment by Model Individualization and Deep Learning," Dissertation, Old Dominion University, 2015.

[14]Pope, A. T., Bogart, E. H., Bartolome, E. S., "Biocybernetic system evaluates indices of operator engagement in automated task," Biological Psychology, Vol. 40, 1995, pp. 187-195.

[15]Harrivel, A., Weissman, D., Noll, D., Huppert, T. and Peltier, S., "Dynamic filtering improves attentional state prediction with fNIRS," Biomed. Opt. Express. Vol. 7, 2016, pp. 979-1002.

[16]Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., "Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload with Functional Near Infrared Spectroscopy," Computer-Human Interaction, Boston, MA, 2009.

[17]Fairclough, S. H., & Gilleade, K., "Construction of the Biocybernetic Loop: A Case Study," *ICMI 2012 Grand Challenge - Brain-Computer Interfaces*, Santa Monica, CA, 2012.

[18]Gross, J. J. and Levenson, R. W. Emotion elicitation using films. Cognition & Emotion, Vol. 9, 1995, pp. 87-108.

[19]Parasuraman, R., & Davies, D. R., "A Taxonomic Analysis of Vigilance," R. R. Mackie, *Vigilance: Theory, Operational Performance, and Physiological Correlates*, Plenum, New York, NY, 1977, pp. 559-574.

[20]Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock, J. R. (2011). The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide. NASA, Langley Research Center. Hampton: NASA/TM-2011-217164, L-20031, NF1676L-12800.

[21]Stephens, C.L., Christie, I.C., and Friedman, B.H. "Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis," Biological Psychology, Vol. 84, 2010, pp. 463-473.

[22]Rivera, J., Jentsch, F., Talone, A. B., Boesser, C. T., and Jimenez, C., "Defining Startle, Surprise, and Distraction: A State-of-the-Art Technical Review to Support the Development of FAA Technical and Advisory Guidance, and of Line-Oriented Simulation Scenarios for Training," University of Central Florida, 2015, FAA CRA 13-G-007.

[23]Von Tscharner, V., "Intensity analysis in time-frequency space of surface myoelectric signals by wavelets of specified resolution," J. Electromyogr Kinesio, Vol. 6, No. 10, 2000, pp. 433–445.

[24]R. N. Khushaba, A. Al-Jumaily, and A. Al-Ani, "Novel Feature Extraction Method based on Fuzzy Entropy and Wavelet Packet Transform for Myoelectric Control", 7th International Symposium on Communications and Information Technologies ISCIT2007, Sydney, Australia, pp. 352 – 357.

[25]R. N. Khushaba, S. Kodagoa, S. Lal, and G. Dissanayake, "Driver Drowsiness Classification Using Fuzzy Wavelet Packet Based Feature Extraction Algorithm", IEEE Transaction on Biomedical Engineering, Vol. 58, No. 1, 2011, pp. 121-131.

[26]MATLAB, Statistics and Signal Processing Toolboxes, R2015a, The MathWorks, Inc., Natick, MA, United States.

[27]Breiman, L., Friedman, J., Olshen, R., Stone, C., "Classification and Regression Trees," Wadsworth, Belmont, CA, 1984.

[28]Lin, J. "Divergence measures based on the Shannon entropy," IEEE Transactions on Information theory," Vol. 37, No. 1, 1991, pp. 145-151.

[29]Gilland, J. *Driving, eye-tracking and visual entropy: Exploration of age and task effects*. ProQuest, 2008.

[30]Hamilton, P. S., Tompkins, W. J., "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," IEEE Trans. Biomed. Eng., Vol. 33, No. 12, 1986, pp. 1157-1165.

[31]Jasper, H., "Report of the committee on methods of clinical examination in electroencephalography: 1957," Electroencephalography and Clinical Neurophysiology, Vol. 10, No. 2, pp. 370–375. doi:10.1016/0013-4694(58)90053-1

[32]Breiman, L., "Random Forests," 2001, URL:https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf [cited 23 October 2015]

[33]Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine," 2001, URL: https://statweb.stanford.edu/~jhf/ftp/trebst.pdf [cited 23 October 2015]

[34]Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C., "Estimating the support of a high-dimensional distribution," Neural Computation, Vol. 13, 2001, pp. 1443-1471.

[35]Liu, X., Wu, J., Zhou, Z., "Exploratory Undersampling for Class-Imbalance Learning," IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS – PART B, 2008.

[36]Li, F., "Improving Engagement Assessment by Model Individualization and Deep Learning," Dissertation, Old Dominion University, 2015.

[37]Wilson G. F. and Russell, C. A., "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," Human Factors, Vol. 45, No. 4, 2003, pp. 635-643.

[38]Nickel, P., Nachreiner, F., "Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload," Human Factors, Vol. 45, No. 4, 2003, pp. 575-590.

[39]Christensen, J. C., Estepp, J. R., Wilson, G. F., Russel, C. A., "The effects of day-to-day variability of physiological data on operator functional state classification," NeuroImage, Vol. 59, 2012, pp. 57–63.

[40]Malkawi, M. and Murad, O., "Artificial neuro fuzzy logic system for detecting human emotions," Human-centric Computing and Information Sciences, Vol. 3, No. 3, 2013. doi: 10.1186/2192-1962-3-3

[41]Lin, C-T., Ko, L-W., Chung, I-F, Huang, T-Y, Chen, Y-C, Jung,T-P, and Liang, S-F., "Adaptive EEG-Based Alertness Estimation System by Using ICA-Based Fuzzy Neural Networks," IEEE Transactions on Circuits and Systems—I: Regular Papers, Vol. 53, No. 11, 2006, pp. 2469-2476.

[42]Moon, B.S., Lee, H.C., Lee, Y.H., Park, J.C., Oh, I.S., and Lee, J.W., "Fuzzy systems to process ECG and EEG signals for quantification of the mental workload," Information Sciences, Vol. 142, 2002, pp. 23–35.