IEEE TRANSACTIONS ON

HUMAN-MACHINE SYSTEMS

A PUBLICATION OF THE IEEE SYSTEMS, MAN, AND CYBERNETICS SOCIETY

DECEMBER 2019 VOLUME 49 NUMBER 6 ITHSA6 (ISSN 2168-2291)

SPECIAL ISSUE ON COMPUTATIONAL HUMAN PERFORMANCE MODELING

REGULAR PAPERS

(Contents Continued on Page 469)

IEEE

(Contents Continued from Front Cover)

IEEE SYSTEMS, MAN, AND CYBERNETICS SOCIETY

The IEEE Systems, Man, and Cybernetics Society is an organization within the framework of the IEEE, with professional interest in the closely interrelated fields of man–machine systems, systems science, systems engineering, and cybernetics. All members of the IEEE are eligible for membership in the Society and will receive this Transactions upon payment of the annual Society membership fee of $12.00 plus an additional subscription fee of $16.00. Members of certain other professional societies are eligible to become Affiliates of the Society. For information on joining, write to the IEEE at the address below. *Member copies of Transactions/Journals are for personal use only.*

IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS

IEEE Transactions on Human-Machine Systems (ISSN 2168-2291) is published bimonthly by The Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society, Council, or its members. **IEEE Corporate Office:** 3 Park, Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $28.00 (first copy only), nonmembers $54.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request.

Postmaster: Send address changes to IEEE Transactions on Human-Machine Systems, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

Digital Object Identifier 10.1109/THMS.2019.2948238

Editorial
Special Issue on Computational Human Performance Modeling

HUMAN performance modeling (HPM) is a method of quantifying human behavior, cognition, and processes; a tool used by human factors researchers and practitioners for both the analysis of human function and for the development of systems designed for optimal user experience and interaction. Different from data-driven approaches (e.g., neural networks), most HPM approaches use "top-down" modeling methods based on the fundamental mechanisms of human cognition and behavior.

This special issue introduces several human performance modeling articles that make use of mathematical modeling, production systems, and formal methods. We hope that readers of the special issue can benefit from the variety of modeling articles using different modeling approaches. In the following paragraphs, we summarize the features of each modeling approach and briefly introduce the articles in this special issue.

I. Introduction to Mathematical Modeling

Mathematical equations can predict, quantify, and analyze human performance, workload, brain waves, and other indices of human behavior in a rigorous way. Compared with computer simulation, mathematical modeling has the following features:

1) Mathematical models and equations of human behavior quantify and extract the mechanisms of human behavior by through quantification of the relationships among variables, including the inputs and outputs of each equation.
2) These mathematical models can be easier for users to understand, and for extracting relationships among variables, than reading computer code.
3) Mathematical models and equations of human behavior are relatively easy to edit, modify, improve, and integrate with other models to develop entirely new equations.
4) Mathematical models and equations of human behavior and performance can easily be implemented with different programming languages and they can be imbedded in different intelligent systems to provide a basis for systems design.
5) Mathematical models and equations can lead to analytical solutions, which are more accurate than simulation (heuristic) results.
6) There are mathematical models and equations quantifying the entire human cognition system [1], [2], which is another unique feature of this particular approach.
7) There are mathematical models and equations that can be proved by derivation with no need for verification in terms of empirical data (e.g., [3]).

This last feature is one of the most important aspects of mathematical modeling. Mathematical modeling discovers relations among variables that may never been studied or explored by experimental studies and it sometimes opens "new doors" in scientific discovery. One typical example of mathematical modeling is the Theory of Relativity by Albert Einstein, which opened new doors in physics and has guided hundreds of experiments. However, mathematical models cannot completely replace production systems (computer code), as production systems can generate simulated behavior.

When a modeler constructs a new mathematical model of human performance, it is important for the modeler to clearly present each step in deriving the equations to minimize circular reasoning. In addition, modelers should clearly list all variable values (how they are set) and any free parameters in the model. The modeling articles in this issue follow these practices (see details in [4]).

The articles in this special issue that make use of mathematical modeling approaches, include: [item 1) in the Appendix] focuses on development of a novel mathematical model of human eye gaze behavior under workload, derived from the basic principle of information constrained control; and [item 2) in the Appendix] provides an example of mathematical modeling of human trust in dynamic human-machine interaction. The proposed model describes human trust levels as a function of experience, cumulative trust, and expectation bias. [Item 3) in the Appendix] describes a mathematical model to quantify drivers' attentional instability on a winding roadway.

II. Introduction to Production Systems

Production systems are a collection of "if-then" rules that represent human information processing of some cognitive task [5]. Newell and Simon [6] introduced production systems into cognitive psychology through their seminal book on human problem solving. A production system is characterized by an architecture that consists of two types of memories. A production memory holds the rules of the system and the content persists for the duration of system execution. Complementing production

Digital Object Identifier 10.1109/THMS.2019.2942742

memory is data memory, which stores dynamic information about the current task.

Early production systems served as models of specific cognitive skills. For example, Just and Carpenter [7] used a production system to characterize reading and comprehension. Also, Rouse *et al*. developed a production system to model operator fault diagnoses [8]. These models not only addressed cognitive skills, but also how tasks were procedurally performed [9].

As the field of cognitive psychology matured, "cognitive architectures" emerged to explicate principles of cognition. Production systems can be considered as a form of task-dependent cognitive architecture. Architectures such as Soar [10], the Adaptive Control of Thought [11], and the Model Human Processor (MHP) [12], [13] have provided not only a means to systematize productions but also a way of validating theories of cognition.

More recent developments include incorporating visual and auditory perception into cognitive architectures [14] and to more accurately represent temporal sequencing of executive processes [15]. Sequencing has been simulated using a Queueing Network-MHP (QN-MHP) architecture, which combines a queueing network with production rules.

In the current issue, Rhie *et al*. [item 4) in the Appendix] extend the QN-MHP framework to explore the relationship between sensory processes and deeper semantic properties. Specifically, they used oculomotor behaviors, such as reaction time and movement patterns, to explain cognitive levels of processing in a driving task.

III. Introduction to Formal Methods

Another active research area with a focus on human performance modeling relates to formal methods. Formal methods are mathematical techniques and tools that enable an analyst to specify, model, and formally verify the operation/behavior of systems [16]. The specification process rigorously describes desirable system properties (usually using different types of modal logic [17]) that engineers and analysts want to be true in a system. Modeling involves analysts using mathematical languages (usually based on state machines) to describe the behavior of the target system that they wish to analyze. The formal verification process then has the analyst mathematically prove whether the model satisfies the specification. This process can take a number of different forms including pen and paper proofs. However, modern approaches use software that enables semi-automated, and fully automated methods [18], [19].

Formal methods have been integrated into the engineering life cycle in a number of different capacities. This includes learning formal models of existing systems, analyzing models of legacy and new designs, generating implementations from formal models, and evaluating/validating implementations as being consistent with proved formal properties. Formal methods are extremely good at addressing unexpected problems caused by interactions between components in complex systems. As such, they have primarily been used in the design and analysis of computer hardware and software systems. Furthermore, a growing body of research has been investigating how they can be applied to human performance modeling to engineer safe and effective human–machine systems [20]–[22].

This special issue contains three articles, [items 5)—7) in the Appendix], that advance the use of formal methods with human performance modeling. In [item 5) in the Appendix], Abbate and Bass describe a novel method that accounts for "connectibility affordance" in formal verification analyses. This allows formal analyses to prove whether the capabilities of people in an environment and possible source-target connections (e.g., tubes connecting pieces of equipment) afforded by environmental objects will prevent unintended configurations. In [item 6) in the Appendix], Joo and Shin present a new formal framework that accounts for human-machine interaction in emerging adaptive automation in smart manufacturing systems. Finally, in [item 7) in the Appendix], Zhu *et al*. make use of a hidden Markov model (a probabilistic formalism) to discover two overriding strategies humans use to detect hacking of unmanned aerial vehicles. These three articles provide useful insights into where research in this area is headed. Traditional formal analyses that incorporate human performance modeling have focused on well-defined human task and cognitive models [20]. While useful, such approaches work best for addressing structured work. They do a poor job of accounting for unstructured work environments in which human behavior may be less predictable. The work on affordances presented by Abbate and Bass, [item 5) in the Appendix], provides knowledge of how formal methods can be used in unstructured environments to prove properties about system safety and resilience. Additionally, legacy formal human factors methods have tended to focus on system automation behavior, which is static and unchanging. However, with the proliferation of increasingly sophisticated automation that can behave autonomously and learn in response to environmental changes, formal methods (like those explored by Joo and Shin in [item 6) in the Appendix] will need to be developed to provide analysts with the capability to prove that human-automation interactions will be safe. Finally, traditional formal methods are discrete and do not account for stochastic behaviors. However, developments in probabilistic formal methods, like those in [item 7) in the Appendix], are beginning to be explored by human performance modelers and will enable analysts to account for uncertainty in human behavior and thus accurately assess the reliability of complex, human-interactive systems when absolute guarantees are not possible.

IV. Statistical Models

The remaining articles broadly fall under the category of statistical models. In [item 8) in the Appendix], Shi *et al*. developed an anomaly detection framework based on Bayes' theorem and metric learning. The authors were able to utilize multiple user feedback simultaneously. Their algorithm was tested in a user study. In [item 9) in the Appendix], Lu and Sarter discovered a relationship between automation reliability and eye movement data. They showed that people dwell more on low-reliable automation interfaces. In [item 10) in the Appendix], Shive *et al*. established statistical saliency as the relationship between the search time and the difference in a target item's features and the distribution of display features.

They then used the statistical saliency model to effectively color code maps by reducing search times.

C. Wu, *Guest Editor*
Department of Systems and
Industrial Engineering
University of Arizona
Tucson, AZ 85719 USA.

L. Rothrock, *Guest Editor*
Department of Industrial and
Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802 USA.

M. Bolton, *Guest Editor*
Department of Industrial and
Systems Engineering
University at Buffalo
Buffalo, NY 14260 USA.

Appendix
Related Work

1) R. M. Hecht, A. B. Hillel, A. Telpaz, O. Tsimhoni, and N. Tishby, "Information constrained control analysis of eye gaze distribution under workload," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2930996.
2) W.-L. Hu, K. Akash, T. Reid, and N. Jain, "Computational modeling of the dynamics of human trust during human-machine interactions," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2018.2874188.
3) R. J. Jagacinski *et al*., "Drivers' attentional instability on a winding roadway," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2906612.
4) Y. L. Rhie, J. H. Lim, and M. H. Yun, "Queueing network based driver model for varying levels of information processing," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2018.2874183.
5) J. Abbate and E. J. Bass, "A formal approach to connectibility affordances," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2018.2886265.
6) T. Joo and D. Shin, "Formalizing human-machine interactions for adaptive automation in smart manufacturing," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2903402.
7) H. Zhu, M. L. Cummings, M. Elfar, Z. Wang, and M. Pajic, "Operator strategy model development in UAV hacking detection," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2018.2888578.
8) Y. Shi, M. Xu, R. Zhao, H. Fu, T. Wu, and N. Cao, "Interactive context-aware anomaly detection guided by user feedback," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2925195.
9) Y. Lu and N. Sarter, "Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2930980.
10) J. Shive, S. Rosichan, S. Davis, C. Wade, J. Ellison, and S. Santoni-Sanchez, "The statistical saliency model can choose colors for items on maps," *IEEE Trans. Human-Mach. Syst.,* to be published, doi: 10.1109/THMS.2019.2901896.

References

[1] C. Wu and Y. Liu, "Queuing network modeling of driver workload and performance," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 528–537, Sep. 2007.
[2] C. Wu and Y. Liu, "Queuing network modeling of the psychological refractory period (PRP)," (in English), *Psychological Rev.*, vol. 115, no. 4, pp. 913–954, Oct. 2008.
[3] C. Wu, M. Berman, and Y. Liu, "Optimization in the brain? Modeling human behavior and brain activation patterns with queuing network and reinforcement learning algorithms," in *Computational Neuroscience*, A. Chaovalitwongse and P. Xanthopoulos, Eds. New York, NY, USA: Taylor & Francis, 2010, pp. 159–179.
[4] C. Wu, "The five key questions in human performance modeling," *Int. J. Ind. Ergonom.*, vol. 63, pp. 3–6, 2018.
[5] R. M. Young, "Production systems in cognitive psychology," in *International Encyclopedia of the Social & Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford, U.K.: Pergamon, 2001, pp. 12143–12146.
[6] J. R. Anderson, "Problem solving and learning," *Amer. Psychologist*, vol. 48, no. 1, pp. 35–44, 1993.
[7] M. A. Just and P. A. Carpenter, *The Psychology of Reading and Language Comprehension*. Needham Heights, MA, USA: Allyn & Bacon, 1987.
[8] W. B. Rouse, S. H. Rouse, and S. J. Pellegrino, "A rule-based model of human problem solving performance in fault diagnosis tasks," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, no. 7, pp. 366–376, Jul. 1980.
[9] P. J. Sticha, "Models of procedural control for human performance simulation," *Human Factors*, vol. 29, no. 4, pp. 421–432, 1987.
[10] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artif. Intell.*, vol. 33, pp. 1–64, 1987.
[11] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1998, p. 490.
[12] S. Card, T. Moran, and A. Newell, *The Psychology of Computer-Human Interaction*. Hillsdale, NJ, USA: Lawrence Erlbaum, 1983.
[13] D. E. Kieras, S. D. Wood, and D. E. Meyer, "Predictive engineering models based on the EPIC architecture for a multimodal high-performance human-computer interaction task," *ACM Trans. Comput. Human Interact.*, vol. 4, no. 3, pp. 230–275, 1997.
[14] D. E. Kieras, G. H. Wakefield, E. R. Thompson, N. Iyer, and B. D. Simpson, "Modeling two-channel speech processing with the EPIC cognitive architecture," *Topics Cognitive Sci.*, vol. 8, no. 1, pp. 291–304, 2016.
[15] L. Yili, F. Robert, and T. Omer, "Queueing network-model human processor (QN-MHP): A computational architecture for multitask performance in human-machine systems," *ACM Trans. Comput.-Human Interact.*, vol. 13, no. 1, pp. 37–70, 2006.
[16] J. M. Wing, "A specifier's introduction to formal methods," *Computer*, vol. 23, no. 9, pp. 8–22, 1990.
[17] E. A. Emerson, "Temporal and modal logic," in *Handbook of Theoretical Computer Science*, J. van Leeuwen, A. R. Meyer, M. Nivat, M. Paterson, and D. Perrin, Eds. Cambridge, MA, USA: MIT Press, 1990, pp. 995–1072.
[18] N. Shankar, S. Owre, J. M. Rushby, and D. W. J. Stringer-Calvert, "PVS prover guide," Comput. Sci. Lab., SRI Int., Menlo Park, CA, USA, 1999.
[19] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking*. Cambridge, MA, USA: MIT Press, 1999.
[20] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Using formal verification to evaluate human-automation interaction in safety critical systems: A review," *IEEE Trans. Syst., Man Cybern., Syst.*, vol. 43, no. 3, pp. 488–503, May 2013.

[21] B. Weyers, J. Bowen, A. Dix, and P. Palanque, *The Handbook of Formal Methods in Human-Computer Interaction*. New York, NY, USA: Springer, 2017.

[22] M. L. Bolton, "Novel developments in formal methods for human factors engineering," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 2017, vol. 61, pp. 715–717.

Changxu (Sean) Wu received the Ph.D. degree in industrial and operations engineering from the University of Michigan, Ann Arbor, MI, USA, in 2007.

He has published 117 papers in his field including 82 journal papers, 36 conference papers, one book chapter, and two patents in intelligent system design authorized. His research interests include integrating cognitive science and engineering system design, especially modeling human cognition system with its applications in system design, improving transportation safety, promoting human performance in human–computer interaction, and inventing innovative sustainable and smart energy systems with human in the loop.

Dr. Wu is an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATIONS SYSTEMS, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, and *Behaviour & Information Technology*. He was the Chair of Human Performance Modeling Technical Group of Human Factors and Ergonomics Society (HFES) in USA. He was the recipient of the Senior Researcher of the Year Award from the Dean of School the Engineering & Applied Sciences at SUNY Buffalo and Outstanding Student Instructor Award from the American Society of Engineering Education (ASEE).

Ling Rothrock (M'90–SM'10) received the Ph.D. degree in industrial engineering from Georgia Tech, Atlanta, GA, USA, in 1995.

He is currently an Associate Professor with the Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA, USA. He has published more than 100 research articles. His research interests include human-in-the-loop simulations, behavioral decision making, display visualization, and human–machine performance assessment.

Dr. Rothrock is the Editor of *Human Factors and Ergonomics in Manufacturing & Service Industries* as well as an Associate Editor for the IEEE TRANSACTIONS ON HUMAN MACHINE SYSTEMS.

Matthew Bolton (S'05–M'10–SM'16) received the B.S. degree in computer science, the M.S. degree in systems engineering, and the Ph.D. degree in systems engineering from the University of Virginia, Charlottesville, VA, USA, in 2004, 2006, and 2010, respectively.

He is currently an Associate Professor with the Department of Industrial and Systems Engineering, the University at Buffalo, the State University of New York, Buffalo, NY, USA. His research interests include the use of human performance modeling and formal methods in the analysis, design, and evaluation of safety-critical systems.

Dr. Bolton has received funding on research projects from the Mozilla Foundation, the European Space Agency, the Agency for Healthcare Research and Quality, the Air Force Research Laboratory, the National Science Foundation, and the Army Research Office (through a Young Investigator Program Award). He was the 2018 recipient of the William C. Howell Young Investigator Award from the Human Factors and Ergonomics Society.

Information Constrained Control Analysis of Eye Gaze Distribution Under Workload

Ron Moshe Hecht, Aharon Bar Hillel, Ariel Telpaz, Omer Tsimhoni, and Naftali Tishby

***Abstract*—We describe a novel model of human eye gaze behavior under workload, derived from the basic principle of information constrained control. The model assumes two distributions over the visual field: A saliency distribution, which is nongoal oriented, and a reward task-related distribution. The eye gaze behavior is determined by the tradeoff between these two distributions, where the goal is to preserve the task-related constraints, while remaining as close as possible to the saliency distribution representing a comfort zone. Based on minimum Kullback–Liebler divergence principles, the model gives rise to a family of gaze distributions controlled by a single tradeoff parameter. The model was evaluated experimentally in a driving simulator that consisted of an immersive environment with clear tasks and accurate monitoring capabilities. The findings confirm the theoretical predictions with respect to the low rank manifold and order relations in the data. We show that the model can be used to visualize the unknown reward function associated with a task, and predict human workload based on gaze pattern.**

***Index Terms*—Eye gazing distribution, information constrained control (ICC).**

I. Introduction

A MAJOR portion of human visual abilities is enabled solely in the receptive field of the fovea, thus making its fixation regions the most important instance of active sensing in humans. The human visual system produces gaze behavior via a complex interaction of bottom-up and top-down processing streams. Although the bottom-up process guides gaze toward salient regions, in the presence of a demanding task, saliency driven processing is partially overridden by a more task-related top-down pattern.

Numerous computational models have been put forward to describe workload-free gaze behavior, mainly using the concept of a non-task-related saliency model [1]–[3]. These have been validated both as components of computer vision algorithms [2] and [3], and models predicting human behavior [1]; see [4] for a review. Task-related attention processes have mainly been addressed using qualitative models in the fields of cognitive psychology and human factors [5]–[10]. Experimental studies [6], [5] have shown that when subjects engage in a task that includes high levels of perceptual load, irrelevant visual distractors are not attended to at all, a phenomenon which was termed "early selection." There is an ambiguous relationship between cognitive workload, working memory (WM) workload, and attending to irrelevant distractors. In [11], cognitive workload was shown to cause irrelevant distractors to be more attended to; however, in [12] and [13], mixed effects were reported. Some WM manipulations were consistent with the effect observed in [11], but others had the opposite effect.

There are many ways to explore the relationship between WM and the visual system under workload. The workload can be imposed by selecting a well-defined controlled environment. Alternatively, a real-world environment, such as driving, can be used. The behavior of the visual system can be measured at both higher levels of attention and lower levels of eye gaze directions. These differences may account for some of the discrepancies in the literature.

Quantitative analysis has gained momentum over the last few years: Databases have been collected, tasks defined, and prediction algorithms proposed and evaluated. The databases cover a broad set of domains. For example, [14] and [15] are composed of everyday activities, such as talking on the phone and eating. The database in [16] recorded eye gaze distributions collected during the classification task of identifying breeds of dogs. The Dr(eye)ve database focuses on driving scenarios, and was collected on the road. It has roughly half a million frames. Some frames include GPS and speed data as well [17]. Eye gaze related databases can be used to obtain insight into human behavior, predict workload, and better understand the basic features of attention. The recent trend toward deep learning in the field of machine learning has led to the introduction of a set of algorithms to model both saliency distributions and eye gaze behavior under task loads. The use of a convolutional neural network (CNN) for salience estimation is surveyed by [18]. An example of task related gaze modeling was presented by Murabito *et al.* [16]. They trained a CNN to predict gaze distribution, while performing a task and showed that this distribution was very different from a distribution collected while no task was present. In addition, they explored the gaze

Manuscript received September 28, 2018; revised February 18, 2019 and April 30, 2019; accepted June 11, 2019. Date of publication August 30, 2019; date of current version November 21, 2019. This article was recommended by Associate Editor L. Rothrock. *(Corresponding author: Ron Moshe Hecht.)*

R. M. Hecht is with the Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel, and also with Advanced Technical Center Israel, General Motors, Herzliya 46733, Israel (e-mail: hadasron@gmail.com).

A. B. Hillel is with the Department of Industrial Engineering and Management, Ben-Gurion University of the Negev—Faculty of Engineering Sciences, Beersheba 8410501, Israel (e-mail: aharon.barhillel@gmail.com).

A. Telpaz is with the Advanced Technical Center Israel, General Motors, Herzliya 46733, Israel (e-mail: ariel.telpaz@gm.com).

O. Tsimhoni is with the General Motors Technical Center, Warren, MI 48092 USA (e-mail: omer.tsimhoni@gm.com).

N. Tishby is with the Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel (e-mail: tishby@cs.huji.ac.il).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2930996

distributions where the resolution of the original image was blurred.

In this article we suggest a simple quantitative model that accounts for the tradeoff between saliency-based and task-related gaze behavior, and show that this model accurately captures human gaze behavior in a set of controlled experiments. We draw on a reinforcement learning variant introduced by [19] and [20], known as the information constrained control (ICC). The ICC can be seen as a partially observed Markov decision process (POMDP) that models the interaction between the perception, states, and actions of an organism. It focuses on the information flow between the organism and its inner state and the environment. The observations are the transmissions of information from the environment to the organism, and the actions and their control are the transmissions in the other directions. Note that the ICC differs from the POMDP in its ability to represent two subgoals pursued by an organism by maximizing the reward, while minimizing the control complexity to perform the action.

Based on the ICC, our model formally describes the tradeoff between saliency-based and task-related gaze behavior as a constrained minimum divergence problem. Saliency-based gaze behavior is formalized as a distribution over the possible gaze locations, which are considered to be the set of possible actions. This is a comfort-zone distribution: The human agent tries to stay as close to it as possible. The task-related signal is introduced using a reward function defined over the action set, such that each possible gaze location is associated with a certain task-related reward. In order to meet task demands, the agent needs to achieve a certain average reward level. The agent's policy is determined by finding the distribution nearest to the comfort zone, in the Kullback–Liebler divergence sense, which still satisfies the task-related linear constraint. Different levels of workload, introduced by more stringent task conditions or performance of a secondary task in parallel, correspond in this framework to changes in a single problem parameter: The required reward level.

The suggested model is simple, but because of this simplicity it has nontrivial consequences leading to nontrivial empirical predictions. First, it intuitively predicts that average saliency levels of human agents under task pressure will drop monotonically with the level of workload introduced. Second, given two or more policies measured using different workload levels, the implicit reward function can be extracted up to multiplicative and additive constants, thus enabling reward visualization and model validation. The third important implication of the model is that the agent policies, viewed as log-probability vectors, constitute a curve in a three-dimensional linear subspace (spanned by the saliency distribution, the reward function, and the unity vector). Importantly, this low manifold structure allows for easy prediction of workload levels from gaze patterns.

We tested this approach empirically in a driving task in a simulator environment, which is realistic and immersive, but still enables the introduction of controlled workloads and exact measurements of eye gaze behavior. Workload in our experiments was manipulated by using driving related parameters: Road curvature level, location along the route, driving speed, and by performance of a parallel secondary task. Our analysis shows the remarkable fit of the model to the empirically recorded gaze behavior of 18 subjects. Three types of higher workload caused the drivers to deviate further from the saliency model. In certain cases, the reward function could be visualized, and corresponded to a task-relatedness intuition. The effective rank of the empirical policy is shown to be close to three. We then tested the model's prediction abilities using a binary classification task: Given two gaze patterns, determine which corresponds to driving with a higher cognitive load. The model-based classifier was able to solve the problem with close to a 90% success rate, and was found to outperform a baseline classifier that relied solely on standard deviations (SD) of eye position considerations.

This article contributes to the literature by introducing a simple quantitative model unifying bottom-up saliency-based gaze models and top-down task-related behavior, in addition to confirming the model's fit to empirical human gaze data. Beyond its empirical success, our model provides an elegant way of deriving gaze behavior from a very basic rationale: The ICC [19], [20] and the minimum discrimination information principles [21], a natural extension of the principle of insufficient reason. Thus, human gaze behavior is "simple" in the sense that it deviates the least from the prior saliency-based gaze distribution.

II. ICC Model

This section presents the formal foundations to the approach used in this article, which is based on [19], [20], and [22]–[25]. In a nutshell, the ICC is a satisfier model. It finds a balance between contradictory goals. The goal of the ICC is not only to maximize the reward an organism receives, but its ultimate goal also takes the motivation to minimize the complexity and information needed to execute a selected behavior into account as well. In the rest of this section, we introduce a single-state derivative to the ICC from scratch and build on top of it. This derivative is formalized using a set of possible actions $A = a_1, \ldots, a_n$ (where n is the number of actions), and two functions defined over this set: A prior distribution $Q(a)$ and a reward function $R(a)$. Intuitively, from a behavioral perspective, Q can be viewed as a comfort zone of behavior. In our case, it is the saliency distribution. The distribution $Q(a)$ is the unintentional behavior, i.e., $Q(a)$ is the probability of performing action a when no task pressure is present. Each action is associated with a specific value of task-related reward $R(a)$. The actual behavior $P(a)$ of the organism according to our model is defined as the solution to the following inference problem:

$$\begin{aligned} \hat{P}(a) = \arg\min_P \quad & \sum_{a\in A} P(a)\log\frac{P(a)}{Q(a)} \\ \text{s.t.} \quad & \sum_{a\in A} P(a) = 1 \\ & \sum_{a\in A} P(a)R(a) \geq \theta. \end{aligned} \tag{1}$$

The term $\sum_{a\in A} P(a)\log\frac{P(a)}{Q(a)}$ is known as Kullback—Leibler divergence (D_{KL}) [26]. The D_{KL} measures the information costs associated with selecting distribution P over the set a. This measure is relative to the distribution Q, i.e., Q is a baseline distribution. The information cost associated with selecting P to be identical to Q ($P = Q$) is zero. Intuitively, the organism

is constrained to achieve a certain level of task-related reward θ ($\sum_{a\in A}P(a)R(a)=\theta$), but under this constraint, it minimizes the D_{KL} of its behavior from prior behavior $Q(a)$. Equation (1) describes a constrained minimization problem. A common way to solve such problem is to transform it into a different unconstrained minimization problem (known as a Lagrangian, and denoted by $L\{P(a)\}$) that has the same solution, and solve it. This methodology is known as Lagrange multipliers [27]. The Lagrangian contains both the constraining terms and the minimized term as a single term that is later minimized. It is shown in the following equation:

$$L\{P(a)\}=\sum_{a\in A}P(a)\log\frac{P(a)}{Q(a)}-\gamma\left(\sum_{a\in A}P(a)-1\right)-\beta\left(\sum_{a\in A}P(a)R(a)-\theta\right). \tag{2}$$

The minimum of the Lagrangian is found by deriving it with respect to β, γ and the probability of each action $P(a)$. The solution with respect to $P(a)$ has the following form:

$$P(a)=\frac{Q(a)e^{\beta R(a)}}{Z(\beta)} \tag{3}$$

where Z is a normalization factor assuring that $\sum_{a\in A}P(a)=1$. Insights can be gained by looking at β as a tradeoff parameter selected implicitly by the organism. It provides a tradeoff between the maximization of the reward and the proximity to the prior distribution Q. In this case, (1) has the following form:

$$\hat{P}(a)=\arg\min_P\ \sum_{a\in A}P(a)\log\frac{P(a)}{Q(a)}-\beta\left(\sum_{a\in A}P(a)R(a)\right)\quad \text{s.t.}\quad \sum_{a\in A}P(a)=1. \tag{4}$$

A. Divergence From Saliency Q

Under this formulation, the organism's policy has one degree of freedom: The change in θ. The distribution of actions should be similar to $Q(a)$ when little reward is needed and should drift away when higher reward demands are set. We associate the reward constraint parameter θ with workload. A high task-related workload means that the task is more demanding, which is formulated as a demand for higher θ values in our model. The indicator for the workload is thus the likelihood of the organism's behavior (actions) according to the distribution $Q(a)$. At a lower workload, the likelihood of $Q(a)$ should remain high because the organism behaves according to the non-task-related behavior, whereas at a higher workload, the likelihood of $Q(a)$ should be lower. For ease of notation, we refer to the likelihood score of samples according to $Q(a)$ as the Q score.

B. Estimating the Reward Function

In cases where the reward function is unknown and several distributions obtained under varying workload conditions are observed, the reward function can be determined up to multiplicative and additive constants, thus enabling its visualization. By taking the log of the result of (1) and rearranging terms, we can see that

$$\log\frac{P(a)}{Q(a)}=\beta R(a)+\log Z(\beta). \tag{5}$$

In other words, there is a linear relationship between the reward and the log likelihood ratio. This relationship can be extended to the likelihood ratio between behavioral distributions produced under different workload conditions: $P_h(a)$, $P_m(a)$, and $P_l(a)$ (where h, m, l are the high, medium, and low workloads, respectively). These three behavior patterns can be expressed as

$$\log\frac{P_{h,m,l}(a)}{Q(a)}=\beta_{h,m,l}R(a)+\log Z_{h,m,l} \tag{6}$$

where $\beta_{h,m,l}$ and $Z_{h,m,l}$ define three different linear equations, one for each choice of h, m, l. Subtracting the equation for low workload from the equation for high workload yields a $Q(a)$ free solution

$$\tilde{R}(a):=\log\frac{P_h(a)}{Q(a)}-\log\frac{P_l(a)}{Q(a)}=\log\frac{P_h(a)}{P_l(a)}=(\beta_h-\beta_l)R(a)+\log(Z_h-Z_l). \tag{7}$$

In (7) we see that the log likelihood ratio between policies that differ in reward level estimates the reward up to multiplicative and additive constants. We denote this estimate by $\tilde{R}(a)$. Intuitively, $\tilde{R}(a)$ maintains a linear relation with respect to $R(a)$. The $\tilde{R}(a)$ values themselves are relative and only useful when compared to other $\tilde{R}(a)$ values. This might seem like a limitation; however, recall that in the larger set of RL tasks, the reward is defined as relative in the first place. More formally, the average property holds for $\tilde{R}(a)$

$$\frac{\tilde{R}(a_i)+\tilde{R}(a_j)}{2}=(\beta_h-\beta_l)\frac{R(a_i)+R(a_j)}{2}+\log(Z_h-Z_l) \tag{8}$$

for any i, j.

C. Low Rank Structure

The results of (1) can be rewritten as

$$\log P(a)=\log Q(a)+\beta R(a)-\log Z(\beta)\quad\forall a. \tag{9}$$

Denote by $\log\vec{P}=[\log P(a_1),\ldots,\log P(a_n)]$ the vector of log likelihood of actions. Similar vectors can be defined for the prior distribution $\log\vec{Q}$, and the reward $\vec{R}$. We have

$$\log\vec{P}=\log\vec{Q}+\beta\vec{R}-\log Z(\beta)\vec{1}. \tag{10}$$

$\log\vec{P}$ located in the subspace spanned by $\log\vec{Q}$, $\vec{R}$, and $\vec{1}$. The latter equation shows that regardless of the number of actions, all the policies generated by an ICC model are located on a three-dimensional subspace within the policy space. This type of expected behavior can be verified on real data.

D. Prediction of Workload From Gaze Pattern—$w_h w_l$ Score

The reward can be eliminated by using the solution to (1) twice (once for a high workload condition and once for a medium workload condition); by rewriting the relationship between $P_h(a)$ and $P_m(a)$

$$\frac{\log \frac{P_h(a)}{Q(a)} + \log Z(\beta_h)}{\beta_h} = R(a) = \frac{\log \frac{P_m(a)}{Q(a)} + \log Z(\beta_m)}{\beta_m} \tag{11}$$

where β_m, β_h are the β values associated with the medium and high workload conditions. By transforming to a vector representation, (11) becomes

$$\log \vec{P}_m = \frac{\beta_m}{\beta_h} \log \vec{P}_h + \left(1 - \frac{\beta_m}{\beta_h}\right) \log \vec{Q} + \left(\frac{\beta_m}{\beta_h} \log Z(\beta_h) - \log Z(\beta_m)\right) \vec{1}. \tag{12}$$

In a manner similar to (12), we can define the following equation:

$$\log \vec{P}_m = \frac{\beta_m}{\beta_l} \log \vec{P}_l + \left(1 - \frac{\beta_m}{\beta_l}\right) \log \vec{Q} + \left(\frac{\beta_m}{\beta_l} \log Z(\beta_l) - \log Z(\beta_m)\right) \vec{1}. \tag{13}$$

$\log \vec{Q}$ can be eliminated by combining (12) and (13)

$$\log \vec{P}_m = (1-w) \log \vec{P}_l + w \log \vec{P}_h + c_{\beta_h \beta_m \beta_l} \vec{1} \tag{14}$$

where the coefficients of $\log \vec{P}_h$ and $\log \vec{P}_l$ are denoted as w and $(1-w)$, respectively,

$$(1-w) = \frac{\beta_m - \beta_l}{\beta_h - \beta_l} \tag{15}$$

and

$$w = \frac{\beta_h - \beta_m}{\beta_h - \beta_l} \tag{16}$$

and the constant

$$c_{\beta_h \beta_m \beta_l} = \frac{(\beta_h - \beta_l)(\beta_m - \beta_l)}{\beta_h \beta_m (\beta_h - \beta_m)} \times (\beta_h \log z(\beta_l) - \beta_l \log z(\beta_h)). \tag{17}$$

The result shows that all possible policies are located in the subspace spanned by the vectors $\log \vec{P}_h$, $\log \vec{P}_l$, and $\vec{1}$. As observed in the previous section, this is a three-dimensional subspace. Furthermore, given that $\beta_h > \beta_m > \beta_l$, the two coefficients of w and $(1-w)$ are between zero and one ($0 \leq w \leq 1$) and sum to one. In other words, all the solutions are a convex combination of the highest and lowest workload solutions. Later, we refer to w as w_h, the high workload component and to $1-w$ as w_l as the low workload component.

Intuitively, this finding makes more sense if we explore the two-dimensional subspace spanned by the coefficients w of $\log \vec{P}_h$ and $\log \vec{P}_l$. In this space, the subspace of possible policies forms a line. Specifically, (14) suggests that one mechanism to estimate workload for a given log-likelihood vector $\log \vec{P}$ would be to calculate a linear regression between it and the vectors $\log \vec{P}_h$, $\log \vec{P}_l$, $\vec{1}$. The workload for $\log \vec{P}$ is w, the coefficient in the regression of $\log \vec{P}_h$.

Fig. 1. Simulator setup—NADS MiniSim with three screens.

III. Method

The experiments were carried out in a driving simulator, an environment that mimics a high workload everyday task. This environment enables accurate gaze pattern estimation while introducing variable visual stimuli and variable levels of task-related workloads.

A. Participants

Twenty healthy volunteers (nine females), ranging in age from 25 to 59 (Mean = 33), were recruited to participate in the study. The participants were required to have a valid driving license for at least three years, and confirm that they drove on a daily basis. All were naïve to the purpose of the study. Participants stated they had normal (or corrected to normal) vision. Prior to the start of the experiment, the participants gave their informed consent in compliance with the guidelines of the General Motors Institutional Review Board. At the end of the experiment, each participant was paid a fixed fee of 200 NIS (about $50).

B. Apparatus

1) Driving Simulator: The experiment was conducted in a low fidelity driving simulator, the NADS Mini-Sim, which is a driving simulator developed by the National Advanced Driving Simulator (NADS) and the University of Iowa [28]. The simulator consists of three $42''$ monitors, which are used to display the frontal and side road views, yielding a 130° field of view (see Fig. 1). An additional screen mounted in the simulator cockpit dashboard was used to present the gauges on the instrument panel. Simulator-related sounds were delivered through a 2.1 audio system. Just like real-world vehicles, the simulator was equipped with a steering wheel, as well as gas and brake pedals, which allowed the participants to operate the simulator vehicle naturally. The simulator software logs information on the participants' driving behavior and location on the route at a 60 Hz rate. In addition, the images displayed on the three simulator monitors are captured and saved at an average rate of about 10 Hz.

2) Eye Tracking: The Smart-Eye pro 5 eye tracking system (manufactured and developed by Smart Eye AB, Gothenburg,

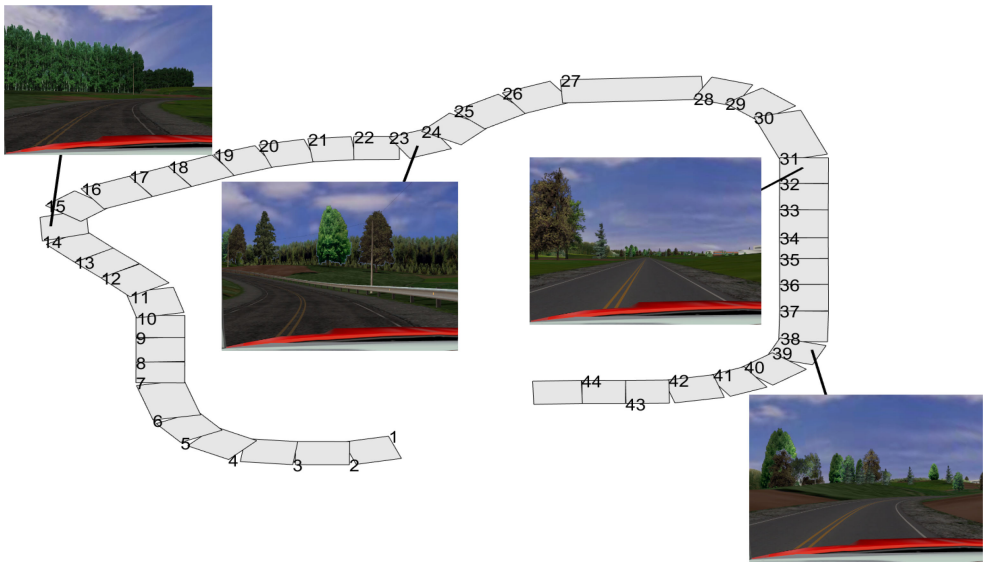


Fig. 2. Route driven in the experiment. The route started at segment 1 and ended at segment 44. The snapshots depict several representative segments along the route.

Sweden; http://smarteye.se/) was used to track the participants' eye movements and point of gaze during the experiment. The system is composed of two IR cameras and IR LEDs. The two cameras were positioned on the upper part of the dashboard facing the participant. This eye-tracking system exploits corneal reflections from the IR LEDs at the center of the pupil to estimate the gaze intersection with the screens. The calibration process of the system takes place in two stages. In the first stage, the system generates a head model of the participant, followed by a gaze calibration. During the gaze calibration phase, the participants are asked to look at 12 different locations in the vehicle cockpit. The process ends when the calibration quality is high enough, as determined by the average and SD of the shift between the recorded gaze points and the 12 target locations. Specifically, the average and SD were set to below 2°. The eye tracking data were collected at a rate of 60 Hz.

C. Design and Procedure

Prior to the start of the experiment, the participants were requested to fill in a simulator sickness questionnaire (SSQ) [29] to help avoid simulator sickness while driving in the simulator. Based on the results, two participants were eliminated. This left us with 18 participants. After the eye tracking calibration was completed, the participants engaged in a 7-min training scenario, which enabled them to experience and learn how to operate the simulator vehicle. After this training session, the participants were given instructions on the experimental scenarios. The study was a within-subjects design consisting of five experimental scenarios divided into two blocks of two-three scenarios each. In all five scenarios, the participants drove an identical route. Fig. 2 shows the driving path used in the experimental scenarios. The route path length was about 5 mi. The entire drive was in cruise control mode so that driving was at constant speed regardless of how the driver behaved. The driving took place in a simulated rural environment. The path involved different curvature levels including straight road segments. The scenery was relatively dull and monotonic without radical visual changes. In addition, for simplification purposes, no other vehicle except for the participant's vehicle were involved in the scenario. The participants were informed that they would not encounter any other vehicles during the scenario.

D. Independent Measures

To evaluate our model predictions for gaze pattern under different levels of workload, three independent factors were manipulated: N-Back task, vehicle speed, and curvature level.

1) N-Back (2-Back) Task: During the two scenarios in the second block, and for each participant, the experimenter read aloud random numbers between 1 and 9 at intervals of about 20 s. After each number that was read to the participants, they were asked to state whether the number was larger or smaller than the number presented to them two steps earlier. This task imposes additional WM storage demands [10] and is known to interfere (i.e., increase drivers' mental workload) on a driving task in dual task settings [8]. In the first block, there was no secondary task. Since driving improves over time, administering the secondary task during the second block would be indicative that N-Back had an even more significant effect.

2) Vehicle Speed: As shown previously, higher speed levels are associated with higher cognitive workload [30]–[32]. For instance, [31] found that the addition of a secondary demanding task caused drivers to reduce their vehicle speed, suggesting that it enabled them to allocate more cognitive resources to the secondary task. Therefore, within each block, in one scenario, the speed was set to 35 mi/h, in another scenario, it was set to 55 mi/h, and in the first blocks, a 45 mi/h scenario was added (see Table I). The order of the scenarios within each block was counter balanced across participants. The speed values were

TABLE I
DRIVING SCENARIOS IN THE EXPERIMENT

Block	Scenario	Speed	N-Back
1	1	35	without
	2	45	without
	3	55	without
2	1	35	with
	2	55	with

Within each block, the order of scenarios was counterbalanced across participants.

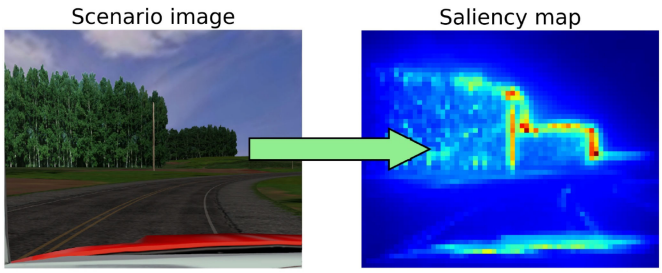


Fig. 3. Illustration of the generation process of the likelihood of Q scores. The figure on the left is a scenario image. The figure on the right shows the corresponding GBVS probability values. The Q score is simply the average of the log value of the pixels in the gaze location coordinates.

selected in such way that it would be possible to drive the entire route without having to press the brake pedal.

3) Curvature Level: Tsimhoni and Green [33] reported a relationship between visual demands and curvature, and showed in particular that there is an increase in task demand even before the curvature starts. For this reason, the route was made of several curved and straight segments. As shown in Fig. 2, the path was divided into 44 driving segments and curvature was estimated in each of them.

E. Dependent Measures

1) Likelihood of Q (Q Score): The likelihood of Q was measured by evaluating the similarity between the observed gaze pattern and the expected gaze pattern according to the graph based visual saliency (GBVS) model [1]. Similarity was estimated on a frame-by-frame basis and then averaged for each of the driving route segments. Specifically, for every frame, a GBVS activation map was generated and the activation values of the pixel in the participant's (x, y) gaze location coordinates (Q score) were averaged for each driving route segment (see example in Fig. 3). In some instances for individual participants, there were several gaze location samples within the same frame. In these cases, the Q score represents the average of the log GBVS probability value for these samples. This analysis was only conducted for the central screen in the simulator because the eye tracking accuracy was lower for the side screens. This did not create bias in the analysis outputs since the proportion of gazes to the side screens was very low. The Q scores were normalized by subtraction of the participants' mean Q score.

2) Statistical Analysis: To examine the statistical significance of the Q score predictions under different levels of workload, we conducted a repeated measures analysis of variance (ANOVA), with N-Back condition (with versus without), speed level (35 versus 55), and driving route segment as within subjects factors. To ensure the quality of the data analysis presented in the results section, any region that did not include data from all participants was eliminated from subsequent analysis. Therefore, due to the technical limitation of the eye tracking system, only 21 regions were taken into account for the analysis presented in the results section. For each analysis, we report the effect size using the partial eta squared (η_p^2) where values from 0.13 to 0.40 typically represent medium to large effects, respectively.

3) Visualization of the Reward: The estimation of the reward $\tilde{R}(a)$ was measured using (7). This was measured by selecting two gaze log likelihood distributions.

4) Dimension of Low Rank Structure: As (10) suggests, all the eye gaze distributions were expected to be located in a three-dimensional subspace. Here, the dimension of the subspace was estimated using principal component analysis (PCA) [34]. More specifically, the drivers' actions, which are the equivalent of sight locations, were measured at high resolution (dimension 1024×1280). Later, we down sampled the raw data to concentrate the probabilistic mass (dimension 21×26). Then, rarely observed actions were removed, yielding the policy vector $\log \vec{P}$ in a low dimensional action space (dimension 21). A $\log \vec{P}$ vector was generated for each driver, driving condition, and road segment, yielding roughly 100 vectors per segment. For each segment, based on its roughly 100 vectors, a covariance matrix was estimated. The PCA algorithm receives a covariance matrix as input, and identifies a set of orthogonal vectors with high variance. These vectors are known as eigenvectors. Each eigenvector is associated with a scalar known as an eigenvalue. The eigenvalue indicates the importance and variance of that direction. The dependent measures are the set of eigenvalues for each segments.

IV. RESULTS

For one participant there were abnormal Q scores exceeding more than three absolute SD from the sample mean. Therefore, this participant was excluded from further analyses, leaving us with 17 participants.

A. N-Back, Speed and Driving Route Segment Main Effects

As seen in Fig. 4, the Q score was lower (M = 0.16, SD = 0.006) in the N-Back condition compared to the without N-Back condition (M = 0.17, SD = 0.008). This difference between the two conditions was significant ($p < 0.05, \eta_p^2 = 0.28$). As stated in [13], the effect of WM manipulation can be similar or opposite to the one observed at [11]. Here the effect was opposite to [11], and was similar to perceptual workload. Similarly, speed had a significant effect on the Q score, with higher Q score values when the speed was set to 35 mi/h (M = 0.17, SD = 0.007) compared to 55 mi/h (M = 0.16, SD = 0.007); $p < 0.05, \eta_p^2 = 0.28$. Not surprisingly, region was also found to be a significant factor

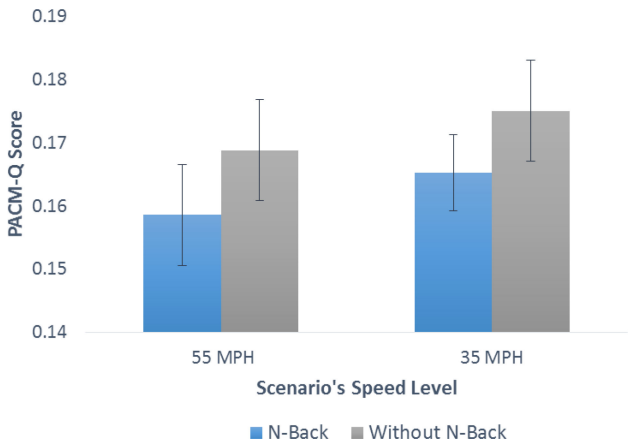


Fig. 4. Q score as a function of speed level and N-Back condition. Error bars denote standard error of the means.

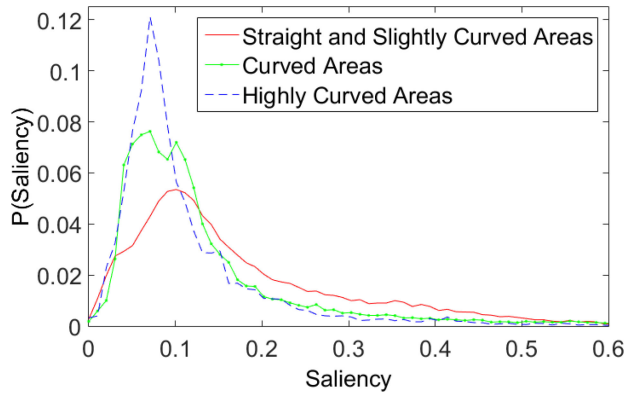


Fig. 5. Distributions of Q scores for three curvature levels.

($p < 0.0001, \eta_p^2 = 0.81$), indicating that the Q scores varied among the 21 driving route segments that were included in the analysis. Since the main differentiator between the driving segments was the curvature level, we conducted further analyses to examine the association between curvature and Q score.

All the participants performed the N-Back task better than random, indicating that the task was not ignored. The worst performer achieved 73% accuracy. There was no significant difference between the performance of the task at 35 and 55 mi/h. There was no monetary award for high performers on the N-Back task.

B. Q Score and Curvature Level

For each frame, we matched its Q score to the road curvature level. The analysis involved four scenarios $[35, 55]X[\text{with, wo}]$ and all participants, totaling more than 100 000 frames. The aim of this analysis was to examine the shape of the Q score distribution under different levels of curvature. We divided the curvature spectrum into three parts: Straight and slightly curved road, curved road, and highly curved road. As can be seen in Fig. 5, the higher the curvature (i.e., the more demanding the task), the lower the mean Q scores (mean of straight and slightly curved: 0.176 mean of curved road: 0.136 mean of highly curved road: 0.115). Interestingly, the Q score distribution became narrower under higher levels of curvature (SD of straight and slightly curved: 0.131 SD of curved road: 0.119 SD of highly curved road: 0.095).

C. Visualization of the Reward

The mechanism to estimate reward was presented in Section III; it boils down to estimations of the log likelihood ratio between two gaze distributions. Fig. 6 shows the visualization capabilities of the ICC. The figure presents the ratio between the gaze log likelihood distributions. The distributions were smoothed based on likelihood and in places where no gazes were observed, only the background is seen. The remainder of the areas could be divided into cold and warm colored blobs. The warm colored blobs represent less important areas. The cold blobs are more important from a reward perspective. These areas were located on the road itself as expected, whereas the low reward areas were located on the sides of the road. The high reward area had a nontrivial shape.

D. Low Rank Structure

As explained in Section III, the vectors used to estimate the covariance matrix were generated using a three-stage process. A visualization of the transformation during this process is shown in Fig. 7. The original high resolution distribution is shown in Fig. 7(a) (dimension 1024×1280). The down sampled distribution is presented in Fig. 7(b) (dimension 21×26). The distribution after the removal of rare actions is presented in Fig. 7(c) (dimension 21).

After the $\log \vec{P}$ vectors were extracted, a covariance matrix was estimated. The eigenvalues of the covariance matrix were used to gain insight into the dimensions of the subspace. Fig. 8 shows the sum of the first N normalized eigenvalues. The sum of the eigenvalues was averaged across the segments and the SD is presented as well. The figure shows that the first three eigenvectors included more than 0.80 of the overall signal variance. It depicts the result when vectors from different participants were used; hence, the unexplained variance tail (the energy in eigenvectors >3) can be attributed to differences between individuals.

E. Load Prediction From Gaze Patterns

As described in Section II, the coefficients of $\log \vec{P}_l$ and $\log \vec{P}_h$ (denoted w_l and w_h) sum to one. If they are chosen as the two extreme conditions, any other intermediate condition $\log \vec{P}_m$ is a convex combination of the two extremes. In our case the two extreme vectors corresponded to 35 mi/h without N-Back and 55 mi/h with N-Back.

A leave-one-out approach was selected to define the vectors $\log \vec{P}_l$ and $\log \vec{P}_h$. for each tested driver and route segment, where the reference vectors $\log \vec{P}_l$ and $\log \vec{P}_h$ were the average vectors in scenarios 35 mi/h without N-Back and 55 mi/h with N-Back of all other drivers in that segment except the tested driver. The tested driver vector was expressed as a linear

Fig. 6. (a)–(c) present $\tilde{R}(a)$ values—the reward visualization capability of the ICC. As suggested by (8), the reward is estimated up to multiplicative and additive constants, i.e., the units themselves are arbitrary and only the relations among them count. In areas where almost no gazes were observed, only the original image is seen. Area colors with hot colors represent less important areas (low $\tilde{R}(a)$ values), and areas with cold colors represent the more important areas [high $\tilde{R}(a)$ values]. The cold colored areas (more important) are mostly located on the road itself, while the hot colored areas (less important) are mostly located off road. Since those values are relative, there is no need to present the exact value. Distributions were estimated based on data from all participants. Three distinct locations along the route where selected to demonstrate the route diversity. In addition, each image depicts reward visualization that was generated from contrast between different driving conditions. (a) was generated from the contrast between 55 mi/h N-Back and 35 mi/h N-Back conditions. (b) was generated from the contrast between 55 mi/h N-Back and 35 mi/h no N-Back conditions. (c) was generated from the contrast between 55 mi/h N-Back and 45 mi/h no N-Back conditions.

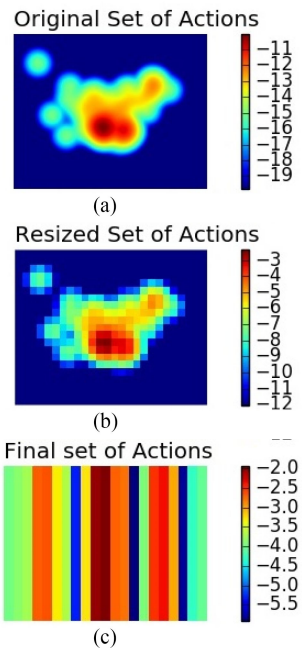


Fig. 7. Action definition. (a) Raw distribution of eye gazes. (b) Distribution after down sampling. (c) Selection of dominant actions. All subfigures present log likelihood values.

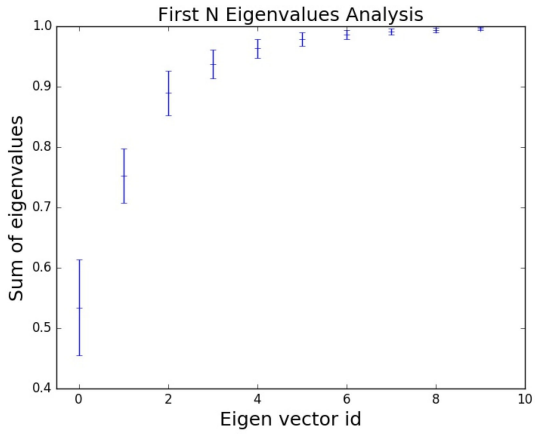


Fig. 8. Sum of the first N normalized eigenvalues of the distributions averaged over route segments.

combination of the two reference vectors and the constant $\vec{1}$. The results are shown in Fig. 9. It is clear that the actual findings are located along the line that was predicted by the model. The color represents the workload condition in different segments. Even a cursory glance shows the different colors at different locations across the curve.

As the model suggests, w_h is a workload measure. When closer to 0, low workload is assumed, and high w_h values are associated with high workload. Another measure for workload is the SD along the x axis [8]. We used the SD to set a reference baseline for recognition performance of workload (see Table II). A comparison within pairs of conditions for that hierarchy of workload was conducted. About 800 comparisons were made within each pair of conditions. On average, our approach outperformed the predictions using SD measurements.

As mentioned in Section I, WM workload is associated with mixed effects. In our experiment, it led to a reduction in the gaze SD. This suggests some disparity between our findings and some of the attentional findings in the psychology literature. However, note that the experiments were not identical and differed in their manipulations, measurements, and environments.

V. Discussion

This study examined whether an ICC approach could account for gaze behavior under workload. Specifically, we modeled eye gaze, while participants drove in a simulator under varying workload conditions. We assumed that the baseline (workload

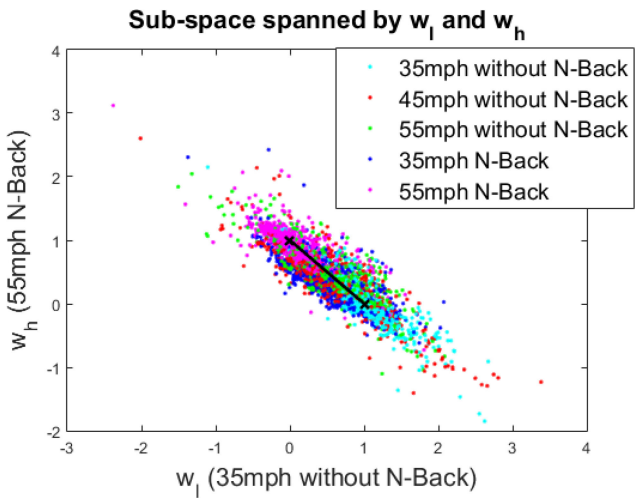


Fig. 9. Expected and observed tradeoff between high and low workload in the w_l and w_h subspace. We refer to w as w_h and to $w-1$ as w_l. Recall that according to(15) and (16), the sum of w_h and w_l equals one. In addition, $0 \leq w_h, w_l \leq 1$. This suggests that theoretically all the possible combinations of w_h, w_l form a line between the point $(1,0)$ and the point $(0,1)$. This line is represented by the solid black line in the figure. The dots represent empirical estimations of w_h, w_l for different segments and driving conditions. The empirical data forms a scatter that is located close to the theoretical black solid line. The colors of the dots represent the driving conditions. The cyan represents the simplest ride condition (35 mi/h no N-Back) and the magenta the hardest one (55 mi/h N-Back). They are located on the opposite ends of the points cloud. The other driving conditions/colors are located more toward the middle of the cloud.

TABLE II
WORKLOAD RECOGNITION PERFORMANCE FOR DIFFERENT PAIRS OF SCENARIOS

Scenarios Comparison	SD	$w_h w_l$ score
35 vs 45	0.64	0.81
35 vs 55	0.82	0.90
45 vs 55	0.77	0.63
35 vs 35 N-Back	0.81	0.85
35 N-Back vs 55 N-Back	0.81	0.96
35 vs 55 N-Back	0.93	0.99
45 vs 55 N-Back	0.88	0.95
55 vs 55 N-Back	0.69	0.91
Overall correct	0.79	0.88

SD is the SD score, and $w_h w_l$ score as presented in (15)

free) distribution Q would be estimated by a visual saliency mechanism (more specifically, by GBVS). We tested the efficiency of the model, in terms of the relationship between workload and deviation from the baseline gaze distribution, estimation and visualization of the reward, existence of a low rank structure, and workload detection capabilities.

A within-subject repeated measures ANOVA was used to test the deviation from the baseline distribution Q. The results showed that curvature, speed, and N-Back, which are known to increase driver workload, caused deviations from the baseline gaze distribution. We increased the workload by two manipulations of the main task and one manipulation by introducing a secondary task. All the manipulations caused the gaze pattern to shift from a more bottom-up pattern to a more top-down pattern, in a way that was clearly and significantly measurable using saliency Q scores.

These findings thus show that the model generates an accurate mechanism for reward visualization. This visualization was possible in cases where we had two gaze distributions that were estimated under two workload conditions (high and low). This visualization is of potential interest to researchers working on task-related perception, in that it can identify the areas in the field of view that are important to the participants. Although the reward is only visualized up to a linear transformation, it is nevertheless a powerful technique, since our estimation maintains the reward relations among the locations in the field of view. Whenever the reward associated with location A is higher than location B, the visualization value of A is higher than the value associated with point B. More specifically, the locations with the highest and the lowest rewards are the ones with the highest and lowest computed and visualized values. Hence the visualization image generated by the model identifies the most and the least important areas in a scene. Our examples show that in our experiments we were able to identify the important gaze locations for a driver entering a curve.

Our entire analysis is focused on the model space (distribution space). In this space, each eye gaze distribution is represented by a single point. For discrete distributions, the dimension of this space is the number of elements in the distribution minus one. For example, all the Bernoulli distributions $(p, 1-p)$ are points in a one-dimensional space. The larger the number of elements, the higher the dimension becomes. Workload manipulation causes a shift from one point in the distribution space (from one distribution) to another point in that space (another distribution). When a set of workload manipulations is applied, each manipulation yields a different point in the distribution space (different distribution). The ICC predicts that all these points are embedded in a low-dimensional subspace in the distribution space. Fig. 8 shows that the majority of the signal is located in the first dimension. A single dimension has more energy than all the rest of the dimensions. The first three dimensions have about 90% of the energy. This suggests an ICC-like behavior where all the distributions are highly bounded and do not tend to exit a small subspace.

Not only are all the points embedded in the small dimensional subspace, but the ICC also suggests that these points form a curved line in that subspace. This line can be projected to a straight line in a certain weighted space. In this geometric description, the task-free distribution is the initial point of the curve, the goal-only distribution is its terminal point, and β is the index of the location along the curve.

The distance between any two points along the curve can be measured. Specifically, one can measure the distance from any point to both ends of the curved line (one end is associated with the lowest workload and the other with the highest workload). Intuitively, normalization of these distances generates the $w_h w_l$ score. This article suggests that the w_h, w_l are a measure of workload. This was compared to the SD of a gaze position measure. We tested these prediction capabilities and found that the w_h, w_l predictions outperformed the commonly used SD

of gaze position measure, and achieved a correct identification rate of 0.88. Fig. 9 visualizes the w_h, w_l distances used in this comparison.

Several workload manipulations were compared, one of which was N-Back manipulation. This manipulation imposes a WM workload and is very different from the cognitive workload presented at [35], which involved memorizing a set of digits before the task began and testing whether the digits were recalled after the task ended, and has no perceptual features, unlike N-Back. Overall, while some cognitive workload [11], [35]–[38] causes participants to pay more attention to goal-irrelevant stimuli, other produce the opposite effect [39], [40]. In addition, it was shown that N-Back is associated with narrower eye gaze distributions [41]–[43]. It is worth bearing in mind that there are differences across the experiments, in terms of the manipulations used, the visual system measures, and experimental environment. Hence, the relationship among these factors is not straightforward, and further work is needed to address and understand these differences.

VI. Limitations and Conclusion

Overall, an ICC model with a saliency baseline emerges as an efficient tool to model driver gaze behavior under varying workload conditions. At a highly practical level, this model provides a set of useful tools for researchers.

One of the drawbacks of our experiments is that the commonly used mechanism of visual distractors and response time measurements was not used, which can preclude generalization. The effect we measured was achieved by varying other conditions, such as curvature. Future experiments should introduce visual distractors and test response time. However, the tendency of distractors to become very salient needs to be overcome, since this can alter the scene's baseline distribution.

In addition, future studies should focus on modeling tasks with more complex reward distributions. This type of approach would be better able to distinguish between the SD-based model and our suggested model. A delta-like or Gaussian distribution of reward is the least useful for discriminating between the two. Future experiments should attempt to generate a bimodal, or otherwise complex reward function, thus providing more interesting test cases for the theory.

More broadly, given the results presented here on the visual system and the findings presented in [22] concerning the role of ICC in word selection for dialog, it is worth exploring whether the ICC principle is effective in modeling other human behavioral contexts.

References

[1] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[2] L. Guanbin and Y. Yizhou, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, 5455–5463.

[3] M. N. Heess and A. Graves, "Recurrent models of visual attention," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[4] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[5] N. Lavie and S. Cox, "On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection," *Psychological Sci.*, vol. 8, no. 5, pp. 395–396, 1997.

[6] N. Lavie, "Attention, distraction, and cognitive control under load," *Current Directions Psychological Sci.*, vol. 19, no. 3, pp. 143–148, 2010.

[7] C. D. Wickens, "Multiple resources and mental workload," *Human Factors: J. Human Factors Ergonom. Soc.*, vol. 50, no. 3, pp. 449–455, 2008.

[8] T. W. Victor, J. L. Harbluk, and J. A. Engström, "Sensitivity of eye-movement measures to in-vehicle task difficulty," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 8, no. 2, pp. 167–190, 2005.

[9] U. Cartwright-Finch and N. Lavie, "The role of perceptual load in inattentional blindness," *Cognition*, vol. 102, no. 3, pp. 321–340, 2007.

[10] N. Cowan, "Processing limits of selective attention and working memory: Potential implications for interpreting," *Interpreting*, vol. 5, no. 2, pp. 117–146, 2000.

[11] N. Lavie, "Distracted and confused?: Selective attention under load," *Trends Cogn. Sci.*, vol. 9, no. 2, pp. 75–82, 2005.

[12] N. Konstantinou, E. Beal, J.-R. King, and N. Lavie, "Working memory load and distraction: Dissociable effects of visual maintenance and cognitive control," *Attention, Perception, Psychophysics*, vol. 76, no. 7, pp. 1985–1997, 2014.

[13] N. Konstantinou and N. Lavie, "Dissociable roles of different types of working memory load in visual detection," *J. Exp. Psychol.: Human Perception Perform.*, vol. 39, no. 4, pp. 919–924, 2013.

[14] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.

[15] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 2929–2936.

[16] F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, "Top-down saliency detection driven by visual classification," *Comput. Vision Image Understanding*, vol. 172, pp. 67–76, 2018.

[17] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "Dr (eye) ve: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016, pp. 54–60.

[18] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[19] J. Rubin, O. Shamir, and N. Tishby, "Trading value and information in MDPS," *Decision Making With Imperfect Decision Makers*. Berlin, Heidelberg, Germany: Springer, 2012, pp. 57–74.

[20] N. Tishby and D. Polani, "Information theory of decisions and actions," *Perception-Action Cycle*. New York, NY, USA: Springer, 2011, pp. 601–636.

[21] C. R. Smith, G. J. Erickson, and P. O. Neudorfer, *Maximum Entropy and Bayesian Methods*, vol. 50. Seattle, WA, USA: Springer Science & Business Media, 2013.

[22] R. M. Hecht, A. Bar-Hillel, S. Tiomkin, H. Levi, O. Tsimhoni, and N. Tishby, "Cognitive workload and vocabulary sparseness: Theory and practice," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3394–3398.

[23] R. M. Hecht, A. Bar-Hillel, A. Telpaz, and N. Tishby, "Adaptation of eye gazing while driving using a perception-action cycle model," in *Poster Without Proc. ICRI-CI Retreat*, 2016.

[24] R. M. Hecht, A. Telpaz, G. Kamhi, A. Bar-Hillel, and N. Tishby, "Disentanglement of top-down and bottom-up processes using information constrained control," in *Poster Without Proc. 5th Conf. Cognit. Res.*, 2018.

[25] R. M. Hecht, A. Telpaz, G. Kamhi, A. Bar-Hillel, and N. Tisbhy, "Information constrained control for visual detection of important areas," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 4080–4084.

[26] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[27] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[28] Y. He, NADS MiniSim driving simulator," Univ. Iowa, Iowa City, IA, USA, Rep. N06-025, 2006. [Online]. Available: http://www.nads-sc.uiowa.edu/publicationStorage/200610111034530.N06-025%%20NADS%20MiniSim.pdf

[29] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, 1993.

[30] C. D. Wickens, "Multiple resources and performance prediction," *Theor. Issues Ergonom. Sci.*, vol. 3, no. 2, pp. 159–177, 2002.

[31] R. Srinivasan and P. P. Jovanis, "Effect of in-vehicle route guidance systems on driver workload and choice of vehicle speed: Findings from a driving simulator experiment," in *Ergonomics and Safety of Intelligent Driver Interfaces*. Hillsdale, NJ, USA: L. Erlbaum Associates, 1997, pp. 97–114.

[32] A. H. Jamson and N. Merat, "Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 8, no. 2, pp. 79–96, 2005.

[33] O. Tsimhoni and P. Green, "Visual demand of driving curves determined by visual occlusion," in *Proc. Vision Vehicles 8 Conf.*, Boston, MA, USA, 1999, pp. 256–264.

[34] I. Jolliffe, "Principal component analysis," in *International Encyclopedia of Statistical Science*. Berlin, Heidelberg, Germany: Springer, 2011, pp. 1094–1096.

[35] N. Lavie, A. Hirst, J. W. De Fockert, and E. Viding, "Load theory of selective attention and cognitive control," *J. Exp. Psychol.: Gen.*, vol. 133, no. 3, pp. 339–354, 2004.

[36] J. W. de Fockert, G. Rees, C. D. Frith, and N. Lavie, "The role of working memory in visual selective attention," *Science*, vol. 291, no. 5509, pp. 1803–1806, 2001.

[37] N. Lavie and J. De Fockert, "The role of working memory in attentional capture," *Psychonomic Bull. Rev.*, vol. 12, no. 4, pp. 669–674, 2005.

[38] J. Rissman, A. Gazzaley, and M. Desposito, "The effect of non-visual working memory load on top-down modulation of visual processing," *Neuropsychologia*, vol. 47, no. 7, pp. 1637–1646, 2009.

[39] M. Rose, C. Schmid, A. Winzen, T. Sommer, and C. Büchel, "The functional and temporal characteristics of top-down modulation in visual selection," *Cerebral Cortex*, vol. 15, no. 9, pp. 1290–1298, 2004.

[40] K. K. Sreenivasan and A. P. Jha, "Selective attention supports working memory maintenance by modulating perceptual processing of distractors," *J. Cogn. Neuroscience*, vol. 19, no. 1, pp. 32–41, 2007.

[41] M. Niezgoda, A. Tarnowski, M. Kruszewski, and T. Kamiński, "Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 32, pp. 23–34, 2015.

[42] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "A field study on the impact of variations in short-term memory demands on drivers visual attention and driving performance across three age groups," *Human Factors*, vol. 54, no. 3, pp. 454–468, 2012.

[43] Y. Wang, B. Reimer, J. Dobres, and B. Mehler, "The sensitivity of different methodologies for characterizing drivers gaze concentration under increased cognitive demand," *Transp. Res. Part F: Traffic Psychol. Behav.*, vol. 26, pp. 227–237, 2014.

Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions

Wan-Lin Hu, Kumar Akash, Tahira Reid, and Neera Jain

***Abstract*—We developed an experiment to elicit human trust dynamics in human–machine interaction contexts and established a quantitative model of human trust behavior with respect to these contexts. The proposed model describes human trust level as a function of experience, cumulative trust, and expectation bias. We estimated the model parameters using human subject data collected from two experiments. Experiment 1 was designed to excite human trust dynamics using multiple transitions in trust level. Five hundred and eighty-one individuals participated in this experiment. Experiment 2 was an augmentation of Experiment 1 designed to study and incorporate the effects of misses and false alarms in the general model. Three hundred and thirty-three individuals participated in Experiment 2. Beyond considering the dynamics of human trust in automation, this model also characterizes the effects of demographic factors on human trust. In particular, our results show that the effects of national culture and gender on trust are significant. For example, U.S. participants showed a lower trust level and were more sensitive to misses as compared with Indian participants. The resulting trust model is intended for the development of autonomous systems that can respond to changes in human trust level in real time.**

***Index Terms*—Affective computing, autonomous systems, behavioral sciences, cultural differences, data models, human–computer interaction, man–machine systems, stability analysis.**

I. Introduction

THE widespread use of autonomy has improved the quality and efficiency of both safety-critical systems (e.g., nuclear power plants, aircrafts) and devices in daily life (e.g., cars, home appliances). In particular, human trust in autonomous systems is critical to improving the collaboration between humans and such systems [1]. Although various efforts have been made to optimize automated processes, the benefits of automation are lost when humans override these systems due to a fundamental lack of trust [2], [3]. Moreover, accidents may occur due to mistrust [4]. Therefore, trust should be appropriately calibrated to avoid disuse or misuse of automation [5]. One way to potentially overcome these negative effects is to design autonomous systems that can adapt to the human in real time based on the human's trust level. However, doing this requires predictive, reliable, and quantitative models of human trust behavior.

Researchers have studied trust behavior in human–machine interactions (HMI) and human–computer interactions (HCI) using experimental methods and modeling techniques from social psychology [2], [6], [7]. Some studies focused on analyzing the statistical significance of demographic factors (e.g., age, gender) on trust behaviors [8], [9]. However, while identifying factors that induce changes in trust is a critical step towards characterizing trust behavior; it is alone insufficient for characterizing a *quantitative* model of this behavior. Moreover, studies have shown that the trust level of humans varies with time due to changing experiences [6], [10] and, as such, any quantitative trust model should be dynamic.

In order to derive a quantitative dynamic model of human trust behavior suitable for HMI contexts, an appropriate experimental design, modeling approach, and model verification are necessary. There is no experimentally verified model for describing the dynamics of human trust level in HMI contexts that: 1) incorporates demographic factors and time-varying experiences; and 2) is built on experiments that elicit multiple transitions in trust level. Existing quantitative models either assume that human trust behavior is fully based on rationale [6] or are nonlinear [10], [11]. While the influence of accumulated effects of past interactions on the future trust level has been modeled in multiagent system contexts, they have not been modeled for independent HMI [11], [12]. Furthermore, existing models of human trust in autonomous systems have not taken into account neither human bias nor attitudes toward the system response bias; here, the *system response bias* is defined in terms of signal detection theory, i.e., liberal (false-alarm-prone) and conservative (miss-prone) system bias.

Finally, human behavior is highly influenced by one's surroundings and past experiences [13], [14], which, in turn, are strongly influenced by demographic factors. With the spread of automation across the globe, it is necessary to model human behavior for different demographics. Several types of autonomous systems such as cars, smart thermostats, and tour-guide robots are designed to interact with unspecified users. In these contexts, a model that describes the trust dynamics of a population (general or grouped by demographics) instead of an individual would facilitate the design of such systems. Unfortunately, a generalized model that is suitable for capturing these variations in human trust behavior does not exist in the current literature.

In this paper, we describe the modeling and experimental methods used to capture *dynamic changes* in human trust,

Manuscript received April 20, 2017; revised January 31, 2018 and May 21, 2018; accepted August 30, 2018. Date of publication October 23, 2018; date of current version November 21, 2019. This work was supported by the National Science Foundation under Award No. 1548616. This paper was recommended by Associate Editor Michael Dorneich. *(Corresponding author: Neera Jain.)*

The authors are with the School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: wanlinhu@stanford.edu; kakash@purdue.edu; tahira@purdue.edu; neerajain@purdue.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2874188

specifically in an HMI context. This paper significantly expands on our previously published conference paper [15] by including a second experiment that investigates the effects of system error types on trust level. We introduce an improved model structure that explicitly incorporates human bias as an input disturbance to the model. Furthermore, we augment the definition of experience, and the input to the model, to enable the integration of other factors that affect trust, such as system error type. We devise two sets of human subject experiments that elicit multiple dynamic transitions in human trust behavior, and the data collected from these experiments are then used to estimate and validate the parameters of the proposed model. We establish a linear model motivated by literature on computational models for the dynamic variation of human trust [6], [16]; the linearity allows for easier control analysis and synthesis aimed at designing adaptive human–machine interfaces, thus, enabling autonomous systems to respond to human trust variations in real time. Furthermore, we systematically analyze the effects of demographic factors, consisting of national culture [17] and gender, as well as system error type.

This paper is organized as follows. First, we provide background on trust modeling and significant factors that affect human trust. In Section III we describe Experiment 1 along with the development of a generalized model of human trust dynamics; we further examine the effects of national culture and gender on these dynamics. In Section IV we describe Experiment 2 in which we incorporate the effects of misses and false alarms into the model. Afterward, we discuss the implications of the estimated parameter values of the trust model in the context of HMI, followed by concluding statements.

II. Background

Comprehensive reviews of the influence of trust in HMI and HCI contexts are provided by Lee and See [5], and Hoff and Bashir [18]. Hoff and Bashir [18] classified trust into three categories: 1) dispositional; 2) situational; and 3) learned [18]. *Dispositional trust* is based on characteristics of the human. Factors that influence dispositional trust do not vary with time, but they still impact human decision-making during interactions with the autonomous system. Studies have shown differences in trust behavior between people of different cultures, age groups, and personality types [19]–[21]. *Situational trust* consists of factors that are external to the operator (e.g., task difficulty, potential risks) and those that are internal to the operator (e.g., self-confidence, domain knowledge) [18]. Finally, *learned trust* is based upon an operator's overall experience with an autonomous system and influences their initial mindset. During a new interaction with an autonomous system, humans' experience affects their established trust level. In this paper, we present a dynamic model that can capture variations in human trust level with respect to automation reliability for various demographic groups, thus, capturing both learned and dispositional trust characteristics.

In the remainder of this section, we first introduce studies that have modeled human trust in various contexts. Second, we review literature on determining the effect of automation reliability and system error types (i.e., misses or false alarms) on human trust. These factors influence learned trust. Finally, we review studies on examining dispositional trust factors, specifically gender and national culture.

A. Studies on Human Trust Modeling

Researchers have developed models for predicting human trust based on past experiences, which strongly influence learned trust. Jonker and Treur [6] suggested two types of functions to model trust dynamics: 1) trust update functions; and 2) trust evolution functions. Trust update functions use the current trust level and current experience to update the future trust level, while trust evolution functions map a sequence of trust related events (experiences) to a current trust level. In order to verify the proposed trust dynamics, Jonker *et al.* [22] conducted follow-up human subject experiments, which presented participants with a sequence of short stories for two scenarios: 1) a photocopier; and 2) a travel agency. Each scenario consisted of five positive and five negative stories, and participants reported their trust level after reading each story. The results suggested that trust dynamics are dependent on positive and negative experiences. However, limited by the number of trials (10 trials in each scenario), these studies only induced a single transition in trust level; therefore, the model did not capture the variations in trust dynamics involving multiple transitions.

Some studies have modeled human trust in the context of HMI. Lee and Moray [7] studied changes in human trust level using a simulated semiautomatic juice plant environment. It was observed that the human trust level was affected by the performance of the system, past trust levels, and faults. The authors used an auto-regressive-moving-average-vector analysis to model the input–output relationship of the trust behavior. They later showed that humans use automation when their trust in the automation exceeds their self-confidence [23]. These early efforts demonstrated the effect of situational and learned trusts on the interactions between humans and autonomous systems. However, due to a small sample size (i.e., four to five participants in each group) and a large standard deviation of the data, the accuracy of their model was limited.

Within the simulation context proposed by Moray and his colleagues, Lewandowsky *et al.* [24] compared trust in automation with trust in human partners in equivalent situations. Similar to the findings of Lee and Moray [23], Lewandowsky *et al.* [24] identified that faults in auxiliary control actions have a strong effect on trust and self-confidence of the human operator, and the difference between trust and self-confidence is a strong predictor of the human operator's reliance on automation as well as his/her reliance on human colleagues.

Factors that are significant in predicting trust level may also be dependent on the application context. Sadrfaridpour *et al.* [25] proposed a time-series model for the dynamics of human–robot trust in assembly lines based on the robot performance, human performance, and fault occurrences. More specifically, the performances were quantified by the robot's working speed and the human's state of muscle fatigue and recovery. How well the robot met the human's pace influenced the workload

and trust perceived by the human. The researchers' experimental results also suggested that the current trust is mainly dependent on the previous trust if there is no dramatic change in performance.

More recently, elements that are not based on rationale have been incorporated into a human trust model. Li *et al.* [26] used the structural equation modeling technique to identify the significance of human attitudes and subjective norm on "trusting intentions". Hoogendoorn *et al.* [10] developed models with biased experience and/or trust to account for this human behavior. They validated their models using a geographical area classification task and showed that a model with a bias term is capable of estimating trust more accurately than models without an explicit bias. However, their model was nonlinear in trust and experience, rendering it more difficult for analysis than linear models.

B. Effect of Misses and False Alarms in Automation

Automation reliability significantly influences human trust in autonomous systems and, in turn, influences human use of these systems [7], [27]. According to signal detection theory, automation errors can be classified as misses or false alarms; failing to detect the presence of a signal constitutes a miss, and incorrectly alerting humans to an absent stimulus constitutes a false alarm [28]. Existing literature shows that these two types of errors have different effects on human trust in automation. Specifically, these error types affect *reliance* and *compliance* to a different degree. On the one hand, reliance is when humans, in the absence of any signal from the system, continue to trust the system and refrain from a response. On the other hand, compliance is exhibited by a human trusting and responding to a signal when the system presents one [29]. An increase in the miss rate reduces reliance, while an increase in false alarms reduces compliance [3], [30]. This distinction is important as it leads the human to react to a signal. For example, a compliance-oriented system (i.e., gives warning when there is a malfunction) increases awareness in humans especially when warnings are spaced close to other indicators [31].

Humans may choose to ignore warnings if they experience high rates of false alarms, which is known as the "cry-wolf" effect [32]. This behavior represents humans' mistrust of autonomous systems and induces disuse of these systems [33]. Some studies suggest that false alarms cause greater negative effects on human trust in automation as they divert humans' attention, causing them to monitor unnecessary information [34]. Pervasive false alarms may make humans respond slower or less frequently to similar alerts in future [35], [36]. However, the high false-alarm rate does not appear to negatively impact trust in the context of en route air traffic controller conflict alerts [37]. Indeed, some studies showed contrary results where false alarm prone systems were more trustworthy than miss-prone systems [38], [39]. In addition, there are studies suggesting that false alarms and misses lead to similar effects on trust [40], [41] or that the effect is dependent on humans' cognitive capabilities [42].

Existing literature shows evident differences in opinions of the effects of misses and false alarms on human trust in automation. In order to resolve these differences, a model for human trust behavior with respect to false alarms and misses is needed. Moreover, the alarm threshold is determined based on the costs associated with each type of error, which means the optimal rate of misses/false alarms varies between systems. Therefore, a model of trust dynamics that connects human trust to autonomous system reliability can help us better understand how reliability-induced trust changes over time. Furthermore, it would allow us to understand how trust recovers with a hit (i.e., system correctly detects the presence of a signal) and/or a correct rejection (i.e., system not alerting the human to an absence of a signal).

C. Demographic Factors That Influence Trust

Autonomous system reliability and error type are external factors that influence learned trust. Apart from experiences accumulated from past interactions with autonomous systems, human trust behavior is also influenced by demographic factors including culture and gender. This is described as *dispositional trust* and is independent of a specific system or the context of an interaction [18].

Gender differences in trust behavior have been studied thoroughly in economic contexts [8], [43], [44]. Furthermore, some studies have shown gender differences in human–robot interaction contexts and technology adoption behavior [9], [45], [46]. For example, males were more likely to develop trust and positive attitudes toward female robots, while women showed little preference [47]. The attitudes of children toward humanoid robots are also influenced by gender. Tung [48] showed that girls favored humanlike, female robots more than boys did. In addition, females perceived highly automated driving systems as significantly less trustworthy than males did [49].

Values and social norms shared by members of a nation that guide people's behaviors and beliefs can be defined as the national culture for each country [50]. These factors also have an influence on the cognitive process of trust formation in humans. Therefore, people from different cultures are likely to use different mechanisms to form trust [51] and show particular trust behavioral intentions [52]. To date, only a few studies have examined the effect of national culture on trust in automation. Rice *et al.* [53] observed that Americans tended to trust less in automated systems as compared to Indians in the context of "auto-pilots." In another study, Americans were found to trust autonomous (decision-aid) systems less than Mexicans in a fraud investigation scenario [54]. Trust can also be seen as "the willingness to take risk" [55]. Considering the influence of national culture, uncertainty avoidance index (UAI) defined in Hofstede's six cultural dimensions [50], [56] is relevant to the construct of trust. Uncertainty avoidance tendency has been found to be significant in influencing trust in web design attributes [57], mobile commerce [58], information technology infrastructure [59], and in the context of simulated unmanned air vehicle control [60]. The higher the UAI number for a country, the less likely their people will tolerate uncertainty or risk.

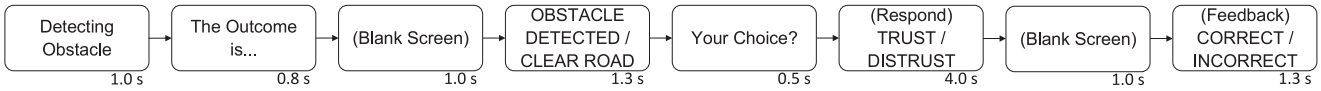


Fig. 1. Sequence of events in a single trial. The time length marked on the bottom right corner indicates the time interval that the information appeared on the screen.

In summary, published quantitative dynamic trust models do not explicitly consider a number of different factors, including the nature of human bias toward the system's response criteria (i.e., liberal and conservative), demographics, and false alarms and misses in autonomous systems. Moreover, although literature in the area of multiagent systems has analyzed the effects of past experiences on future trust level, this effect needs to be modeled for independent HMI. Therefore, the influence of these factors on human trust *dynamics* remains unexplored. To address these key gaps, we present two experiments that test the trust factors mentioned above and aid us in establishing a dynamic model of human trust.

III. Experiment 1

The first experiment was designed to understand human trust dynamics induced by the autonomous system performance, and to identify the effects of humans' national culture and gender on trust behaviors. The rates of misses and false alarms were controlled; so, participants encountered approximately equal numbers of these two error types. In addition, the order of these two error types was randomized within faulty trials. Therefore, the trust behavior induced by a specific error type was neutralized in Experiment 1.

A. Method

Stimuli and Procedures. The experiment was conducted online, and each participant accessed the study through a computer interface. Participants were told that the experiment was a simplified simulation of driving a car equipped with an obstacle detection sensor. The sensor was based on an image-recognition algorithm that would detect obstacles on the road in front of the car. During each trial, participants' task was to decide whether or not to trust the algorithm report, based on their previous experience with the algorithm. The instructions informed participants that the image-recognition algorithm used in the sensor was in beta testing.

An experiment session consisted of four initial practice trials followed by 100 trials comprising a sequence of events including stimulus, response, and feedback (see Fig. 1). There were two stimuli: 1) "obstacle detected"; and 2) "clear road," each having a 50% probability of occurrence. After receiving the stimulus, participants were asked to determine whether they "trusted" or "distrusted" the report provided by the algorithm. The system then gave feedback to the participants on the correctness of their responses (i.e., "correct" and "incorrect"). In order to examine how system reliability influences human trust level, the system was "reliable" in half of the trials and was "faulty" in the remaining half. Here, reliability is defined as the degree to which the algorithm report can be depended on to be accurate. In reliable trials, the algorithm correctly identified the road condition. This meant that "obstacle detected" was a hit, and "clear road" was a correct rejection. Accordingly, it would be marked as "correct" if the participant trusted the report, and "incorrect" if the participant distrusted the report. In faulty trials, there was a 50% probability that the algorithm *incorrectly* identified the road condition. A report of "obstacle detected" could be a false alarm, and "clear road" could be a miss (see Fig. 2). For the participant, this meant that responding "trust" to a false alarm or a miss would be marked as "incorrect." On the one hand, we implemented the 100% accuracy condition for reliable trials because it is the ideal performance a sensor can achieve. On the other hand, 50% accuracy for a binary decision would be a pure random chance. Therefore, it should result in the lowest possible trust level that a human has in the simulated sensor.

		Actual Scenario	
		Obstacle Present	Obstacle Absent
System Response	Obstacle Detected	Hit	False Alarm
	Clear Road	Miss	Correct Rejection

Fig. 2. Actual scenario and the system response form a 2 × 2 matrix. A system response of "clear road" in the presence of an obstacle constitutes a miss, and a system response of "obstacle detected" in the absence of an obstacle constitutes a false alarm.

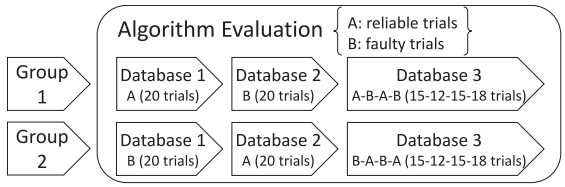


Fig. 3. Participants were randomly assigned to one of the two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty trials, were counterbalanced across groups.

All the trials in the study (100 in total) were divided into three phases, called "databases," as shown in Fig. 3. There was a 30-s break before the start of each database. Databases 1 and 2 were used to induce responses to constant system reliability—either reliable or faulty. Database 3 was used to excite all possible dynamics of the participants' trust responses by switching the accuracy of the algorithm between reliable and faulty according to a pseudorandom binary sequence. Stimuli in each trial were individually randomized for each participant and database. Participants were randomly assigned to one of two groups, which differed in the order of reliable and faulty trials to counterbalance possible ordering effects.

Participants: A total of 581 individuals (ages 20–73) were recruited using Amazon Mechanical Turk [61] to participate in

the study. Among the participants, 340 were males, 235 were females, and 6 did not provide gender information. These participants were randomly assigned to one of the two experimental groups. Participants in Groups 1 and 2 were initially faced with reliable trials and faulty trials, respectively. The participants were paid $0.50 each for their participation in the study. Before starting the study, participants electronically provided their consent. The Institutional Review Board at Purdue University approved the study. We collected participants' demographic information via a poststudy survey, which included questions about their gender along with the country in which they grew up. The latter is defined as national culture in this study.

Data Processing: To preprocess the collected data, we identified and removed outliers from the dataset. Each participant completed all 100 trials, but they were allowed to skip a trial if they could not make a decision within the given time frame (4 s). We considered excessive "no responses" (i.e., when participants skipped a trial) as well as excessive trust or distrust responses as outliers, determined by the interquartile range (IQR) rule (the 1.5× IQR rule) [62]. As a result, we removed 63 outliers from the dataset (out of 581 participants), which resulted in 518 valid participants.

To investigate the effects of national culture and gender on human trust, we categorized the collected data into four demographic bins; two were based on nationality: United States (U.S.) and India, and two were based on gender: male and female. Ideally, the selected sample would be representative of the population it came from. However, practically it was not possible to have an equal representation of each demographic group in the collected sample. In order to correct this anomaly in the selection probability of each demographic group in the population, the variables of each bin were adjusted using sampling weights such that each group had an equal representation in the sample population [63]. We calculated sampling weights for each demographic group in all of the bins as follows:

$$\text{Sampling weight} = \frac{\text{Population percentage}}{\text{Sample percentage}}. \tag{1}$$

Trust Model Description: For Groups 1 and 2, we computed the *probability of trust response* for each trial and across all subjects in each of the groups. This probability is defined as the percentage of people in the group who trust the algorithm report. At each trial, for calculating this probability, we assume that the response of each participant is like a Bernoulli trial with "trust" response as success and "distrust" response as failure. Given that for each trial, the responses of all participants are independent from one another, the random variable X, defined as the number of participants responding "trust" on a given trial, has a binomial distribution $B(k, p)$. The parameter k is the total number of participants in the bin and the parameter p, binomial proportion, can be estimated using a normal approximation as $\hat{p} = \frac{k_S}{k}$. Here k_S is the number of successes, i.e., number of trust responses in the given trial across participants. Therefore, at each trial, the probability of trust response can be estimated as $\hat{p}$ for that trial. The range of estimated probabilities was 0.5 to 1, where 0.5 represented low trust (i.e., the report was perceived as random by the participants; therefore, they responded randomly) and 1 represented high trust. These trust probabilities varied as the decision scenario changed with time and represent the trust level for the sample population. Henceforth, the trust probability will be labeled as *trust level* $T(n)$, where $n \in [1, 100]$ is the trial number. Similarly, we calculated the *probability of misses* $M(n)$ and the *probability of false alarms* $F(n)$ for each trial, across all subjects, in Groups 1 and 2. For Experiment 1, $M(n) = F(n)$ and varied from approximately 0 to 0.25, with 0.25 representing faulty trials leading to negative experience and 0 representing reliable trials leading to positive experience. Therefore, we define *experience* $E(n)$ as a function of $M(n)$ and $F(n)$ given by

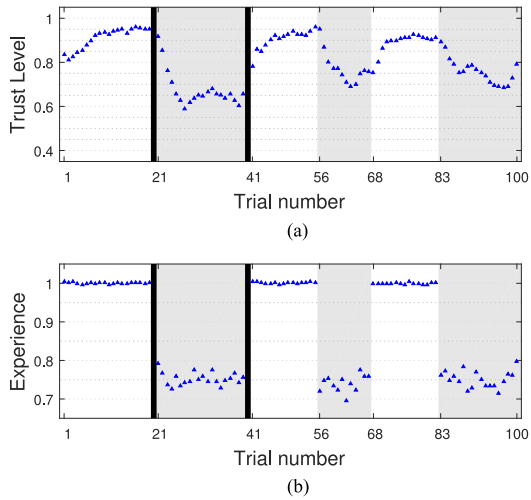


Fig. 4. Trust level (probability of trust response) and the experience for all participants. Faulty trials are highlighted in gray, and black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials. (a) Variation of trust level as a function of trial number. (b) Variation of experience as a function of trial number.

$$E(n) = 1 - [(1 - \beta)M(n) + \beta F(n)]. \tag{2}$$

Here, $\beta \in (0, 1)$ is the weighting factor for evaluating the relative effect of misses and false alarms on experience. Beta is the coefficient of the probability of false alarms in the model and, thus, can be called the *cry-wolf factor*. The higher the value of the cry-wolf factor β, the greater the effect of false alarms on the experience, and the lesser the effect of misses on the experience. Since the probability of misses and false alarms was equal for Experiment 1 [$M(n) = F(n) = K(n)$], (2) reduces to

$$E(n) = 1 - K(n) \tag{3}$$

where $K(n)$ is the probability of a miss or a false alarm. Thus, we obtain the dynamic variation of trust level $T(n)$ with experience $E(n)$ for all participants as shown in Fig. 4. In order to reduce noise from the dynamically varying signal $T(n)$, we used the Savitzky–Golay filter with order 3 and a window of size 5 [64].

Most of the existing human trust models showed trust to be directly related to experience. Jonker and Treur [6] presented change in trust to be directly proportional to the difference of experience and past trust. However, along with experience, we identified the significance of *cumulative perception of trust* and

the human's expectations of the autonomous system in formulating human trust behavior. Therefore, we adapt the model used by Jonker and Treur and introduced two additional terms—*cumulative trust* C_T and *expectation bias* B_X—to propose a second-order model as shown in (4). The states of the model are defined as trust level $T(n)$ and cumulative trust $C_T(n)$, and the input is defined as experience $E(n)$ along with a constant input disturbance called expectation bias B_X

$$T(n+1) - T(n) = \alpha_e[E(n) - T(n)] \tag{4a}$$
$$+ \alpha_c[C_T(n) - T(n)] \tag{4b}$$
$$+ \alpha_b[B_X - T(n)] \tag{4c}$$
$$C_T(n+1) = [1-\gamma]C_T(n) + \gamma T(n). \tag{4d}$$

In the model (4), change in trust $T(n+1) - T(n)$ depends linearly on three terms: 1) $E(n) - T(n)$ (4a); 2) $C_T(n) - T(n)$ (4b); and 3) $B_X - T(n)$ (4c), where each term is bounded between -1 and 1. We call the parameters α_e, α_c, and α_b the *experience rate factor*, *cumulative rate factor*, and *bias rate factor*, respectively, since they control the rate by which each individual difference affects the predicted trust level. Details on the estimation of these parameters are described in the *Parameter Estimation* section.

As shown in (4d), we define cumulative trust C_T as an exponentially weighted moving average of past trust level. Cumulative trust incorporates the learned trust in the model using a weighted history of past trust levels. A higher value of parameter γ discounts older trust levels faster, and thus γ can be called the *trust discounting factor*. The expectation bias B_X accounts for a human's expectation of a particular interaction with an autonomous system. This is modeled as an input disturbance, which remains constant during an interaction. The state $T(n)$ represents a probability, and the state $C_T(n)$ represents an exponentially weighted moving average of probability; therefore, both belong to $[0, 1]$. $B_X(n)$ must belong to $[0, 1]$ so that $T(n+1)$ remains bounded within $[0, 1]$ in (4a). Moreover, $E(n)$ belongs to $[0, 1]$, based on (2).

The linearity of the proposed model allows us to represent the model in state space form as

$$x(n+1) = \begin{bmatrix} 1-\alpha & \alpha_c \\ \gamma & 1-\gamma \end{bmatrix} x(n) + \begin{bmatrix} \alpha_e \\ 0 \end{bmatrix} u(n) + \begin{bmatrix} \alpha_b \\ 0 \end{bmatrix} d(n)$$
$$y(n) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(n) \tag{5}$$

where $x = \begin{bmatrix} T & C_T \end{bmatrix}^T$, $u = E$, $d = B_X$, and $\alpha = \alpha_e + \alpha_c + \alpha_b$. The linearity of the model also simplifies analysis of the trust dynamics as well as potential synthesis of model-based control algorithms for improved HMI.

Proposition 1: The linear state-space model given in (5) is stable if the parameters α_e, α_c, α_b, α, and γ belong to $(0, 1)$.

Proof: The eigenvalues of the proposed discrete model are given by

$$\lambda_{1,2} = 1 - \frac{\alpha+\gamma}{2} \pm \sqrt{\frac{\alpha-\gamma}{2} + \alpha_c\gamma} \tag{6}$$

and must lie inside the unit circle, i.e., $|\lambda_{1,2}| < 1$ to guarantee asymptotic stability. Therefore, it is sufficient to prove that

$$\lambda_1 = 1 - \frac{\alpha+\gamma}{2} - \sqrt{\left(\frac{\alpha-\gamma}{2}\right)^2 + \alpha_c\gamma} > -1 \tag{7a}$$

$$\lambda_2 = 1 - \frac{\alpha+\gamma}{2} + \sqrt{\left(\frac{\alpha-\gamma}{2}\right)^2 + \alpha_c\gamma} < 1 \quad . \tag{7b}$$

By rearranging and squaring both sides, (7a) and (7b) can be reduced to show that $\forall\ \gamma \in (0, 1)$, the following hold true:

$$2 > \alpha + \alpha_c \quad \text{and} \tag{8a}$$
$$0 < \alpha_e + \alpha_b. \tag{8b}$$

Equations (8a) and (8b) are satisfied if $\alpha_e, \alpha_c, \alpha_b \in (0, 1)$ and $\alpha \in (0, 1)$. Therefore, the trust model (5) is asymptotically stable. ■

Remark 1: The physical interpretation of these bounds on the parameters can be obtained by a closer examination of (4). The parameters α_e, α_c, and α_b are weighting factors for each of the terms and should be less than 1 so that the trust level remains stable. The variable γ is an exponential weighting factor that belongs to $(0, 1)$. Additionally, $\alpha = \alpha_e + \alpha_c + \alpha_b$ belongs to $(0, 1)$. This ensures that the net coefficient of the term $T(n)$ for calculating $T(n+1)$, i.e., $1 - \alpha$, belongs to $(0, 1)$ and is not negative. Consequently, a higher previous trust level will have a positive influence on the current trust level.

Proposition 2: The steady-state values of trust T_{ss} and cumulative trust C_{Tss} for a stable system given by (5) are a weighted average of steady-state experience E_{ss} and expectation bias B_X. The weights are proportional to α_e and α_b.

Proof: By substituting $x(n+1)$ and $x(n)$ with $x_{\text{ss}} = [T_{\text{ss}} \quad C_{T\text{ss}}]^\top$ and $u(n)$ with $u_{\text{ss}} = E_{\text{ss}}$, in (5), we can solve for the steady-state values T_{ss} and $C_{T\text{ss}}$ as follows:

$$T_{\text{ss}} = C_{T\text{ss}} = \frac{\alpha_e}{\alpha_e + \alpha_b} E_{\text{ss}} + \frac{\alpha_b}{\alpha_e + \alpha_b} B_X \quad . \tag{9}$$

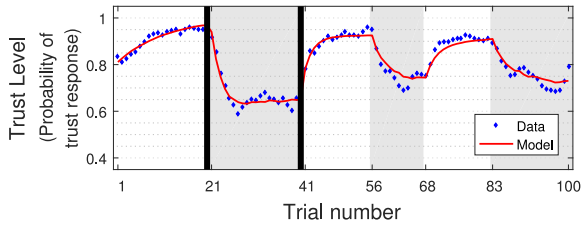
Here, the subscript $\bullet_{\text{ss}}$ represents the steady-state value of the variable. ■

Remark 2: Consider the case when $E_{\text{ss}} = 1$, which indicates that the system interacting with the human is consistently accurate. If the expectation bias is less than 1 ($B_X < 1$), the steady-state trust level T_{ss} of the human will be less than 1. Alternatively, consider the case when $E_{\text{ss}} = 0$, which indicates that the system interacting with the human is consistently faulty. If $B_X > 0$, the steady-state trust level T_{ss} will also be greater than 0. Therefore, the inclusion of human bias in the proposed model enables us to characterize this important effect on human trust level.

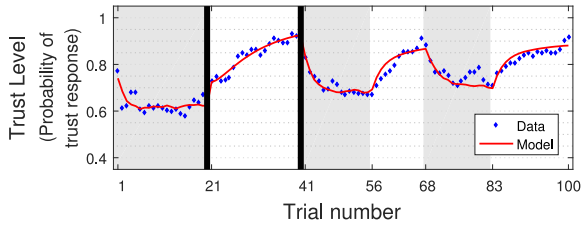
Parameter Estimation. For estimating the optimal set of model parameters, we used a nonlinear least squares estimation function *nlgreyest* from MATLAB 2016a. We identified the parameters using: 1) the data of all participants; and 2) the data in each of the four demographic bins. Each dataset consisted of data from each of the three "databases" in both Group 1 (in which participants were initially faced with reliable trials) and Group 2 (in which participants were initially faced with faulty trials). It is well known that the quality of any empirical

TABLE I
ESTIMATED MEAN PARAMETER VALUES WITH 95% CI FOR ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

Bin	Experience rate factor α_e	Cumulative rate factor α_c	Bias rate factor α_b	Trust discounting factor γ	Fit% Grp 1	Fit% Grp 2
All	0.2169 ± 0.0007	0.0755 ± 0.0005	0.0428 ± 0.0004	0.1148 ± 0.0012	95.7138 ± 0.0286	92.5567 ± 0.0555
US	0.2157 ± 0.0007	0.0635 ± 0.0007	0.0394 ± 0.0004	0.1270 ± 0.0029	94.5923 ± 0.0407	87.5857 ± 0.1061
India	0.2177 ± 0.0010	0.0996 ± 0.0011	0.0515 ± 0.0007	0.0942 ± 0.0008	91.9665 ± 0.0612	90.1372 ± 0.0793
Female	0.2277 ± 0.0009	0.0783 ± 0.0007	0.0373 ± 0.0005	0.1042 ± 0.0017	91.4777 ± 0.0589	89.1182 ± 0.0891
Male	0.2085 ± 0.0009	0.0817 ± 0.0007	0.0491 ± 0.0005	0.1375 ± 0.0018	93.9327 ± 0.0447	89.1748 ± 0.0804



(a)



(b)

Fig. 5. Participants' trust level (blue dots) and the prediction (red curve) based on past behavioral responses and the experience of all participants. Faulty trials are highlighted in gray, and black lines mark the breaks between databases. (a) Group 1: $R^2 = 95.74\%$. (b) Group 2: $R^2 = 92.53\%$.

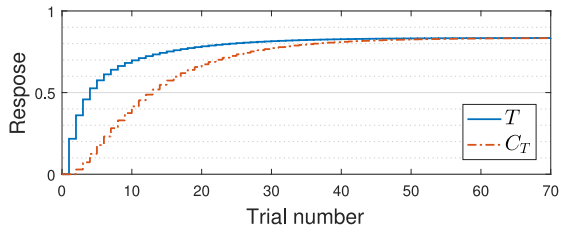


Fig. 6. Step response of the trust model with expectation bias $B_X = 0$ for all participants.

TABLE II
RISE TIMES (IN NUMBER OF TRIALS) FOR STEP RESPONSES CALCULATED USING THE ESTIMATED PARAMETER VALUES FOR ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

Bin	Rise Time (Number of Trials) T	C_T
All	15	27
US	13	24
India	20	34
Female	13	27
Male	15	23

parameter estimation is dependent on the data itself. A sample of human subject data cannot completely represent the human population, and the derived inferences may vary based on the selected sample. Therefore, in order to verify the robustness of the parameter estimation relative to the sample selection, we iterated the estimation 1000 times, with each iteration using a new, randomly selected subset of data representing 90% of the total datasets for all participants and each demographic bin. There was less than 2.5% error in the estimated parameter values caused by the variation in sample selection for a 95% confidence interval (CI) (see Table I), signifying a robust estimation.

B. Results

In order to verify whether our proposed model of trust level is valid, we estimated the model parameters for a general population, which included all 518 valid participants in our experiment. The fit between the trust model and the experimental data is shown in Fig. 5. Table I shows the optimal parameter values and the goodness of fit between the data and the model calculated using R-squared. The goodness of fit was 95.71% and 92.56% for all participants in Groups 1 and 2, respectively. Note that all the estimated parameter values satisfy the stability criteria defined in Proposition 1.

We observed that participants from different demographic groups required different amounts of time to adapt to changes in the system performance and attained different steady-state trust levels. In order to analyze these differences, we simulated the step response of each parameterized model. A sample step response for all participants with expectation bias $B_X = 0$ is shown in Fig. 6. The calculated rise time for the step response (see Table II) is an indicator of the rate of change of the trust dynamics. Rise time is defined as the time required for the response to increase from 10% to 90% of its final value. Therefore, a longer rise time implies slower trust dynamics.

Fig. 7 shows the experimentally obtained trust level and the predicted trust level of participants grouped by their national culture. Upon visual inspection, the U.S. participants trusted the system report less during the trials in Database 3, than during trials in Databases 1 and 2, in which the accuracy of the algorithm was switched between reliable and faulty; see Fig. 7(a) and (c). Moreover, in response to changes in system reliability, the trust level of U.S. participants changed at a faster rate, and approached an overall lower level than that of Indian participants. These observations are supported by the calculated rise time of the models (see Table II). The rise time of the state T for Indian participants is 53.8% higher than that of U.S. participants. This implies that Indian participants' trust level increased or decreased more slowly than that of U.S. participants after the system performance changes. Additionally, the rise time of the state C_T for U.S. participants is 29.4% shorter than that of Indian participants, which implies that their cumulative trust changed

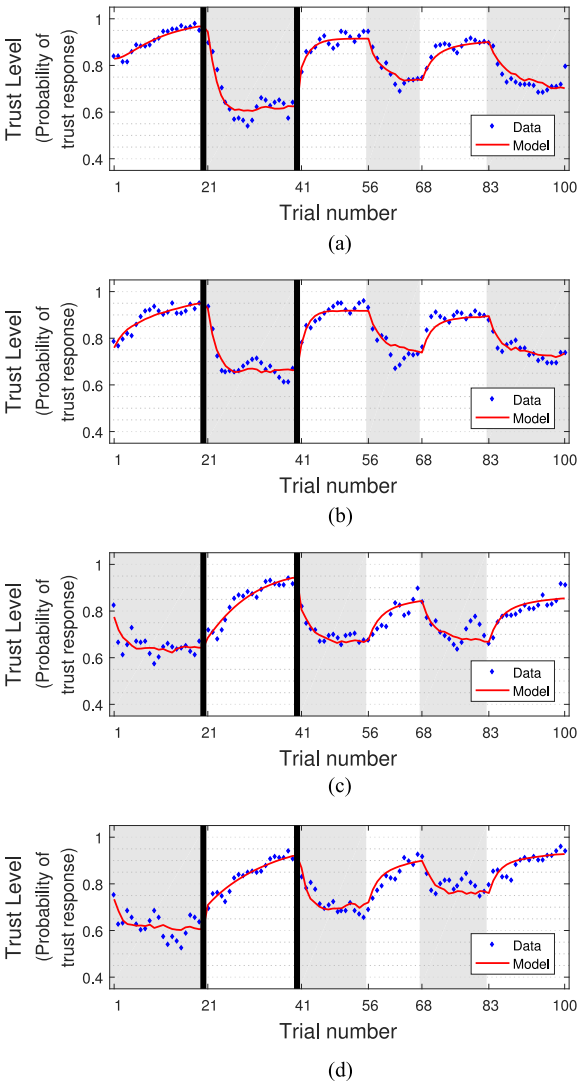


Fig. 7. Participants grouped by national culture. Blue dots are the reported trust level, while the red curve is the prediction from model. Faulty trials are highlighted in gray, and black lines mark the breaks between databases. (a) U.S. Group 1: $R^2 = 94.51\%$. (b) India Group 1: $R^2 = 92.00\%$. (c) U.S. Group 2: $R^2 = 87.56\%$. (d) India Group 2: $R^2 = 90.08\%$.

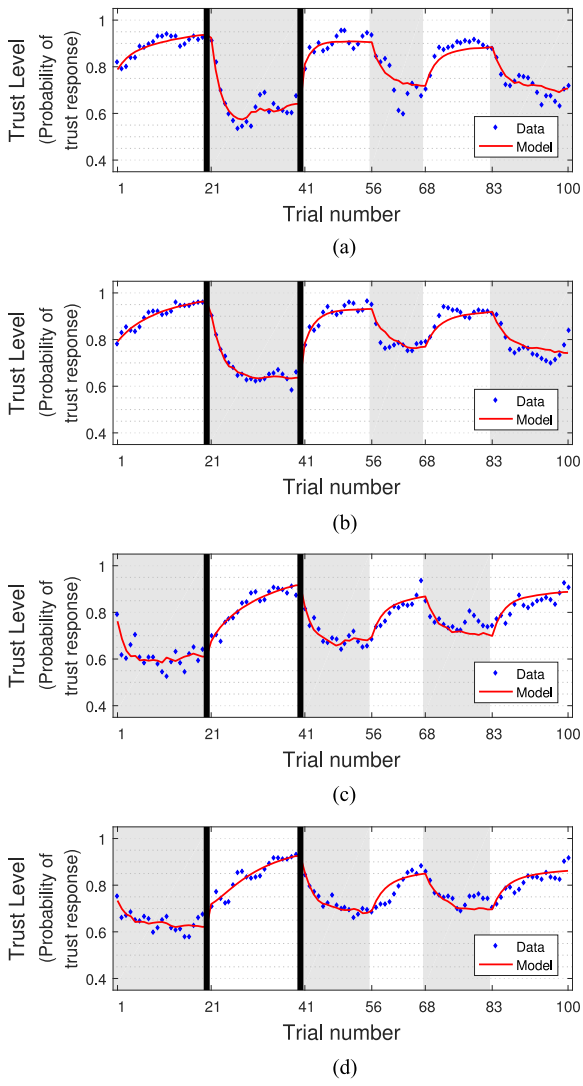


Fig. 8. Participants grouped by gender. Blue dots are the reported trust level, while the red curve is the prediction from model. Faulty trials are highlighted in gray, and black lines mark the breaks between databases. (a) Female Group 1: $R^2 = 91.57\%$. (b) Male Group 1: $R^2 = 93.98\%$. (c) Female Group 2: $R^2 = 88.94\%$. (c) Male Group 2: $R^2 = 89.22\%$.

relatively faster. This observation can also be attributed to the trust discounting factor γ, which is 34.8% larger for U.S. participants, indicating that U.S. participants relied on their recent trust level and experience more as compared to Indian participants.

Fig. 8 shows the experimentally obtained trust level and the prediction of participants grouped by their gender. The plots show that male participants exhibited greater trust in the system than female participants, especially when the system did not perform well [see Fig. 8(b) and (d)]. However, the trust level of female participants changed more rapidly than that of male participants. Similarly, when comparing the step responses, the rise time of state T for male participants is 15.4% longer than that of female participants, implying that the trust level of male participants changed more slowly than that of female participants. Furthermore, the rise time of the state C_T for male participants is 14.8% shorter than that of female participants, which implies that their cumulative trust changed relatively faster. This observation can also be attributed to the trust discounting factor γ, which is 32.0% larger for male participants, indicating that they relied on their recent trust level more as compared to female participants.

Based upon the high fit percentages achieved between the model and experimental data after parameter estimation, these results suggest that human trust in autonomous systems can be modeled as a function of their experience (which varies with system performance), cumulative trust, and expectation bias. Moreover, the estimated model parameters capture the effects of national culture and gender on trust behaviors.

IV. Experiment 2

As an extension of Experiment 1, we designed Experiment 2 to conduct an in-depth study on the effects of misses and false alarms on participants' trust levels. In this experiment,

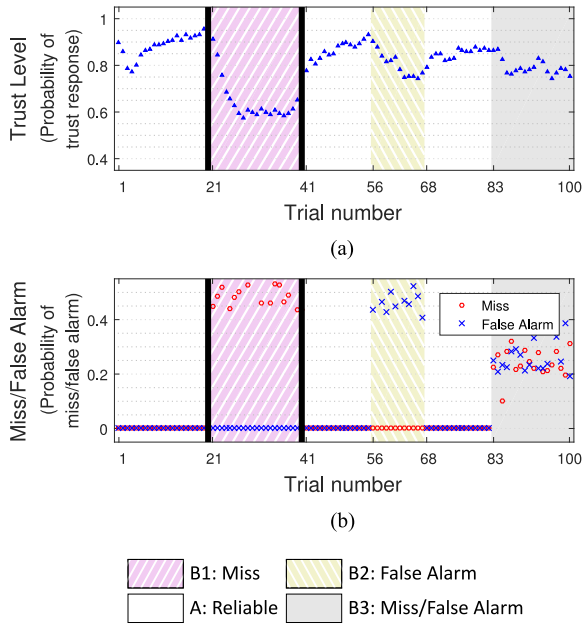


Fig. 9. Trust level (probability of trust response) for all participants and the probability of misses/false alarms that affect the experience. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials. (a) Variation of trust level as a function of trial number. (b) Variation of misses/false alarms as a function of trial number.

we present participants with trials containing 100% of misses and 100% of false alarms, unlike the 50–50 split used in Experiment 1 (see Fig. 9).

A. Method

We followed the same methodologies from Experiment 1 in terms of data collection, data processing, and modeling. We revised the stimuli to elicit trust reactions in response to misses and false alarms and analyzed the resulting data that were collected. We then expanded the general trust model to incorporate the effects of misses and false alarms.

Stimuli and Procedures. In comparison to Experiment 1, the only additional factor incorporated into Experiment 2 was the error type during faulty trials. More specifically, we manipulated the probability of misses and false alarms in faulty trials. In Experiment 1, a system error was equally probable to be a miss or a false alarm in each faulty trial. In Experiment 2, we examined the following three conditions:

1) an error was always a miss in a session of faulty trials;
2) an error was always a false alarm in a session of faulty trials;
3) an error was equally probable to be a miss or a false alarm in a session of faulty trials.

Fig. 10 shows the condition and trial orders in each database. Participants were randomly assigned to one of two groups in the interest of testing whether the experience of misses or false alarms affects the other.

Participants: A total of 333 individuals (ages 19–74) participated in Experiment 2. Among the participants, 171 were

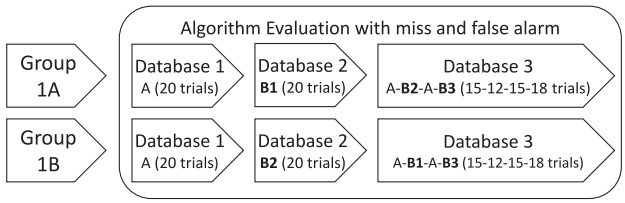


Fig. 10. Participants were randomly assigned to one of the two groups. The system reliability was varied between databases and groups. *A* consisted of reliable trials (miss = 0%, false alarm = 0%); *B*1 consisted of faulty trials with misses (miss = 50%, false alarm = 0%); *B*2 consisted of faulty trials with false alarms (miss = 0%, false alarm = 50%); *B*3 consisted of faulty trials with both misses and false alarms (miss = 25%, false alarm = 25%).

TABLE III
ESTIMATED MEAN PARAMETER VALUES WITH 95% CI FOR THE CRY-WOLF FACTOR β FOR ALL PARTICIPANTS AND EACH DEMOGRAPHIC BIN

Bin	Cry-wolf factor β	Fit% Grp 1	Fit% Grp 2
All	0.3956 ± 0.0012	91.7593 ± 0.0734	91.2061 ± 0.0652
US	0.3209 ± 0.0016	90.6699 ± 0.0953	87.4562 ± 0.0882
India	0.4758 ± 0.0019	82.4059 ± 0.1667	87.1428 ± 0.1110
Female	0.4276 ± 0.0015	83.8733 ± 0.1557	80.2373 ± 0.1644
Male	0.3623 ± 0.0018	89.8757 ± 0.1062	90.8326 ± 0.0800

males, 158 were females, and 4 did not provide gender information. These participants were randomly assigned to one of the two experimental groups. The recruitment procedure and the survey used to collect demographic information were the same as those in Experiment 1.

Data processing: We used the IQR rule as introduced in Experiment 1 to identify and remove outliers. The procedure resulted in 293 valid datasets (out of a total of 333 participants) to be analyzed.

Parameter Estimation: Using the data collected in Experiment 2, we estimated the *cry-wolf factor* β by setting all other factors (*experience rate factor* α_e, *cumulative rate factor* α_c, *bias rate factor* α_b, and *trust discounting factor* γ) to the values estimated in Experiment 1. The robustness of the estimated value of β was verified by 1000 iterative estimations, with each iteration using a new randomly selected subset of data representing 90% of the total dataset for all participants and each demographic bin. The errors caused by the variation in sample selection for a 95% CI were less than 2.5%. Table III shows the parameter values and the goodness of fit.

B. Results

We first investigated whether the system error type (i.e., miss and false alarm) affects the trust dynamics of the general population, which included all 293 valid participants in the experiment. Fig. 11 shows the experimental trust level compared against the model. Participants responded differently to misses and false alarms, and in some cases, the experience of one error type affected later responses to the other error type. The proposed trust model was able to predict the trust dynamics while taking into account the rate of misses and false alarms. The goodness of fit was measured using the R-squared value of the data; the results

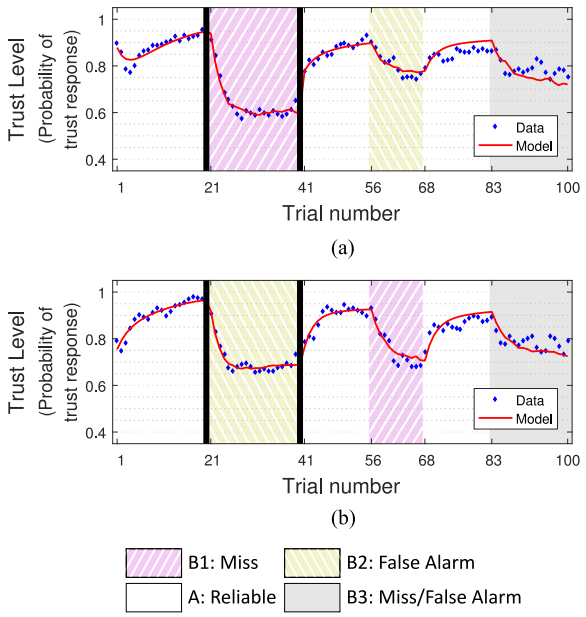


Fig. 11. Participants' trust level (blue dots) and the prediction (red curve) based on past behavioral responses and the experience of all participants. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases. (a) Group 1A: $R^2 = 91.83\%$. (b) Group 1B: $R^2 = 91.25\%$.

were 91.76% and 91.20% for all participants in Groups 1 and 2, respectively.

The results suggest an interaction effect between risk-taking behavior and demographic factors on trust. Fig. 12 shows the experimentally obtained trust level and model predictions for participants grouped by their national culture. U.S. participants trusted less than Indian participants when encountering system misses [see Fig. 12(a) and (b)]. Moreover, U.S. participants trusted less in miss-prone trials than false alarm prone trials regardless of whether they encountered false alarms first or not [see Fig. 12(a) and (c)]. By contrast, Indian participants trusted less in miss-prone trials than false alarm prone trials only when they encountered misses first [see Fig. 12(b)]; their trust level in miss-prone trials decreased less if they encountered system false alarms prior to misses [see Fig. 12(d)]. The cry-wolf factor β of the model is 48.3% larger for Indian participants than that for U.S. participants. The larger the value of β, the weaker the negative effect of misses on trust, indicating that misses have a stronger negative effect on trust for U.S. participants as compared with Indian participants.

We also observed gender differences in response to system misses and false alarms. Fig. 13 shows the experimental trust level and the prediction of participants grouped by their gender. Male participants trusted less in miss-prone trials than female participants if they had not encountered system false alarms first [compare Fig. 13(a) and (b)]. However, if participants encountered false alarms first, females reached a lower trust level than males [compare Fig. 13(c) and (d)]. In general, male participants were more sensitive to system misses. The cry-wolf factor β of the trust model supports this observation; β is 18.0% larger for female participants than for male participants, which

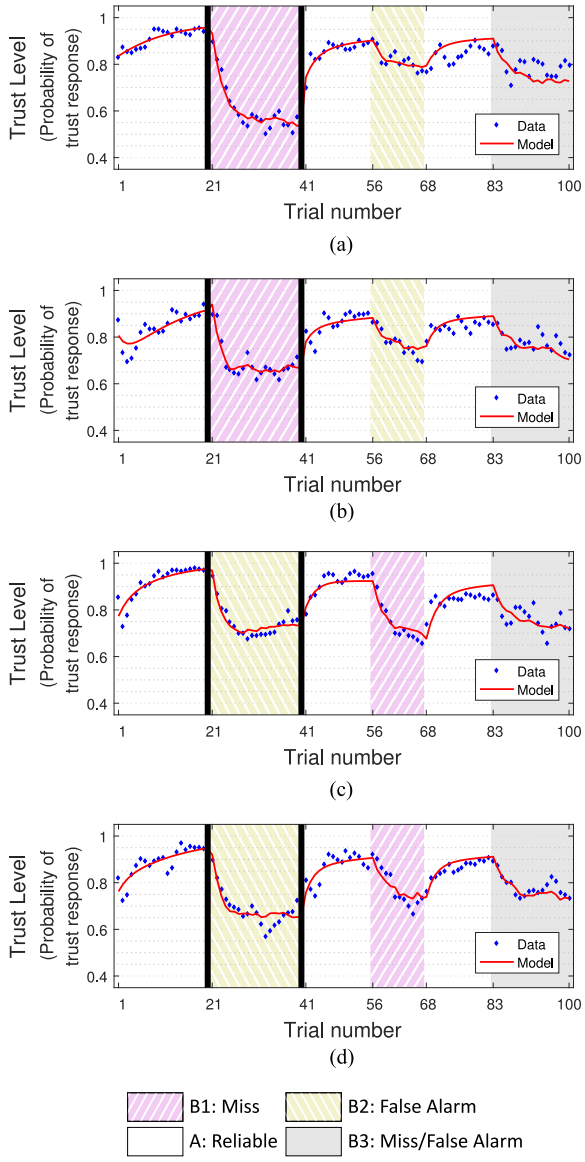


Fig. 12. Participants grouped by national culture. Blue dots are the reported trust level, while the red curve is the prediction from model. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases. (a) U.S. Group 1A: $R^2 = 90.67\%$. (b) India Group 1A: $R^2 = 82.41\%$. (c) U.S. Group 1B: $R^2 = 87.46\%$. (d) India Group 1B: $R^2 = 87.14\%$.

implies that misses have a stronger negative effect on trust for male participants as compared with female participants.

V. Discussion

In this section, we provide a more in-depth discussion of the main results of the two experiments. The two experiments presented in this paper elicited the *variation* of a human's trust response to system reliability. Participants attained a high trust level in reliable trials and a low trust level in faulty trials; this was achieved without training participants or providing them with specific information (e.g., a game rule or background stories). The trust dynamics were modeled based on *past behavioral*

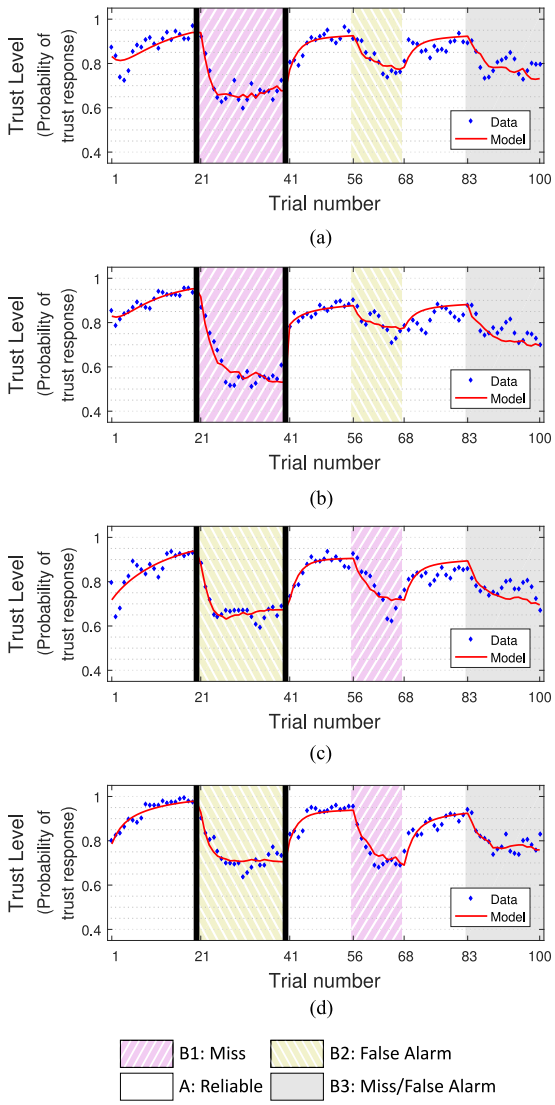


Fig. 13. Participants grouped by gender. Blue dots are the reported trust level while the red curve is the prediction from model. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases. (a) Female Group 1A: $R^2 = 93.87\%$. (b) Male Group 1A: $R^2 = 89.88\%$. (c) Female Group 1B: $R^2 = 80.24\%$. (d) Male Group 1B: $R^2 = 90.83\%$.

responses of the human, *human trust bias*, and the system reliability. The system reliability was further described by *the rate of misses and false alarms*. The model was verified using the collected human subject data that accounted for ordering effects with respect to system reliability. In other words, the prediction capability of the model was consistent for both Groups 1 and 2. Thus, the model can describe human trust irrespective of the initial condition of the system reliability. Moreover, the interaction between the human and machine was the most significant factor in temporal variations in trust level. Therefore, the developed study is effective for modeling dynamic human trust behavior in HMI contexts.

The proposed study design induced trust dynamics by manipulating multiple transitions between positive and negative experiences. We observed that it took approximately 8 to 10 trials for the participants to establish a new trust level. Moreover, in some cases [e.g., Fig. 13(a)] the trust response still increased or decreased near this newly attained level in both reliable and faulty trials. This finding was contrary to Jonker *et al.* [22] who asserted that "after a negative experience an agent will trust less or the same amount, but never more." Jonker's study was composed of only two sets of five trials, each with one transition in between. However, we found this to be less than the required number of trials to reach a new trust level.

The aggregated trust response and the trust model enhanced our understanding of dispositional trust and learned trust in autonomous systems. Participants from the U.S. exhibited a lower trust level than Indian participants. This is consistent with the findings from Rice *et al.* [53] and Huerta *et al.* [54] that Americans trust autonomous systems less than Indians and Mexicans, respectively. Moreover, system misses induced stronger distrust in U.S. participants than in Indian participants, suggesting that U.S. participants are less willing to take risks. This agrees with the smaller UAI of Indian culture as compared to that of U.S. culture (40 versus 46) [17], where the literature demonstrated that humans from higher uncertainty avoidance cultures are less likely to trust or implement new technology [58], [59].

Regarding gender, male participants appeared to trust more than female participants, especially when the system was not reliable. This is supported by Feldhütter *et al.* [49], but is contrary to the findings of Haselhuhn *et al.* [65], which showed that women's trust decreases less than men after transgressions as they prefer to maintain interpersonal relationships. These results highlight that the dynamics of human trust behavior in HMI contexts is different from interpersonal trust behavior between humans, thus, creating a need for human trust models in HMI contexts. Additionally, the variation in trust responses of female participants was noticeably higher than that in male participants. This variation indicates that the female participants have diverse perceptions of autonomous systems and, therefore, other factors such as personality and expertise should be investigated in future studies. Finally, there were gender differences in the responses to misses and false alarms as discussed in Section IV-B. Along with the observations of U.S. and Indian participants, demographic effects can partially explain the inconsistency between previously published results on the effects of system error type on human trust.

VI. Conclusions

We developed a study composed of two experiments to elicit a dynamic change in human trust with respect to HMI contexts. Furthermore, we established a quantitative trust model, motivated by literature on computational models and parameterized using human subject data. This model was verified using data collected from more than 800 participants and has a prediction accuracy higher than 92% for the general population. We introduced the effect of cumulative trust, expectation bias, and misses/false alarms, to accurately capture human trust dynamics during HMI.

It has been established that trust plays an important role in human–system interactions. Therefore, to establish collaborative interactions between humans and autonomous systems, it is essential to adapt the human–system interaction based on the human's trust level. This, in turn, requires autonomous

systems to utilize quantitative models of dynamic human trust behavior. Existing human trust models are typically nonlinear or predict trust solely based on experiences. Moreover, others were developed using experimental data in which the stimulus–a transition in system performance–occurred only once in each experiment. Therefore, their ability to predict trust variations is limited. We addressed this gap by identifying the significance of cumulative trust and expectation bias through experiments that elicited multiple dynamic transitions in human trust, and then incorporated these two variables in the proposed linear model.

In addition to proposing a general trust model structure, we characterized the effects of both dispositional and learned trust factors, specifically national culture, gender and system error type, using estimated model parameters. We also characterized the effects of misses and false alarms on the dynamics of human trust behavior and compared differences between demographics. We found that system misses induce a stronger distrust in U.S. participants than that in Indian participants and have a stronger negative effect on trust for male participants than that for female participants. While the proposed model is representative of a population of individuals rather than trained to a specific human, such a model could be used to design machines that are required to interact with unspecified users grouped by demographics.

One limitation of this study is that a computer-based interface system was used in the experiment, and therefore, the ecological validity could be improved. Future work will involve conducting experiments in real-life settings. The model could also be generalized for use in a wider range of domains by expanding the definition of experience to incorporate other significant factors beyond the probability of misses and false alarms, such as system transparency and the level of automation. Other extensions of the work will include consideration of additional demographics and validation of the model in other HMI contexts.

REFERENCES

[1] J. Y. C. Chen and M. J. Barnes, "Human-agent teaming for multirobot control: A review of human factors issues," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 1, pp. 13–29, Feb. 2014.

[2] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, no. 5–6, pp. 527–539, 1987.

[3] B. T. B. Sheridan, R. Parasuraman, T. B. Sheridan, and R. Parasuraman, "Human-automation interaction," *Rev. Human Factors Ergonom.*, vol. 1, no. 1, pp. 89–129, 2005.

[4] M. Richtel and C. Dougherty, "Google's driverless cars run into problem: Cars with drivers," *New York Times*, Sep. 1, 2015. [Online]. Available: http://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html. Accessed on: Jan. 5, 2017.

[5] J. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *Proc. Euro. Workshop Model. Auton. Agents a Multi-Agent World*, 1999, pp. 221–231.

[7] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[8] R. Croson and N. Buchan, "Gender and culture: International experimental evidence from trust games," *Amer. Econ. Rev.*, vol. 89, no. 2, pp. 386–391, May 1999.

[9] T. Nomura and S. Takagi, "Exploring effects of educational backgrounds and gender in human-robot interaction," in *Proc. Int. Conf. User Sci. Eng.*, Nov. 2011, pp. 24–29.

[10] M. Hoogendoorn, S. W. Jaffry, P. P. Van Maanen, and J. Treur, "Modelling biased human trust dynamics," *Web Intell. Agent Syst.*, vol. 11, no. 1, pp. 21–40, Aug. 2013.

[11] M. Hoogendoorn, S. W. Jaffry, P. P. Van Maanen, and J. Treur, "Design and validation of a relative trust model," *Knowl.-Based Syst.*, vol. 57, pp. 81–94, Feb. 2014.

[12] J. Sabater and C. Sierra, "Review on computational trust and reputation models," *Artif. intell. Rev.*, vol. 24, no. 1, pp. 33–60, Sep. 2005.

[13] C. Carver and M. Scheier, *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*, (Springer Series in Social Psychology). New York, NY, USA: Springer-Verlag, 2012.

[14] G. Klein, "Naturalistic decision making," *Human Factors*, vol. 50, no. 3, pp. 456–460, Jun. 2008.

[15] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *Proc. Amer. Control Conf.*, Seattle, WA, USA, May 2017, pp. 1542–1548.

[16] M. Hoogendoorn, S. W. Jaffry, P.-P. van Maanen, and J. Treur, "Modeling and validation of biased human trust," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2011, pp. 256–263.

[17] G. H. Hofstede and G. Hofstede, *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2nd ed. Newbury Park, CA, USA: Sage, 2001.

[18] K. A. Hoff and M. Bashir, "Trust in automation integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, May 2015.

[19] J. A. Colquitt, B. A. Scott, and J. A. LePine, "Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance," *J Appl. Psychol.*, vol. 92, no. 4, pp. 909–927, 2007.

[20] G. Ho, L. M. Kiff, T. Plocher, and K. Z. Haigh, "A model of trust & reliance of automation technology for older users," in *Proc. AAAI-2005 Fall Symp., Caring Mach., AI Eldercare*, 2005, pp. 45–50.

[21] M. Naef, E. Fehr, U. Fischbacher, J. Schupp, and G. Wagner, "Decomposing trust: Explaining national trust differences," *Inst. J. Psychology*, vol. 43, no. 3, p. 577, 2008.

[22] C. M. Jonker, J. J. P. Schalken, J. Theeuwes, and J. Treur, "Human experiments in trust dynamics," in *Proc. Trust Manage., 2nd Int. Conf., iTrust*, 2004, pp. 206–220.

[23] J. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, Jan. 1994.

[24] S. Lewandowsky, M. Mundy, and G. P. A. Tan, "The dynamics of trust: Comparing humans to automation." *J. Exp. Psychol., Appl.*, vol. 6, no. 2, pp. 104–123, 2000.

[25] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and control of trust in human-robot collaborative manufacturing," in *Proc. Robust Intell. Trust Auton. Syst*. Boston, MA, USA, 2016, pp. 115–141.

[26] X. Li, T. J. Hess, and J. S. Valacich, "Using attitude and social influence to develop an extended trust model for information systems," *ACM SIGMIS Database, Database Adv. Inf. Syst.*, vol. 37, no. 2–3, pp. 108–124, Sep. 2006.

[27] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, Jun. 2003.

[28] H. Poor, *An Introduction to Signal Detection and Estimation* (Springer Texts in Electrical Engineering). New York, NY, USA: Springer-Verlag, 1998.

[29] E. T. Chancey, J. P. Bliss, Y. Yamani, and H. A. H. Handley, "Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence," *Human Factors*, vol. 59, no. 3, pp. 333–345, May 2017.

[30] S. R. Dixon and C. D. Wickens, "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload." *Human Factors*, vol. 48, no. 3, pp. 474–486, Sep. 2006.

[31] J. Meyer, "Effects of warning validity and proximity on responses to warnings." *Human Factors*, vol. 43, no. 4, pp. 563–572, Dec. 2001.

[32] S. Breznitz, *Cry Wolf: The Psychology of False Alarms*, 1st ed. London, U. K.: Psychology Press, May 2013.

[33] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997.

[34] S. R. Dixon, C. D. Wickens, and J. S. McCarley, "On the independence of compliance and reliance: Are automation false alarms worse than misses?" *Human Factors*, vol. 49, no. 4, pp. 564–72, Aug. 2007.

[35] J. P. Bliss and G. Capobianco, "Collective mistrust of alarms," in *Proc. Int. Symp. Aviation Psychol., Apr*. 2003, pp. 14–18.

[36] J. D. Johnson, "Type of automation failure: The effects on trust and reliance in automation," M.S. thesis, Sch. Psychol., Georgia Inst. Technol., 2004.
[37] C. D. Wickens, S. Rice, D. Keller, S. Hutchins, J. Hughes, and K. Clayton, "False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect?" *Human Factors*, vol. 51, no. 4, pp. 446–462, Aug. 2009.
[38] N. S. Stanton, S. A. Ragsdale, and E. A. Bustamante, "The effects of system technology and probability type on trust, compliance, and reliance," in *Proc. Human Factors Ergonom. Soc. 53rd Ann. Meet.*, Jun. 2009, pp. 1368–1372.
[39] R. B. Davenport and E. A. Bustamante, "Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task," in *Proc. Human Factors Ergonom. Soc. 54th Ann. Meet.*, 2010, pp. 1513–1517.
[40] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, "Automation failures on tasks easily performed by operators undermine trust in automated aids," *Human Factors*, vol. 48, no. 2, pp. 241–256, 2006.
[41] J. Sauer, A. Chavaillaz, and D. Wastell, "Experience of automation failures in training : Effects on trust, automation bias, complacency and performance," *Ergonomics*, vol. 59, no. 6, pp. 767–780, Jun. 2016.
[42] J. Y. C. Chen and P. I. Terrence, "Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment." *Ergonomics*, vol. 52, no. 8, pp. 907–920, Aug. 2009.
[43] A. Chaudhuri and L. Gangadharan, "Gender differences in trust and reciprocity," Department of Economics–Working Papers Series, The University of Melbourne, no. 875, 2003. [Online]. Available: https://EconPapers.repec.org/RePEc:mlb:wpaper:875
[44] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games Econ. Behav.*, vol. 10, no. 1, pp. 122–142, Jul. 1995.
[45] V. Venkatesh, M. G. Morris, and P. L. Ackerman, "A longitudinal field investigation of gender differences in individual technology adoption decision-making processes," *Organizational Behav. Human Decis. Process.*, vol. 83, no. 1, pp. 33–60, Sep. 2000.
[46] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Prediction of human behavior in human - robot interaction using psychological scales for anxiety and negative attitudes toward robots," *IEEE Trans. Robot.*, vol. 24, no. 2, pp. 442–451, Apr. 2008.
[47] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive robotics: The influence of robot gender on human behavior," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 2563–2568.
[48] F.-W. Tung, "Influence of gender and age on the attitudes of children towards humanoid robots," *Human-Comput. Interact.*, vol. IV, pp. 637–646, 2011.
[49] A. Feldhütter, C. Gold, A. Hüger, and K. Bengler, "Trust in automation as a matter of media and experience of automated vehicles," in *Proc. Human Factors Ergonom. Soc. 2016 Ann. Meet.*, 2016, pp. 2024–2028.
[50] G. Hofstede, *Culture's Consequences: International Differences in Work Related Values*. Newbury Park, CA, USA: Sage, 1980.
[51] P. M. Doney, J. P. Cannon, and M. R. Mullen, "Understanding the influence of national culture on the development of trust," *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 601–620, Jul. 1998.
[52] D. Gefen and T. H. Heart, "On the need to include national culture as a central issue in e-commerce trust beliefs," *J. Global Inf. Manage.*, vol. 14, no. 4, pp. 1–30, Oct. 2006.
[53] S. Rice *et al.*, "Passengers from India and the United States have differential opinions about autonomous auto-pilots for commercial flights," *Int. J. Aviation, Aeronaut., Aerosp.*, vol. 1, no. 1, p. 3, 2014.
[54] E. Huerta, T. Glandon, and Y. Petrides, "Framing, decision-aid systems, and culture: Exploring influences on fraud investigations," *Int. J. Accounting Inf. Syst.*, vol. 13, no. 4, pp. 316–333, Dec. 2012.
[55] F. D. Schoorman, R. C. Mayer, and J. H. Davis, "An integrative model of organizational trust: Past, present, and future," *Acad. Manage.*, vol. 32, no. 2, pp. 344–354, 2007.
[56] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and Organizations: Software of the Mind*, 3rd ed. New York, NY, USA: McGraw-Hill, 2010.
[57] C. M. N. Faisal, M. Gonzalez-Rodriguez, D. Fernandez-Lanvin, and J. de Andres-Suarez, "Web design attributes in building user trust, satisfaction, and loyalty for a high uncertainty avoidance culture," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 847–859, Dec. 2017.
[58] A. Vance, C. Elie-Dit-Cosaque, and D. W. Straub, "Examining trust in information technology artifacts: The effects of system quality and culture," *J. Manage Inf. Syst.*, vol. 24, no. 4, pp. 73–100, Apr. 2008.
[59] I. P. L. Png, B. C. Y. Tan, and K. L. Wee, "Dimensions of national culture and corporate adoption of IT infrastructure," *IEEE Trans. Eng. Manage.*, vol. 48, no. 1, pp. 36–45, Feb. 2001.
[60] S.-Y. Chien, M. Lewis, K. Sycara, Jyi-Shane Liu, and A. Kumru, "Influence of cultural factors in dynamic trust in automation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Budapest, Hungary, Oct. 2016, pp. 002 884–002 889.
[61] Amazon, "Amazon mechanical turk," *Amazon Mech. Turk - Welcome*, 2005. [Online]. Available: https://www.mturk.com/. Accessed on Feb. 20, 2016.
[62] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdiscip. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 73–79, Jan. 2011.
[63] D. Pfeffermann, "The role of sampling weights when modeling survey data," *Int. Statist. Rev.*, vol. 61, no. 2, pp. 317–337, Aug. 1993.
[64] S. J. Orfanidis, *Introduction to Signal Processing* (Prentice Hall International Editions). Upper Saddle River, NJ, USA: Prentice-Hall, 1995.
[65] M. P. Haselhuhn, J. A. Kennedy, L. J. Kray, A. B. Van Zant, and M. E. Schweitzer, "Gender differences in trust dynamics: Women trust more than men following a trust violation," *J. Exp. Soc. Psychol.*, vol. 56, pp. 104–109, 2015.

Wan-Lin Hu received the B.S. degree in bio-industrial mechatronics engineering and the M.S. degree in electrical engineering from National Taiwan University, Taipei City, Taiwan, in 2006 and 2008, respectively. She received the Ph.D. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2017.

She is a Postdoctoral Scholar with the SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA. Her research interests include human–machine interaction, psychophysiology, decision-making, and design methodology.

Kumar Akash received the B. Tech. degree in mechanical engineering (ME) from the Indian Institute of Technology Delhi, New Delhi, India, in 2015, and the M.S. degree in ME from Purdue University, West Lafayette, IN, USA, in 2018. He is working toward the Ph.D. degree at the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA.

His research interests focus on dynamic modeling and control of human behavior in human–machine interactions, brain–computer interfaces, and machine learning.

Tahira Reid received the B.S. and M.S. degrees in mechanical engineering (ME) from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2000 and 2004, respectively. She received the Ph.D. degree in design science from the University of Michigan, Ann Arbor, MI, USA, where ME and psychology were her focus areas.

She is an Associate Professor with the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA. Her research interests focus on the quantification and integration of human-centered considerations in engineering, the design process, and human–machine systems.

Neera Jain received the S.B. degree in mechanical engineering (ME) from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, and the M.S. and Ph.D. degrees, also in ME, from the University of Illinois at Urbana-Champaign in 2009 and 2013, respectively.

She is an Assistant Professor with the School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA. Her research interests include dynamic modeling and control theory with applications to human–machine interactions, thermodynamic systems, and advanced manufacturing.

Drivers' Attentional Instability on a Winding Roadway

Richard J. Jagacinski, Emanuele Rizzi, Benjamin J. Bloom, O. Anil Turkkan, Tyler N. Morrison, Haijun Su, and Junmin Wang, *Senior Member, IEEE*

***Abstract*—The spatiotemporal distribution of drivers' attention to preview was inferred from their steering movements while tracking a winding roadway in a laboratory setting. For most subjects, the average driving attentional distribution over six daily sessions was relatively stable and generalized across different control devices. However, there was considerable day-to-day variability in the attentional distributions. This variability was modeled as a strong interaction between two dynamic processes, the attentional emphasis of selected regions and inhibition of surrounding regions. The model combines a novel application of a reaction-diffusion model of biological pattern formation with an optimal control model of attention to preview. The combined model treats attentional dynamics as an example of the biological spacing of a limited cognitive resource, which is also shaped by the demands of action.**

***Index Terms*—Attention, driving, manual control, optimal control, pattern formation, preview, reaction-diffusion, tracking.**

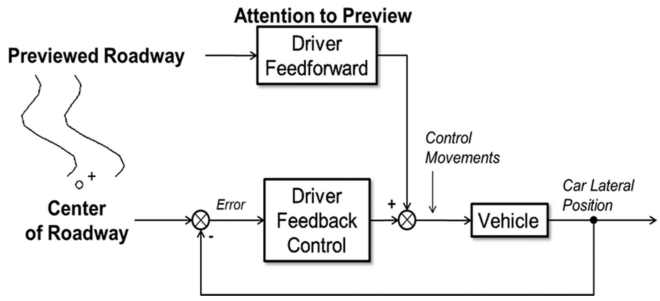


Fig. 1. Roadway display (upper left) and a conceptual model of movement control with preview. Preview extends 1 s into the future at the top of the display. The cross represents the present center of the roadway. Subjects were instructed to keep the circular cursor directly below the cross.

I. Introduction

In TRACKING, a winding roadway performance is improved by the availability of a preview of the upcoming roadway [1] [2] (see Fig. 1). Jagacinski *et al*. [3] have recently introduced a methodology for measuring the spatiotemporal distribution of attention to preview by analyzing human movements used to track a winding roadway. Building on previous work by Johnson and Phatak [4], Sheridan [5], Levison [6], and others, the previewed roadway was perturbed with a sinusoid of a different frequency at each of ten preview times ranging from 0.1, 0.2, up to 1.0 s. Fourier analysis of the person's tracking movements revealed an amplitude at each of these frequencies. The ratio of this amplitude to the amplitude occurring at that same frequency in a control condition without perturbations provided a signal-to-noise ratio that is interpreted as a measure of attention allocation. It is assumed that the driver has a limited cognitive capacity to process preview and selectively emphasizes or attends to certain preview times depending on the dynamics of the control task. The present methodology measures to what degree various preview times are coupled to the driver's control motions, and does not rely on eye movements to infer attention. The present experimental task used a display that did not require shifting one's gaze. This measurement technique reveals a detailed spatiotemporal distribution of attention, which may be helpful in assessing individual differences in cognitive aspects of driving skill.

The data of Jagacinski *et al.* [3] were obtained from naive subjects over 2 or 3 sessions of approximately 1 h each. The present paper examines this measure of attention with a convenience sample of four subjects who were researchers familiar with the measurement technique, and who tracked for an extended period of 18 sessions in two different laboratory settings. Questions of interest are as follows.

1) Do subjects reach stable asymptotic performance in their distribution of attention to preview?
2) Do subjects exhibit similar patterns of attention in two different laboratory settings, one using a joystick and another using a steering wheel?

Manuscript received June 9, 2018; revised November 2, 2018 and February 4, 2019; accepted March 5, 2019. Date of publication April 11, 2019; date of current version November 21, 2019. This work was supported by the National Science Foundation under Grant 1645657 to Ohio State University. This paper was recommended by Associate Editor L. Rothrock. *(Corresponding author: Richard J. Jagacinski.)*

R. J. Jagacinski is with the Department of Psychology, Ohio State University, Columbus, OH 43210 USA (e-mail: jagacinski.1@osu.edu).

E. Rizzi was with the Department of Psychology, Ohio State University, Columbus, OH 43210 USA. He is now with the Department of Psychology, Florida International University, Miami, FL 33199 USA (e-mail: erizzi@fiu.edu).

B. J. Bloom is with Department of Computer Science, Ohio State University, Columbus, OH 43210 USA. He is now with Magellan Health, Worthington, OH 43085 USA (e-mail: bbloom22@gmail.com).

O. A. Turkkan was with the Department of Mechanical and Aerospace Engineering, Ohio State University, Columbus, OH 43210 USA. He is now with ServiceNow, Santa Clara, CA 95054 USA (e-mail: anil.turkkan@servicenow.com).

T. N. Morrison and H. Su are with the Department of Mechanical and Aerospace Engineering, Ohio State University, Columbus, OH 43210 USA (e-mail: morrison.730@osu.edu; su.298@osu.edu).

J. Wang was with the Department of Mechanical and Aerospace Engineering, Ohio State University, Columbus, OH 43210 USA. He is now with the Department of Mechanical Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: jwang@austin.utexas.edu).

Digital Object Identifier 10.1109/THMS.2019.2906612

3) Do subjects' exhibit a continuously decreasing pattern of attention for preview times farther into the future as suggested by Miller's optimal control analysis [7], or do they exhibit two discrete points of attention, one close and one far, as suggested by empirical research on car driving and piloting [8]–[11].

II. Method

A. Participants

Four researchers familiar with the measurement technique of Jagacinski *et al*. [3] participated in the study. They ranged in age from 22 to 30 years. Subjects 1 and 3 had participated in the research program for more than 2 years and had significant practice on the experimental task involving the joystick (see below). Subjects 2 and 4 had a slight amount of practice prior to the experiment. All participants passed a test for 20/25 corrected vision. The research was approved by the Institutional Review Board at Ohio State University. Informed consent was obtained from each participant.

B. Apparatus

A winding roadway was tracked in two different laboratories. In Laboratory A, subjects viewed the winding roadway on a 24-in BenQ LED monitor from a distance of 26 in (66 cm). They manipulated a one-dimensional joystick (Measurement Systems 525 constrained to one axis) to keep a circular cursor below a cross that indicated the center of the roadway. Preview of the roadway was provided at 0.05, 0.10, 0.20, 0.30, …,1.00 s into the future and was displayed as two curvy edge lines (see Fig. 1). The horizontal separation of the edge lines decreased by 19% from 0 to 1.00 s of the preview to give an impression of depth. In Laboratory B subjects viewed the winding roadway on a Christie Digital HoloStage Minicave from a distance of 74 in (183 cm). This display provided a low fidelity simulation of driving down a country road at a challenging speed with no additional traffic. The drivers sat in a Playseat and manipulated a steering wheel (Logitech G29 Racing Wheel with no force feedback) to keep the circular cursor below the cross indicating the center of the roadway [see Figs. 1 (left) and 2]. The depiction of the winding roadway was the same as in Laboratory A. The horizontal range of the roadway center was approximately 4.8° of visual angle to the right and left in both laboratories, and both displays were updated at 60 Hz. Both roadway displays could be viewed without shifting one's gaze. The sensitivity of the joystick was 0.27° of visual angle per 1° of joystick movement, which was controlled by finger and wrist movements. The sensitivity of the steering wheel was 0.14° of visual angle per 1° of steering wheel movement, which was controlled with arm movements. Both systems directly controlled the position of the circular cursor.

C. Procedure

Each experimental session consisted of three blocks of four trials. Each trial began with 10 s of warm-up tracking followed by 164 s of data collection. The roadway consisted of the sum of ten sinusoids, six with high amplitudes and four with amplitudes that were one-fifth larger. The six high amplitude sinewaves determined the input bandwidth, which was approximately 3 rad/s. One block had additional sinusoidal observation-noise disturbances added to the display of the roadway, one block had sinusoidal wind gust disturbances of the cursor, and one block had no added disturbances. A full counterbalance of the ordering of the blocks was used across each set of six sessions. The initial phases of the sinusoids varied from trial to trial. There were three sets of six sessions for a total of 18 sessions. Subjects 1 and 2 had 12 sessions with the steering wheel followed by six sessions with the joystick. Subjects 3 and 4 had 12 sessions with the joystick followed by six sessions with the steering wheel. There was a one-to-five-week break between the second and third sets. After each block subjects received feedback on their median root-mean-squared error over the four trials.

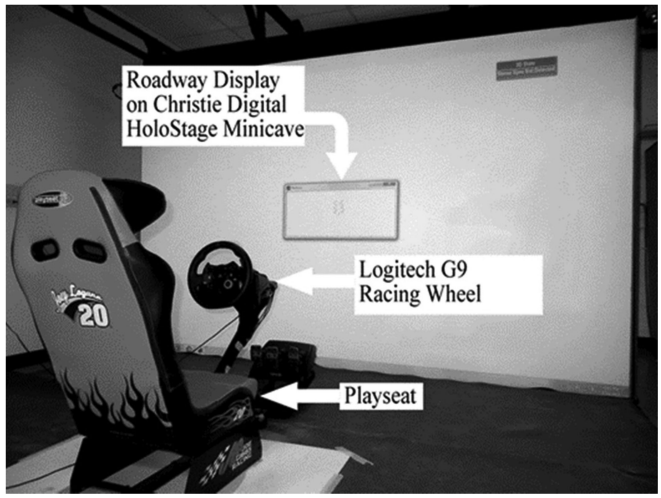


Fig. 2. Laboratory B with a steering wheel controller.

D. Measure of Attention to Preview

In the block with additional sinusoidal observation noise, ten different sinusoids with distinct frequencies perturbed the road at ten preview times ranging from 0.1 to 1.0 s into the future. Fourier analysis of the joystick or steering wheel movements determined the median amplitude at each of these ten frequencies. The ratio of this amplitude to the median amplitude at that same frequency when there was no additional observation noise provided a signal-to-noise ratio that was interpreted as a measure of attention to each of the ten previewed roadway positions. The observation noise frequencies interleaved the frequencies of the roadway and were arranged so that the fastest perturbation frequency was at the 0.1 s preview time, and the slowest was at the 1.0 s preview time. Previous research revealed that this ordering resulted in higher signal-to-noise ratios than the reverse ordering [3].

E. Measure of Error Nulling

Wind gust disturbances consisting of ten frequencies different from the roadway frequencies were added to the cursor position

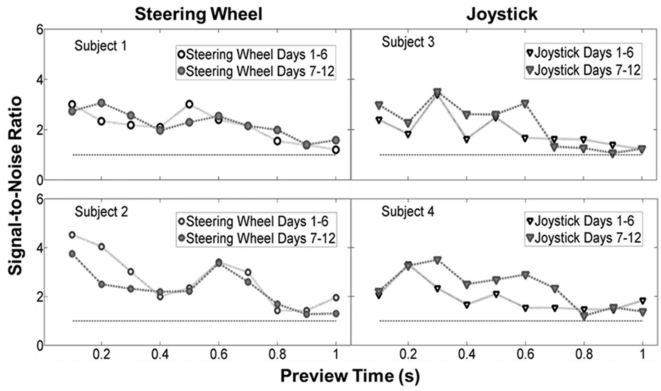


Fig. 3. Attentional signal-to-noise ratios for Days 1–6 and Days 7–12. Circles represent the steering wheel, and inverted triangles represent the joystick (pivoting about a fixed axis). The unfilled symbols are for Days 1–6, and the gray symbols are for Days 7–12. Subjects used the same control device for Days 1–12.

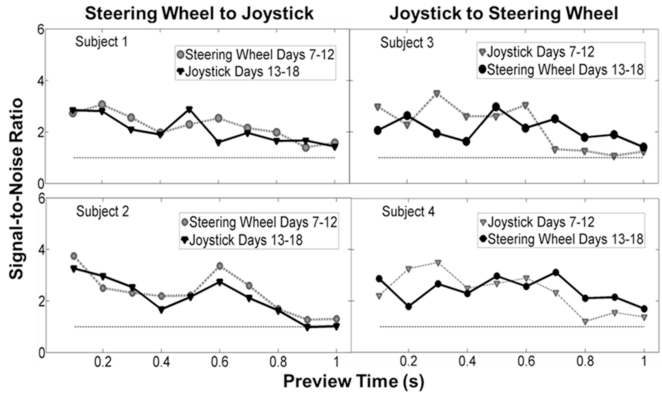


Fig. 4. Attentional signal-to-noise ratios for Days 7–12 and Days 13–18. Circles represent the steering wheel, and inverted triangles represent the joystick. The gray symbols are for Days 7–12, and the black symbols are for Days 13–18 when the subjects switched to a different control device.

for one of the three blocks of trials. Because these disturbances were not previewed, the response of the subject at these frequencies provided a measure of error nulling independent of the response to preview. Fourier analyses were conducted to determine the median amplitude ratio and phase shift from error to system output (commanded cursor position). A simplified McRuer crossover model was fit to these data [12]. This model posits that the describing function for the person plus the control system can be approximated as a gain, an integrator, and a time delay in a feedback loop. The log–log plot of output to error amplitude ratio versus frequency for this model is linear and has a slope of –1 due to the integrator. The frequency in rad/s at which the amplitude ratio is equal to 1 is numerically equal to the gain K. A plot of phase shift versus frequency is also linear and has an intercept of –90° due to the integrator in the crossover model. The slope of the linear trend, rad versus rad/s, is equal to the time delay. The median amplitude ratio and phase shift were calculated across the four trials in a block at each measurement frequency. The medians at the middle six frequencies out of ten were used to estimate the gain and time delay for a block of trials.

III. Results

A. Attention Distribution

A comparison of the attention signal-to-noise ratios at the ten measured preview times for Days 1–6 and Days 7–12 is shown in Fig. 3. All subjects exhibited significant effects of the preview time ($p < .01$) and generally exhibited stable patterns across the two sets of six days. Only Subject 4 showed a statistically significant practice effect ($F(1,10) = 4.95$, $p = .05$) corresponding to higher signal-to-noise ratios on Days 7–12.

A comparison of the attention signal-to-noise ratios for Days 7–12 and Days 8–13 is shown in Fig. 4. All subjects exhibited significant effects of the preview time ($p < .01$). Only Subject 3 exhibited a significant effect of joystick versus steering wheel, an interaction ($F(9,90) = 2.18$, $p < .05$) reflecting a greater emphasis of longer preview times with the steering wheel. The other three subjects showed relatively stable patterns of attention across the joystick and steering wheel.

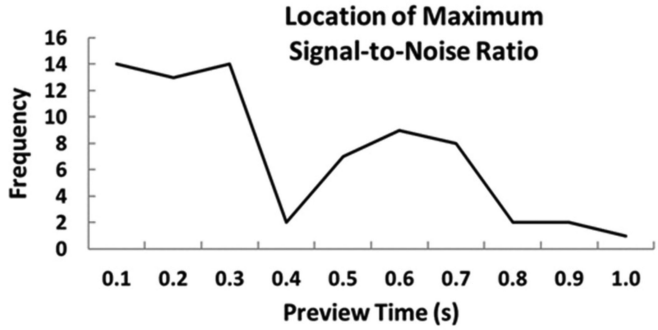


Fig. 5. Preview times at which the *maximum* signal-to-noise ratio occurred across all subjects and days (4 subjects × 18 days = 72 signal-to-noise attentional distributions).

Despite the relative stability of the six-day average attentional patterns, there was striking variability in the day-to-day measures of attention. Fig. 5 shows the preview times corresponding to the highest daily signal-to-noise ratio across all 18 days and four subjects. There are two groups of preview times with similarly high frequencies. Preview times 0.1, 0.2, and 0.3 s had the maximum signal-to-noise ratio on more than half of the sessions (57%), and preview times 0.5, 0.6, and 0.7 s had the maximum signal-to-noise ratio on one-third of the sessions (33%).

Plots of the signal-to-noise ratios for Days 8–11 are shown in Figs. 6 and 7 for the two subjects with the lowest error scores on Days 7–12. Subject 2 showed a strong peak at 0.6 s on Days 8 and 11; on Days 9 and 10 there was a strong peak at 0.1 s. Subject 4 showed strong peaks at 0.5 and 0.6 s on Days 8 and 9; peaks at times 0.1, 0.2, or 0.3 s occurred for all four days. These patterns of instability across days and the two regions of maximal signal-to-noise ratio in Fig. 5 suggest that the peaks in the 0.1–0.3 s range and the 0.5–0.7 s range are distinct foci of attention. They may alternate across days as illustrated in Fig. 6, or they may co-occur as illustrated in Fig. 7.

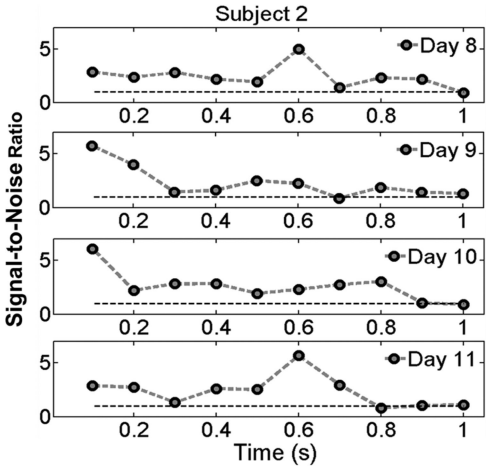


Fig. 6. Attentional signal-to-noise ratios for Days 8–11 for Subject 2 using a steering wheel.

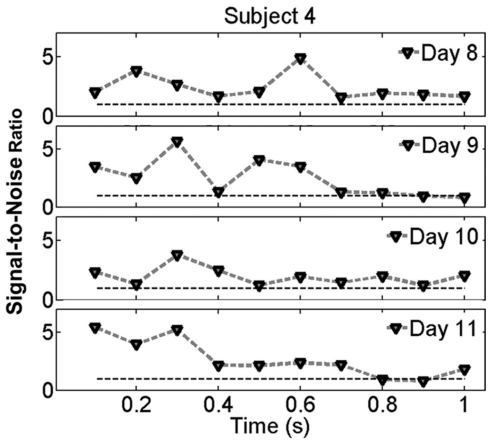


Fig. 7. Attentional signal-to-noise ratios for Days 8–11 for Subject 4 using a joystick.

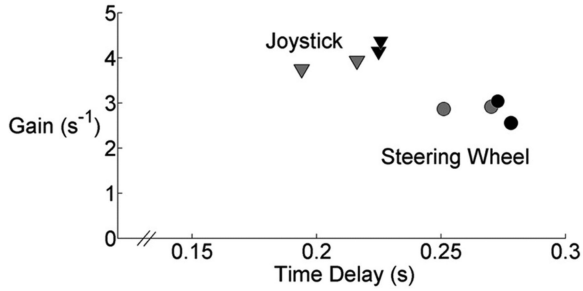


Fig. 8. Error nulling gains (K) and time delays estimated from the crossover model. Each symbol represents a single subject. Gray symbols are for Days 7–12. Black symbols are for Days 13–18.

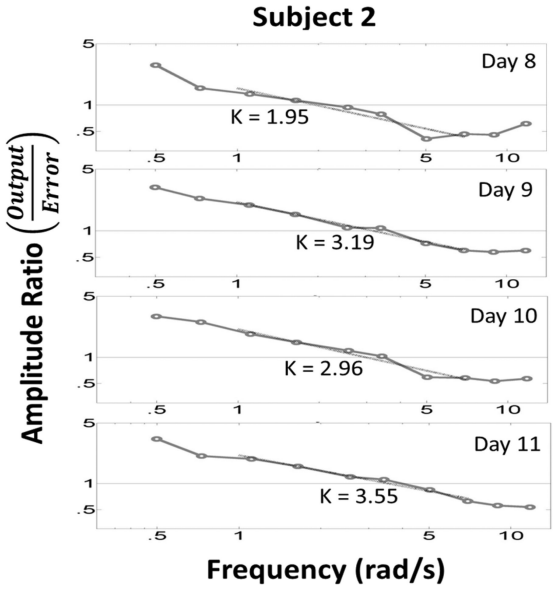


Fig. 9. Amplitude ratios for the relationship between system output (commanded cursor position) and error for Days 8–11 for Subject 2 using a steering wheel.

Another possibility is that these peaks are a result of measurement noise. However, similar experiments [13] in which only part of the preview was visible provided an estimate of measurement noise at preview positions that were not visible to the subject. Signal-to-noise ratios of magnitude 3 or greater spuriously occurred less than 1% of the time at these hidden preview positions for the data in single daily sessions. These experiments used a rate control, which has exhibited similar signal-to-noise patterns as a position control system. Given the large magnitudes of the signal-to-noise peaks in the daily sessions in Figs. 6 and 7, it is highly unlikely that they reflect measurement noise.

B. Error Nulling

An analysis of error nulling on Days 7–12 and Days 13–18 using the McRuer Crossover Model revealed systematic differences in feedback control between the joystick and steering wheel for each subject ($p < .01$). The crossover frequency at which the amplitude ratio equals 1.0 was estimated from a linear fit to the middle six measurement frequencies plotted logarithmically against log amplitude ratio.

The crossover frequency is numerically equal to the gain K in the McRuer Crossover Model. The time delay was estimated from the slope of a linear fit to phase lag plotted against the middle six measurement frequencies. The joystick controller resulted in higher gains K and lower time delays than the steering wheel (see Fig. 8). These differences can be attributed to biomechanical differences between the limbs (fingers and wrist for the joystick versus arms for the steering wheel) and hardware (light joystick handle versus more massive steering wheel). Similar differences in effective time delay are reported in [14]. The pattern of feedback control performance was highly stable across days, in contrast to the attentional signal-to-noise ratios. This stability is exemplified in the decreasing pattern of amplitude ratios for Subject 2 in Fig. 9.

Given the 3 rad/s input bandwidth, one would expect the crossover model gain to be 3 or higher [12]. The gain K is numerically equal to the frequency in rad/s at which the output to error amplitude ratio "crosses over" from above 1.0 to less than 1.0. With the steering wheel subjects exhibited gains near 3. With the joystick their gains were around 4 (see Fig. 8).

IV. Modeling Attentional Pattern Formation

A dynamic model of selective attentional emphasis and inhibition of surrounding regions is proposed to interpret the patterns of spacing and day-to-day instability in the attentional signal-to-noise ratios (see Figs. 5–7). Broadly speaking, the function of attention is to emphasize particular aspects of a creature's informational environment that are relevant to achieving present goals and/or the need to switch goals. Selective emphasis is needed because of a creature's limited cognitive and action capabilities in information-rich environments. Sustained stable attention to fixed spatiotemporal loci may be necessary for particular tasks like tracking a moving target. However, if attention was very stable, it might limit responses to new stimuli that indicate a need to interrupt the present task and switch goals to address some imminent danger or opportunity. It would therefore not be surprising if attention had a level of stability that could be easily interrupted by environmental perturbations.

Previous efforts to model aspects of attentional dynamics include the metaphor of a spotlight which can be quickly moved to different locations [15], [16], a spotlight or zoom lens with adjustable width [17], [18], and internal oscillators which can become entrained with external rhythmic patterns as in musical contexts [19]. The present modeling effort posits two component processes, selective attentional emphasis (*A*) and inhibition of surrounding regions (*I*) that have been widely discussed by attentional researchers (see [20] for a review). This model describes the shapes of average attentional distributions over the spatiotemporal display of the preview as well as variability over successive blocks of trials. Leber [21] noted strong trial-to-trial variations in the degree to which attention could be captured by irrelevant stimuli, and also noted correlated changes in brain activity in the middle frontal gyrus. The present model emphasizes attentional instability as a key aspect of behavior and tries to exploit the detailed spatiotemporal structure revealed by the present measurement technique to understand attentional dynamics.

The present model of attention is an adaptation of a type of biological model of dynamic pattern formation that has been used to model the development of embryos [22] and cortical feature maps [23], the rhythmic spacing of color striations on seashells [24], and similarly many other examples of biological pattern formation (see [25] for a review). This type of model was introduced by Turing [26] to describe the emergence of features in embryos (morphogenesis) from relatively uniform initial conditions. It is called a reaction-diffusion model and consists of two processes, an activator and an inhibitor (e.g., [24, p. 23], Table I) whose dynamics are described by a system of two partial differential equations

$$\begin{aligned}\partial A/\partial t &= \left(e^{-M(x-0.1)/0.9}\right)\left(sA^2/I + sb_A\right) - r_A A \\ &\quad + D_A \partial^2 A/\partial x^2 \\ \partial I/\partial t &= \left(e^{-M(x-0.1)/0.9}\right)\left(sA^2 + b_I\right) - r_I I \\ &\quad + D_I\, \partial^2 I/\partial x^2 \end{aligned} \tag{1}$$

TABLE I
GMM Model

Symbol	Quantity	Values
A	Attentional emphasis	$0 \le A \le 5$ initial conditions = 1.5
I	Inhibition of surrounding regions	$1 \le I$ initial conditions = 1.5
x	Spatial display of preview actively attended	$0.1 \le x \le 1.0$ in Fig. 10 $0.1 \le x \le 0.9$ in Fig. 11 $0.1 \le x \le 0.9$ in Fig. 12
s	Production rate amplifier	0.15[1 + 0.03norm. distrib. (0,1)]
b_A	Production rate constant for A	0.01
b_I	Production rate constant for I	0.00
r_A	Decay rate for A	0.03
r_I	Decay rate for I	0.06
D_A	Diffusion coefficient for A	0.0011 in Fig. 10 0.01 in Fig. 11 0.01 in Fig. 12
D_I	Diffusion coefficient for I	0.0440 in Fig. 10 0.40 in Fig. 11 0.40 in Fig. 12
M	Exponential shaping of production rate	0.00 in Fig. 10 0.45 in Fig. 11 0.90 in Fig. 12

The activator (*A*) has positive feedback such that it grows in strength over time once started by a sufficiently large perturbation in growth rate or initial condition. The activation slowly spreads through space or diffuses at a rate determined by D_A. The activator also creates an inhibitor (*I*) at that location that limits its growth. The inhibitor spreads much more quickly through space to limit the growth of other nearby activators ($D_I > D_A$). The activator and inhibitor dynamics also include decay rates, r_A and r_I, that limit their growth and constants s, b_A, and b_I that adjust the production rates (see Table I). This model has qualitative characteristics of emphasizing certain regions in a patterned manner and of stochastic variability. These are qualitative characteristics of human attention, so we wanted to test whether this common dynamic found in various species could also describe aspects of human attention.

For modeling attention, the activator process (*A*) will be attentional emphasis, and the inhibitor process (*I*) will be inhibition of attention to immediately surrounding regions. The general intent is to consider the spatial distribution of attention as an instance of a general class of biological pattern formation processes. This model has a rhythmic spacing of areas of attentional emphasis that is directly related to the magnitude of the diffusion coefficients, D_A and D_I [27]. In Fig. 10 the diffusion coefficients are relatively small (see Table I), and multiple peaks of attentional emphasis are tightly spaced. The model

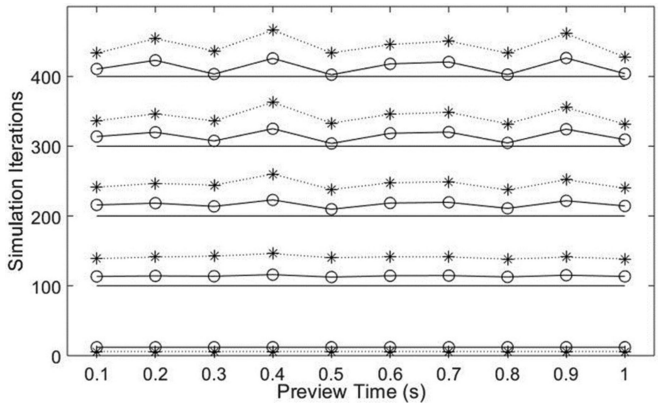


Fig. 10. Rhythmic spacing of the attentional emphasis evolving from uniform initial conditions and random perturbations of growth rate, s in (1). Circles represent attentional emphasis (A), and $*$'s represent inhibition ($I/2$). Small values of the diffusion coefficients, D_A and D_I, produce tight spacing of the regions of the attentional emphasis after 400 temporal iterations. See Table I for parameter values.

was simulated with ten discrete values for x, which correspond to the positions of the various preview times in the display in Fig. 1.

To extend this model to the tracking task, an additional decreasing exponential function multiplies the attentional emphasis (A) and inhibitor (I) production rates (1). The constants 0.1 and 0.9 scale the exponential multiplier to equal 1.0 when $x = 0.1$ and e^{-M} when $x = 1.0$. This term reflects the optimal control solution for tracking with finite preview developed by Miller [7]. Namely, Miller showed that the optimal attentional distribution to preview is a decreasing exponential for a velocity control system. The present experiment used a position control. If a position control is approximated as a lag with a high bandwidth, Miller's method gives the corresponding optimal attentional weighting of the preview as an exponential function that rapidly decreases with increasing preview times. To reflect this task demand to emphasize shorter preview times, both the attentional emphasis and inhibitor production rates are multiplied by a decreasing exponential. This function will therefore favor the growth of attentional emphasis at short preview times. A similar positional gradient was used by Gierer and Meihnardt [22] in models of embryo development. The equations for the combined model will be referred to as the Gierer–Meinhardt–Miller Model or GMM model.

A second way of controlling the inherent rhythmicity of spatial attention is to limit the range of attention. The locus of spatial attention has been previously described as having an adjustable range or width [15], [17], [18]. The spatial rhythmicity of the reaction-diffusion dynamics is more evident when the spatial range of attention is larger. Therefore, a control strategy for limiting this rhythmicity in the present task and emphasizing short preview times would be to limit the range of attention to less than the full range of ten positions (preview times) shown in Fig. 1. In the present modeling effort (see Figs. 11 and 12), the active range of attention was limited to nine positions to better approximate the subjects' data.

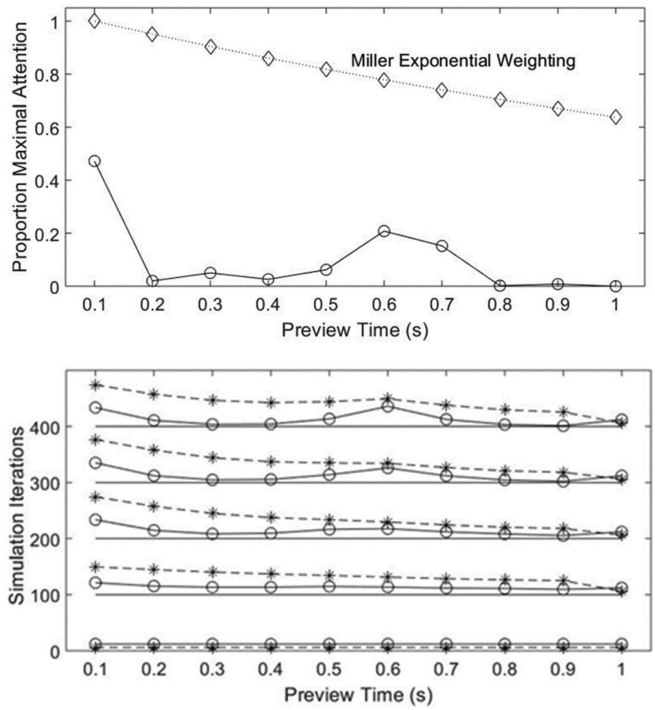


Fig. 11. (Bottom) Evolution of an attentional distribution to preview with a *weak* exponential decrease ($M = 0.45$) in the GMM model. Circles represent attentional emphasis (A), and $*$'s represent inhibition ($I/2$). (Top) Frequency distribution of the location of the *maximal* attention (A) over 500 simulations of the model with 400 temporal iterations each.

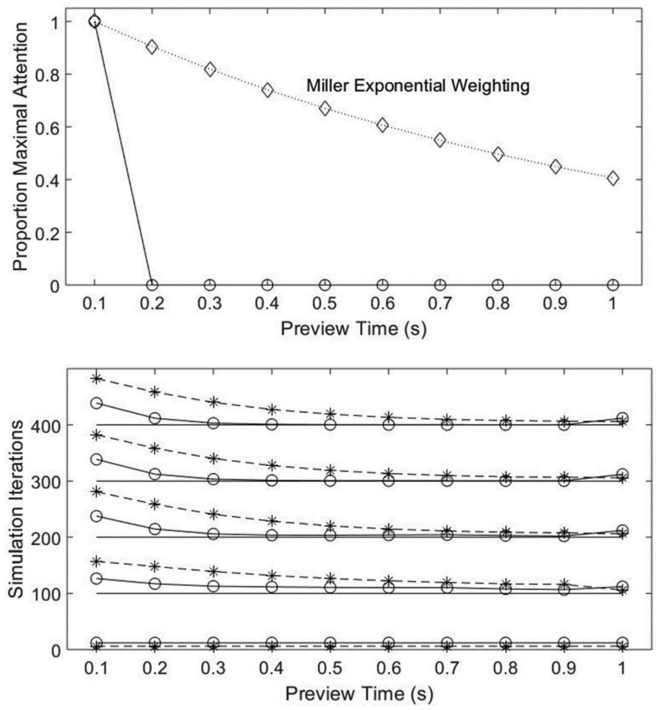


Fig. 12. (Bottom) Evolution of an attentional distribution to preview with a *strong* exponential decrease ($M = 0.90$) in the GMM Model. Circles represent attentional emphasis (A), and $*$'s represent inhibition ($I/2$). (Top) Frequency distribution of the location of the *maximal* attention (A) over 500 simulations of the model with 400 temporal iterations each.

Figs. 11 and 12 show the development of attentional distributions from uniform initial conditions of A and I. The model simulation begins with a random set of perturbations of the attentional emphasis and inhibition production rates, s in (1), at each of nine preview times (positions on the tracking display in Fig. 1). As some of these attentional foci start to grow, they simultaneously send out rapidly diffusing inhibition that limits the attentional growth at other positions. As these spatially distributed processes interact over time they can produce patterns of primary emphasis at short preview times (see Fig. 12) or primary emphasis at short and/or intermediate preview times (see Fig. 11) depending on the exponent of the Miller function. Fig. 11 (top) mimics the two-clump pattern in Fig. 5, although the clump at the shortest preview time does not show the variability exhibited by the subjects. This attentional pattern formation is presumed to occur early in a trial and then continue throughout the remainder of the trial.

This model attributes the daily instability in attentional focus in Figs. 5–7 to the reaction-diffusion dynamics and the form of the Miller shaping function. The interaction of the attentional emphasis and inhibition of surrounding regions has a level of stability that is sensitive to small random variations in the production rates. The Miller exponential shaping is a way of trying to control these sensitive processes to emphasize short preview times. The Miller function is a smooth exponential; however, because it acts on complex reaction-diffusion dynamics, the resulting attentional distributions (see Fig. 11) may not be smoothly decreasing, but instead form clumps of emphasis. The clumps are due to the reaction-diffusion dynamics, which is a model of biological spacing. In the case of Meinhardt's seashells [24], the spacing dynamics lead to highly differentiated patterns of coloration. In the case of attention, the advantage of spacing may be to limit focusing too much of a limited cognitive resource in a single delimited region.

V. Discussion

A. Detailed Measure of Attention

This longitudinal study revealed that the feedforward attentional pattern of signal-to-noise ratios was generally stable when averaged across multiple six-day periods and transferred across physically different control devices and limb movements for most subjects (see Figs. 3 and 4). In contrast, feedback parameters of error nulling (gain and time delay) were strongly influenced by the control devices and limbs (see Fig. 8). This pattern of selective influence supports the common modeling assumption that feedforward control and feedback control are distinct behavioral processes [1], [2]. Converging evidence is that this same measurement technique found roadway bandwidth to selectively influence feedback control and not feedforward control [3]. With much more complex dynamic systems, feedforward also exhibits longer learning times [28]. Therefore, feedback and feedforward are behaviorally distinct processes.

In contrast to the stability of six-day average attentional signal-to-noise distributions, day-to-day variability was quite marked. One interpretation of these data is that the peaks that occurred at 0.1–0.3 s and at 0.5–0.7 s of the preview (see Figs. 5–7) are a form of attentional spacing that results from the influences of two complementary, but somewhat unstable processes. A dynamic theory of the attentional emphasis and inhibition of surrounding regions was proposed as underlying processes to account for this pattern of instability. The instability of the feedforward attentional dynamics contrasts sharply with the relatively stable pattern of feedback movement dynamics (see Fig. 9) that were approximated as a simple lag with an internal time delay, i.e., the crossover model [12].

One objection to this interpretation might be that some researchers have proposed two separate loci of attention to preview for car driving (e.g., [8]–[10]; for helicopter control, see [11]). For example, the steering dynamics of a car can be approximated as an acceleration control system. Feedback control requires lead compensation, which can be implemented by using yaw error to extrapolate current lateral position error [29]. The perception of these angular and lateral errors would occur close to the present car position. In contrast, Land and Horwood [8] estimated that the perception of curvature for feedforward control is at about 0.8 s of preview. Miller [7] used optimal control theory to predict heightened feedforward attention in the region of 0.7 s of preview depending on the relative emphasis of error versus effort with an acceleration control system. Land and Lee [30] and Macadam [31] have estimated that preview even greater than 1 s is useful in car driving. Therefore, the information requirements for feedforward and feedback control would *primarily* create the need to attend to two distinct regions of preview, one close and one far. This interpretation would not preclude a behavioral role for the presently proposed attentional dynamics to influence the relative positions of the two attentional regions and their relative stability.

In contrast to typical car dynamics, the present experiment used a position control system. If one approximates a position control as a lag with a high bandwidth, one can use Miller's method [7] to determine the optimal attentional weighting for feedforward control with a position control system. The distribution has a high attention weighting at short preview times and exponentially decreases at a rapid rate with increasing preview time. In this case, one would expect feedback and feedforward control to rely on information close to the vehicle. Converging empirical evidence is provided by Ito and Ito [32]. They found that tracking error decreased with up to about 0.25 s of the preview for a position control, and that longer preview was beneficial for higher order dynamics. The two separated regions of attention found in the present experiment might therefore arise *primarily* from the proposed attentional dynamics rather than from the information requirements of the task, which are different from typical car dynamics.

Another objection to the present modeling might be that the peaks in the 0.5–0.7 s range are a result of measurement noise. As noted above, estimates of measurement noise from previous studies indicate that a signal-to-noise ratio greater than 3 occurs less than 1% of the time [13]. Second, the range of the peaks, 0.5–0.7 s, is rather delimited. 0.4 s is rare, and 0.8–1.0 s is rare (see Fig. 5). This limited range is consistent with the proposed model of attentional dynamics. Namely, the balance between the

two separate processes, *A* and *I*, leads to a relatively consistent spacing.

A fundamental structural aspect of attention is whether it is unitary or whether it can be allocated to separate spatial locations [33]. For the present model, the answer is "both." Namely, the dynamic structure of attention leads to the emphasis of separate spatial regions. However, there are continuous fields of both attentional emphasis and inhibition that span the range of attention (see Fig. 11). The spacing of the attentional regions of emphasis is not arbitrary; rather, it is constrained by the underlying dynamics. Research on eye movements has found that in walking, driving, and other tasks people tend to direct their gaze toward where they will be acting 0.5 to 1 s into the future to strengthen anticipatory responding (see [34] for a review). The present model suggests that attention can be directed at such a spatiotemporal region and simultaneously at a closer region by controlling attentional dynamics.

B. Optimal Control Considerations

Subjects in the present experiment emphasized preview positions from 0.1 to 0.7 s (see Figs. 3 and 4) rather than only the short preview times as would be expected from Miller's [7] optimal control model. There are at least three possible interpretations of this difference.

First, Miller's optimal control modeling demonstrates that the relative emphasis on minimizing mean-squared error and mean-squared effort (control stick movement) should influence the shape of the attentional distribution. The exponential function emphasizing short preview times should decrease more quickly with an emphasis on error minimization and more slowly with an emphasis on effort minimization. If these experienced subjects were emphasizing error minimization, then their attentional distributions would be expected to be short and steep. The wide range of the experienced subjects' attentional distributions suggests that they were minimizing movement effort. However, this conclusion seems unlikely given that their effort scores (root-mean-squared control stick movement) closely matched the root-mean-squared pathway excursion.

A second possible interpretation is that using the Miller exponential function to emphasize short preview times may be attentionally effortful [35]. These experienced subjects may have been trying to lessen attentional effort, which allowed the reaction-diffusion dynamics to have more influence in forming clumps of attentional emphasis. Kahneman [35] has argued that attention to multiple sources of information can occur in parallel when the overall level of attention is low. In contrast, at high levels of effort, attention is likely to be concentrated on a single source of information. The peak signal-to-noise ratios produced by these experienced subjects are high (see Figs. 6 and 7), which suggests the subjects were not minimizing attentional effort.

A third possible interpretation is that the observation noise added to the display of the path to measure the signal-to-noise distributions created greater uncertainty in the previewed path position. These experienced subjects may have used a wide range of attentional emphasis in order to obtain a more reliable estimate of path position by combining information from multiple positions. This wide range of attention then allowed the reaction-diffusion dynamics to create clumps of attentional emphasis that would not occur with a more restricted range of attention. The simulation in Fig. 11 used a range of attention from 0.1 to 0.9 s, which was sufficient to demonstrate the secondary clump of attention around 0.6 s of preview. If the range of attention is restricted to 0.5 s in the simulation, there is no secondary clump of emphasis. The range of attention may be an important behavioral variable in allowing the spatiotemporal rhythmic nature of attention to be evident.

C. Conclusions and Future Directions

In summary, from an abstract perspective, the spatiotemporal distribution of attention may be considered an example of spacing dynamics similar to those that have been investigated by biologists in contexts ranging from the striation of seashells to the formation of branch structures [25]. This pattern formation is a type of complex spatial rhythm. In the present context, the implication is that attentional dynamics are inherently rhythmical, a point that has been previously raised by Jones and colleagues in their studies of attention to sound patterns [19], [36], [37] and in more physiological theorizing [38], [39]. Two types of control strategies for dealing with attention's inherent rhythmicity in the present task are spatial biasing of the growth processes for attention and inhibition (Miller exponential) and limiting the range of attention. The details of how these attentional control strategies may interact with movement control in other contexts is a topic for future research.

This new technique for examining attention to preview should be explored with higher order dynamics more representative of automotive control. The present method does not rely on eye movements, but instead measures which aspects of the preview are coupled to the driver's control movements. Eye movements are correlated with attention and can indicate brief changes in attentional focus. However, as noted by Land [40] it is often not clear what aspects of the visual field are being emphasized for a given gaze direction. The present methodology can be used in conjunction with eye movements to provide additional detail regarding drivers' attention. Future research should also develop techniques to shorten the present 3-min measurement period and reduce the amount of observation noise needed to assess the attentional distributions.

The present study used nonnaive subjects and measured their performance over an extended period of time. These conditions maximize the likelihood of stationary performance and strengthen the argument that observed variations in the attention distribution reflect an inherent instability. Future research can proceed to examine how individual differences in driving skill [41], [42] are related to attentional distributions in various populations with the goal of improving driver safety.

ACKNOWLEDGMENT

The authors thank R. A. Miller for helpful comments and S. Ruland and D. Findlay for technical support.

REFERENCES

[1] E. Donges, "A two-level model of driver steering behavior," *Human Factors*, vol. 20, no. 6, pp. 691–707, Dec. 1978.

[2] D. T. McRuer, R. W. Allen, D. H. Weir, and R. H. Klein, "New results in driver steering control models," *Human Factors*, vol. 19, no. 4, pp. 381–397, Aug. 1977.

[3] R. J. Jagacinski, G. M. Hammond, and E. Rizzi, "Measuring memory and attention to preview in motion," *Human Factors*, vol. 59, no. 5, pp. 796–810, Aug. 2017.

[4] W. W. Johnson and A. V. Phatak, "Modeling the pilot in visually controlled flight," *IEEE Control Syst. Mag.*, vol. 10, no. 5, pp. 24–26, Aug. 1990.

[5] T. B. Sheridan, "Three models of preview control," *IEEE Trans. Human Factors Electron.*, vol. HFE-7, no. 2, pp. 91–102, Jun. 1966.

[6] W. H. Levison, "A model for mental workload in tasks requiring continuous information processing," in *Mental Workload: Its Theory and Measurement*, N. Moray, Ed. New York, NY, USA: Plenum, 1979, pp. 189–218.

[7] R. A. Miller, "On the finite preview problem in manual control," *Int. J. Syst. Sci.*, vol. 7, no. 6, pp. 667–672, 1976.

[8] M. F. Land and J. Horwood, "Which parts of the road guide steering?," *Nature*, vol. 377, no. 6547, pp. 339–340, Sep. 1995.

[9] D. D. Salvucci and R. Gray, "A two-point visual control model of steering," *Perception*, vol. 33, no. 10, pp. 1233–1248, 2004, doi: 10.1068/p5343.

[10] C. Sentouh, P. Chevrel, F. Mars, and F. Claveau, "A sensorimotor driver model for steering control," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Piscataway, NJ, USA, 2009, pp. 2462–2467.

[11] R. A. Hess and K. K. Chan, "Preview control pilot model for near-earth maneuvering helicopter flight," *J. Guid., Control, Dyn.*, vol. 11, pp. 146–152, 1988.

[12] D. T. McRuer and H. R. Jex, "Review of quasi-linear pilot models," *IEEE Trans. Human Factors Electron.*, vol. HFE-8, no. 3, pp. 231–249, Sep. 1967.

[13] E. Rizzi, "The relationship between attention to preview and action during roadway tracking," unpublished, Ph.D. dissertation, Dept. Psychol., Ohio State Univ., Columbus, OH. [Online]. Available: https://etd.ohiolink.edu/pg_10?::NO:10:P10_ETD_SUBID:169098

[14] M. Martinez-Garcia, T. Gordon, and L. Shu, "Extended crossover model for human-control of fractional order plants," *IEEE Access*, vol. 5, pp. 27622–27635, 2017.

[15] P. L. Wachtel, "Conceptions of broad and narrow attention," *Psychol. Bull.*, vol. 68, no. 6, pp. 417–429, Dec. 1967.

[16] M. I. Posner, "Orienting of attention," *Quart. J. Exp. Psychol.*, vol. 32, no. 1, pp. 3–25, Feb. 1980, doi: 10.1080/00335558008248231.

[17] D. LaBerge, "Spatial extent of attention to letters in words," *J. Exp. Psychol.: Human Perception Perform.*, vol. 9, no. 3, pp. 371–379, Jun. 1983.

[18] C. W. Eriksen and J. D. St. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception Psychophys.*, vol. 40, no. 4, pp. 225–240, Oct. 1986.

[19] E. W. Large and M. R. Jones, "The dynamics of attending: How people track time-varying events," *Psychol. Rev.*, vol. 106, no. 1, pp. 119–159, Jan. 1999.

[20] C. I. Folk, "Controlling spatial attention: Lessons from the lab and implications for everyday life," in *The Handbook of Attention*, J. M. Fawcett, E. F. Risko, and A. Kingstone, Eds. Cambridge, MA, USA: MIT Press, 2015, ch.1, pp. 3–25.

[21] A. B. Leber, "Neural predictors of within-subject fluctuations in attentional control," *J. Neurosci.*, vol. 30, no. 34, pp. 11458–11465, Aug. 2010, doi: 10.1523/JNEUROSCI.0809-10.2010.

[22] A. Gierer and H. Meinhardt, "A theory of biological pattern formation," *Kybernetik*, vol. 12, pp. 30–39, 1972.

[23] S. P. Wilson and J. A. Bednar, "What, if anything, are topolgical maps for?" *Developmental Neurobiol.*, vol. 75, no. 6, pp. 667–681, Jun. 2015.

[24] H. Meinhardt, "Pattern formation by local self-enhancement and long range inhibition," in *The Algorithmic Beauty of Sea Shells*, Berlin, Germany: Springer, 1995, ch. 2, pp. 19–39.

[25] P. Ball, "Bodies," in *The Self-Made Tapestry: Pattern Formation in Nature*. New York, NY, USA: Oxford Univ. Press, 1999, ch. 4, pp. 77–109, 132.

[26] A. Turing, "The chemical basis of morphogenesis," *Philos. Trans. Roy. Soc.*, vol. B 237, pp. 37–72, 1952.

[27] D. S. Jones and B. D. Sleeman, "Turing diffusion driven instability and pattern formation," in *Differential Equations and Mathematical Biology*, Boca Raton, FL, USA: Chapman & Hall, 2003, ch. 12, sec. 5, pp. 282–296. (Originally published in Boston, MA, USA: Allen & Unwin, 1983.)

[28] X. Zhang, S. Wang, J. B. Hoagg, and T. M. Seigler, "The roles of feedback and feedforward as humans learn to control unknown dynamic systems," *IEEE Trans. Cybernetics*, vol. 48, no. 2, pp. 543–555, Feb. 2018, doi: 10.1109/TCYB.2016.2646483.

[29] D. T. McRuer, D. H. Weir, H. R. Jex, R. E. Magdaleno, and R. W. Allen, "Measurement of driver-vehicle multiloop response properties with a single disturbance input," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, no. 5, pp. 490–497, Sep. 1975.

[30] M. F. Land and D. N. Lee, "Where we look when we steer," *Nature*, vol. 369, pp. 742–744, Jun. 1994.

[31] C. C. Macadam, "Understanding and modeling the human driver," *Vehicle Syst. Dyn.*, vol. 40, nos. 1–3, pp. 101–134, 2003.

[32] K. Ito and M. Ito, "Tracking behavior of human operators in preview control systems," *Elect. Eng. Japan*, vol. 95, no. 1, pp. 120–127, 1975.

[33] M. Bay and B. Wyble, "The benefit of attention is not diminished when distributed over two simultaneous cues," *Attention, Perception Psychophys.*, vol. 76, no. 5, pp. 1287–1297, Jul. 2014.

[34] B. W. Tatler and M. F. Land, "Everyday visual attention," in *The Handbook of Attention*, J. M. Fawcett, E. F. Risko, and A. Kingstone, Eds. Cambridge, MA, USA: MIT Press, 2015, ch. 17, pp. 391–421.

[35] D. Kahneman, "Attention divided among inputs," in *Attention and Effort*. Englewood Cliffs, NJ, USA: Prentice Hall, 1973, ch. 8, pp. 136–155.

[36] M. R. Jones, "Time, our lost dimension: Toward a new theory of perception, attention, and memory," *Psychol. Rev.*, vol. 83, no. 5, pp. 323–355, Sep. 1976.

[37] M. R. Jones, "Attending to sound patterns and the role of entrainment," in *Attention and Time*, A. C. Nobre and J. T. Coull, Eds. New York, NY, USA: Oxford University Press, 2010, ch. 23, pp. 317–330.

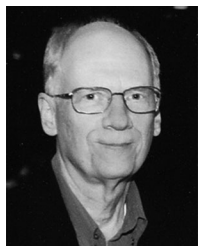
[38] A. C. Nobre and S. G. Heideman, "Temporal orienting of attention," in *The Handbook of Attention*, J. M. Fawcett, E. F. Risko, and A. Kingstone, Eds. Cambridge, MA, USA: MIT Press, 2015, ch.3, pp. 57–78.

[39] A. N. Landau and P. Fries, "Attention samples stimuli rhythmically," *Current Biol.*, vol. 22, pp. 1000–1004, Jun. 2012.

[40] M. F. Land, "The visual control of steering," in *Vision and Action*, L. R. Harris and M. Jenkin, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1998, pp. 163–179.

[41] W. Wang, J. Xi, A. Chong, and L. Li, "Driving style classification using a semisupervised support vector machine," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 5, pp. 650–660, Oct. 2017.

[42] W. Wang, J. Xi, and D. Zhao, "Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches," *IEEE Trans. Intell. Transp. Syst.*, to be published.

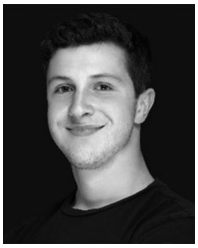


Richard J. Jagacinski received the B.S. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1968, and the Ph.D. degree in experimental psychology from the University of Michigan, Ann Arbor, MI, USA, in 1973.

He is currently a Professor with the Psychology Department, Ohio State University, Columbus, OH, USA. His research and teaching explore human attention and memory in movement control and rhythm in athletic and musical performance.

Emanuele Rizzi was born in Italy. He received the B.S. degree in psychology and statistics from Florida State University, Tallahassee, FL, USA, and the M.A. and Ph.D. degrees in psychology from Ohio State University, Columbus, OH, USA, in 2010, 2015, and 2018, respectively.

He is currently a Faculty Instructor with the Department of Psychology, Florida International University, Miami, FL, USA. His research interests include perceptual-motor control, rhythmic coordination, driving performance, and attention.



Benjamin J. Bloom received the B.S. degree in computer science and engineering from The Ohio State University, Columbus, OH, USA, in 2017.

From 2016 to 2017, he was a Software Engineer Intern with Healthy Roster, working on mobile applications for student athletes. Since 2018, he has been a Software Engineer with Magellan Health, Columbus, OH, USA, working on mobile and cloud technologies.

O. Anil Turkkan received the B.S. degree in aerospace engineering and computer science from Middle East Technical University, Ankara, Turkey, in 2010, the M.S. degree in mechanical engineering from Illinois Institute of Technology, Chicago, IL, USA, in 2013, and the Ph.D. degree in mechanical engineering from The Ohio State University, Columbus, OH, USA, in 2017.

Since 2018, he has been a Machine Learning Engineer with ServiceNow, Santa Clara, CA, USA. His research interests include text-based machine learning and natural language understanding.

Tyler N. Morrison received the B.S. degree in mechanical engineering from the University of Tulsa, Tulsa, OK, USA, in 2017. He is currently working toward the Ph.D. degree in mechanical engineering at The Ohio State University, Columbus, OH, USA.

His current research focuses on design and control of robots and automated systems, especially as they relate to physical interaction with humans.

Haijun Su received the Ph.D. degree in mechanical engineering from the University of California at Irvine, Irvine, CA, USA, in 2004.

He is currently an Associate Professor with the Mechanical and Aerospace Engineering Department, The Ohio State University, Columbus, OH, USA.

Dr. Su is currently or was an Associate Editor of *ASME Journal of Mechanical Design* and *Mechanism and Machine Theory*. In 2017, he was elected a Fellow of ASME.

Junmin Wang (SM'14) received the B.E. degree in automotive engineering and the M.S. degree in power machinery and engineering from Tsinghua University, Beijing, China, in 1997 and 2000, respectively, the M.S. degrees in electrical engineering and mechanical engineering from the University of Minnesota Twin Cities, Minneapolis, MN, USA, in 2003, and the Ph.D. degree in mechanical engineering from The University of Texas at Austin, Austin, TX, USA, in 2007.

He is currently the Accenture Endowed Professor of Mechanical Engineering with the University of Texas at Austin. His research interests include control, modeling, estimation, optimization, and diagnosis of dynamical systems, especially for automotive systems.

Dr. Wang is an IEEE Vehicular Technology Society Distinguished Lecturer, SAE Fellow, and ASME Fellow.

Queueing Network Based Driver Model for Varying Levels of Information Processing

Ye Lim Rhie, Ji Hyoun Lim, *Member, IEEE*, and Myung Hwan Yun, *Member, IEEE*

***Abstract*—With growing interest in the topic of smart computing and context-aware services, identifying driver's intention has become important in automotive industry. Among existing studies on driver assistance systems, few studies have used the concept of the levels of processing (LOP) to understand driver's information processing. This paper introduces experimental and computational studies that connect human behavior data to internal goal of a driver when an in-vehicle task involves visual search. In the case of consuming information displayed on a car instrument cluster, we considered two levels of information processing, perceptual and cognitive. Through an empirical study, we observed different human oculomotor behaviors that depend on the LOP by evaluating the reaction time (RT) and eye movement patterns. Further, we suggested different ways of information processing that can be represented as different routes in the underlying cognitive architecture of a queueing network. Simulation study demonstrated that trends in simulated oculomotor behavior were similar to those observed in the experiment, that is, RT at the cognitive LOP was shorter than that at the perceptual LOP, while eyes fixated for shorter duration and with lower frequency. In terms of application, this study reveals the possibility of using human-generated data for evaluating the innate purposes of drivers. Together with further development of sensors and invasive computing, the approach proposed in this paper could assist the realization of cognitive cars by understanding drivers' intent using eye tracking technology.**

***Index Terms*—Computational modeling, context-aware services, driver model, levels of processing (LOP).**

I. Introduction

With growing interest in the topic of smart computing and context-aware services, use of biometric sensors to recognize physiological and mental state of drivers has been increasing among manufacturers. Ford [1] has developed sensors in driver's seat to detect driver fatigue by observing heart rate and breathing, while the MIT Media Lab has successfully implemented a project called AutoEmotive, which identifies stress and emotion of drivers using a heart monitoring system, face recognition, and eye tracking [2]. With development of advanced driver assistance systems, the term "cognitive car" was proposed by Heide and Henning [3], the purpose of which is to detect errors and prevent accidents by identifying the cognitive status of drivers [4], [5]. Conventionally, the mental status of drivers has been inferred by the eye movement analysis, since drivers obtain most of their information through vision [6]. For example, Windridge *et al.* [7] classified driver intentions based on human behavior-related variables including eye-gaze location, Kujala *et al*. [8] measured event density by occlusion distance, and Yamani *et al*. [9] predicted the type of tasks by a sequential glance distribution. Driver distraction was also measured by observing glance sequences and gaze patterns [10]–[12].

Recognizing the importance of a driver's internal status [13], [14], researchers have adopted data mining techniques to analyze drivers' behaviors, using Markov-chain models [15] and classification algorithms such as support vector machines [7], [16]. This data-driven approach explains a driver's behavior in a specific application by evaluating different parameters, but provides a limited understanding of the underlying cognitive process. In contrast, computational models such as SEEV (salience, effort, expectancy, and value) [17], [18], adaptive control of thought-rational (ACT-R) [19], [20], and the queueing network–model human processor (QN-MHP) [21]–[25] explain a driver's scanning behavior and driving performance based on cognitive architectures. Existing computational models have focused on the impact of visual stimuli and environments that are directly related to the driving task, but have paid little attention to the internal status or task goal, which also affect a driver's scanning behavior [24].

Levels of processing (LOP) classifies the depth of mental processing based on semantic properties that are required for a stimulus to be processed [26]. This concept is important in the human–machine interaction because it is closely related to the cognitive workload and sequential behaviors. Therefore, this paper utilizes the concept of the LOP and proposes a computational model based on QN-MHP to explain the LOP in visual search while driving.

First, driver's internal goals are classified by applying the concepts of the perceptual and cognitive LOP and analyzing their influence based on the reaction time (RT) and eye movement characteristics. We report on an experimental study carried out to observe human performance at different LOPs, which is followed by a description of computational models developed for the comprehensive understanding of a driver's cognitive processes.

Manuscript received January 28, 2017; revised July 26, 2017, January 5, 2018, and May 12, 2018; accepted September 3, 2018. Date of publication October 31, 2018; date of current version November 21, 2019. This work was supported by the Korean Federation of Science and Technology Societies under Grant NRF - 2012R1A1A30110320. This paper was recommended by Associate Editor C. Wu. *(Corresponding author: Ji Hyoun Lim.)*

Y. L. Rhie was with the Department of Industrial Engineering, Seoul National University, Seoul 08826, Korea. She is now with the Agency for Defense Development, 305-152 KR Daejeon, Korea (e-mail: bsinitsaz@gmail.com).

J. H. Lim is Human Factors Engineer in San Francisco/Bay area, CA, USA (e-mail: smilelim@gmail.com).

M. H. Yun is with the Department of Industrial Engineering, Seoul National University, Seoul 08826, Korea (e-mail: mhy@snu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2874183

II. Background

A. LOP

A person's internal goal influences the encoding and cognitive processes of visual information [27] as the person typically searches for the useful information to achieve a goal. The concept of the LOP, suggested by Craik and Lockhart [26], defines the depth of processing from initial sensory analysis to deeper semantic properties [28]. They found that an individual conducts "shallow" sensory processing, "intermediate" conceptual processing, and "deep" semantic processing depending on the purpose for which the given information will be used.

This distinction between the three stages was proven by memory performance: deeper LOP improves the degree of elaboration and memory duration [26]. The difference between LOPs was also shown in the field of neuroscience, as different neural activities were observed for perceptual and cognitive LOPs [29], [30]. In particular, Vanrullen and Thorpe [31] revealed temporal differences in the event-related potentials, and quantitatively proved that high-level semantic properties were extracted faster than low-level properties. The stages of LOP do not interact with each other in a linear manner, in that individuals sometimes apply higher LOPs before going through the lower level conceptual stages [32], [33]. However, owing to the clear differences between perceptual and cognitive processes, further processes within the sensory level cannot improve the memory performance [34].

For driving contexts, researchers have distinguished tasks involving simple perception from tasks involving higher LOPs [25], [35], [36]. Greenstein and Rouse [35] defined a driving task with "monitoring" and "decision making," Salvucci [20] classified driving techniques into "control," "monitoring," and "decision making" following Michon's [37] hierarchical control structure, and Wickens [38] defined eight driving tasks based on information access effort. Understanding the process of perception, decision, and action of drivers has been an important subject for the cognitive car [13].

Traditionally, researchers have investigated LOP effects on different types of tasks; an explicit memory task involves a conscious recollection of prior knowledge, while an implicit memory task only focuses on the physical features of stimuli [39]. They empirically proved that LOP has a strong influence on explicit memory tasks, small but significant influence on cognitive implicit tasks (implicit memory tasks involving conceptual processing), and no influence on perceptual implicit tasks (implicit memory tasks without conceptual processing) [40]–[44]. Considering there is a clear difference between an explicit memory task and implicit memory task, we assume that the LOP can be observed for these tasks. We classify LOPs as "perceptual" or "cognitive"; a perceptual level-task involves no conceptual processing but does involve physical processing; therefore, it corresponds to the shallow level of the LOP. Meanwhile, a cognitive level-task involves decision making based on prior knowledge, indicating a deeper LOP.

B. Visual Information Processing

Many studies have focused on bottom–up and top–down processes for visual searching [45]–[47]. In the bottom–up approach, eye movement is composed of eye fixation and saccadic movements. The retina analyzes the fine details in the fovea and limited properties in the parafovea during eye fixation. The information from the parafoveal area is used to find the next moving location [24]. A conscious process on the exact eye location is called overt attention, and the attention preceding an eye movement is defined as covert attention [48]–[50]. Because of this bottom–up mechanism, the number of a driver's eye fixations increases and eye blink frequency decreases when visual complexity is high [51], [52].

Several studies also have revealed the effects of the top–down processing. Yarbus [27] observed that a human intentionally looks at the location that contains the necessary information, and Myers and Gray [53] introduced visual scan adaptation, whereby the number of fixations decreases and scan pattern becomes similar because of the learning effect. Neisser [54] defined perception as a cyclic process in that an individual's internal scheme directs where he/she looks, and the information subsequently modifies the internal scheme. One illustration of the top–down processing for drivers is seen by a higher gaze concentration for a higher cognitive workload [55], [56]. Overall, the internal LOP process can be identified by analyzing eye movement characteristics, which is closely associated with visual stimuli and cognitive processes.

C. Queueing-Network-Based Cognitive Architecture

The cognitive model aims to elucidate human behaviors based on a cognitive architecture composed of the cognitive units that work toward a coherent result [57]. To understand the cognitive process, the "symbolist" model was suggested to represent the components and their interconnected arrangements in a human. However, symbolist models such as the model human processor (MHP) [58], ACT-R [19], and executive-process/interactive control (EPIC) [59] lack mathematical frameworks that represent the overall architecture. Meanwhile, "mathematical" models, such as the information transmission process [21], focus on presenting algebraic expressions on specific mechanisms, but do not provide a comprehensive understanding for the process of perception, cognition, and action. In contrast, QN-MHP [22] bridges the gap between these approaches because it directly represents information processing using the mathematical framework of the queueing network theory [20], [22], [60].

Using symbolic and mathematical models, the computational approach produces practical models that imitate human performance. The representative computational cognitive models are ACT-R [19], EPIC [59], and the queueing network model [21], which has been further extended as the QN-MHP [22]. To implement a computational model, appropriate production rules or procedural models are defined for each subgoal (i.e., monitoring lanes and determining lane changes [20]), and eventually are applied to the cognitive structure [22], [35], [61].

The QN-MHP uses both a production and procedural system approach, integrating studies on simulation methods for queueing networks and symbolic/procedural methods for goal, operator, method, and selection (GOMS) style descriptions of the MHP. Although EPIC and ACT-R have achieved multiple success in predicting the human performance [20], [62]–[64],

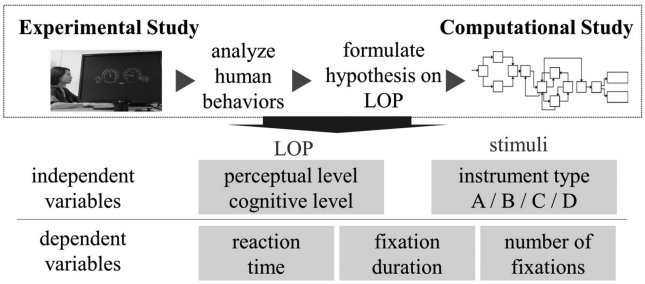


Fig. 1. Overview of the research.

QN-MHP has the advantage of requiring less complex assumptions. Whereas the models of EPIC and ACT-R are based on production rules to specify delays and interruptions on perceptual and motor processing time, QN-MHP requires less task-specific assumptions but relies on the mathematical fundamentals originated from the queueing network architecture. Wu and Liu [65] modeled the psychological refractory period [66] using QN-MHP and revealed that the method requires no task-specific assumption and the same number of free parameters as ACT-R and EPIC or less. In addition, Tsimhoni and Liu [67] emphasized the advantage of the QN-MHP approach in adding concurrent activities without limiting or predefining the order of occurrence. In a line of research, QN-MHP was used to simulate LOP effects, defining servers as the basis of the cognitive function and the depth of processing as the routes of entities.

The QN-MHP incorporates the concept of the queuing theory, which has been widely adopted in the field of operations engineering. Entities, which indicate visual stimuli, denote jobs or customers in operations research. Servers correspond to employees who provide services, and a route represents the sequence of visited servers. Based on this theoretical background, QN-MHP has proven its applicability in predicting human behaviors from simple visual search tasks [68], [69] to complex driving tasks; Tsimhoni and Liu [67] created steering models, Wu and Liu [70] measured driver's cognitive workloads for various contexts, and Lim *et al.* [23] structured a night-vision enhancement system assisting pedestrian-detection tasks. More recently, Bi *et al.* [71] predicted a driver's lateral motor control using physiological parameters.

This paper presents experimental and computational studies predicting an individual's depth of processing (perceptual and cognitive) while interacting with a car's instruments. As illustrated in Fig. 1, the first part of this paper introduces the experimental study conducted to evaluate human performances in terms of RT and eye fixation characteristics such as the number and duration of eye fixations in a controlled environment. Participants were asked to answer two questions for the same visual stimulus of a car's instruments. Based on the experimental results, we suggest hypotheses on the routes of entities to explain the differences in the LOP. According to the suggested hypotheses, we propose computational models and compare their performance with the experimental results. Through this research, we expect to understand a driver's cognitive process with respect to effects of the LOP.

III. Experiment on Driver's RT and Eye Movement

To verify the impact of the LOP on RT and eye fixation characteristics, we classified perceptual- and cognitive-level tasks and observed performance in the same experimental environment.

A. Method

1) Participants and Setup: Initial participants were 13 drivers ranging in age from 20 to 29 years. However, excluding two participants whose eye-tracking trials failed to collect tracking data more than 30% of the time, the performances of seven male and four female drivers (average age: 25.33 years) were analyzed. Subjects' RT was measured through video analysis, while eye fixation characteristics were recorded using an eye tracker (T120, Tobii, Sweden). The minimum fixation duration was set to 60 ms. The eye tracker allowed the head movement within 50–80 cm; therefore, a head positioner was not used during the experiment because the viewing distance was 60 cm.

2) Procedure: Before starting the experiment, participants were asked to provide verbal answers to two questions. The first question required subjects to perform a perceptual-level task, reading the exact values displayed on the speedometer, tachometer, fuel gauge, engine temperature gauge, and odometer, while the second required a cognitive-level task in which the participants had to decide what to do in a given situation. For example, the perceptual-level task involved answering questions such as "What is the speed of this car?" or "Read the number displayed on the odometer," while the cognitive-level task was based on questions such as "If you were in an 80 km/h speed limit zone, what would you do? Decide whether to accelerate or decelerate." Although we only used eye tracking data for the speedometer in further analysis, questions related to different instruments were also posed to prevent drivers from predicting where to look. The indicators were manipulated to display different values of the speedometer, tachometer, fuel gauge, engine temperature gauge, and odometer. In total, 192 questions consisting of 96 perceptual tasks and 96 cognitive tasks were formed to ensure the number of questions were evenly distributed over the various components of the instrument layout. Half of the questions asked for the speed of the car, and these data were actually utilized in further analysis.

To prevent eye fatigue and the learning effect, we grouped questions into four blocks and provided breaks between them. The processing level and types of instrument layouts were counterbalanced in each block. Before starting the experiment, participants were trained with one set of randomly chosen questions composed of 24 perceptual- and 24 cognitive level tasks for less than 5 min. Then, participants were asked if they felt ready to perform the tasks. During the experiment, the instruments were shown right after each question, with a black screen presented every 1.5 s. This blocking is based on ISO guidelines [72], which state that the driver's view should be visible for 1.5 s and occluded for 1.5 s. The calibration was carried out before starting another block of questions.

3) Experimental Design: There are two primary within-subject factors in the experimental design: instrument layout types and the given tasks. Four different layouts of instruments were used as stimuli, as illustrated in Fig. 2. When selecting

Fig. 2. Four instrument layouts used as stimuli.

visual stimuli, high-resolution images of car instruments were collected from 2010 concept cars. After screening the electronic instruments to control the effects of analog/digital presentation, four instrument images were selected with different features in their layouts. First, the speedometer was located at different positions: on the left, as shown in A, in the center, as shown in B, and on the right, as shown in C and D. Second, the overall layouts of A and D show four distinctly separate areas for each function, whereas B and C show three distinct areas. The size of the speedometer also varies. The speedometer has the same size as the revolution counter in A and D, whereas it occupies the largest area in B, and is as small as the fuel gauge in C.

For each stimuli type, tasks requiring two different LOPs were provided. The number of tasks for each LOP and instrument layout were evenly distributed in each block, while the sequence was randomly chosen. To observe the impact of the LOP on human performance, we assumed that the average RT, eye fixation duration, and number of eye fixations would be affected by the LOP. For the analysis, we collected data on RT and eye movement characteristics and tested their mean differences. RT was measured from the onset of the stimulus up to the moment that the participant started to voice his/her answer by analyzing sound waves from the recorded video. Fixation duration is defined as the total time eye fixation occurs, while the number of fixations was calculated by how many eye fixations were made.

B. Experimental Results

Because normal distributions were not observed on RT and eye fixation characteristics, nonparametric statistical methods were applied in analysis. As a result of the Mann–Whitney U test, we observed significant effects of the LOP on RT ($U = 59522.0$; $p < 0.001$; $r = 0.469$), fixation duration ($U = 76388.5$; $p < 0.001$; $r = 0.334$), and eye fixation characteristics ($U = 91799.5$; $p < 0.001$; $r = 0.200$). Here, r represents the rank-biserial correlation coefficient, which is used to measure the effect size of nonparametric statistics [93]. For more detailed results, we conducted a statistical test for each sample pair of RT- and eye fixation-related variables (see Tables I and II).

1) Reaction Time (RT): On average, the RTs for the perceptual-level task were 0.871 s longer than those for the cognitive-level task. As a higher cognitive LOP can be processed faster than the perceptual LOP [31], we can infer from the results that cognitive tasks require less detail in the visual information than perceptual tasks. In addition, RT differed depending on the instrument layout (see Table I). Significant differences were observed only when conducting perceptual-level tasks, for example, RT for B is longer than that of A and C ($p < 0.005$). The results also reveal that design features of instrument layout C forces drivers to conduct both perceptual- and cognitive-level tasks faster.

2) Eye Fixation Characteristics: Fixation duration is calculated as the total time eye fixation occurs. As seen in Table II, the mean fixation duration for different LOP differed significantly ($p < 0.05$); fixation duration for cognitive-level tasks is 1.050 s, while that for perceptual-level tasks is 0.730 s. In addition, the frequency of eye fixation was also significantly different for the two types of tasks ($p < 0.05$). There were 5.036 fixations for perceptual-level tasks and 4.139 for cognitive-level tasks. The eye movements were also affected by the layout of the instruments, inferring significant differences in the number and duration of eye fixations for different types of stimuli, as presented in Table II. Instrument layout B has a lower fixation duration and higher number of fixations for both perceptual- and cognitive-level tasks, whereas there are almost no significant differences for instrument layouts A, C, and D. Because shorter fixation durations and lower fixation counts indicate, respectively, the difficulty of the tasks at a semantic level [73] and efficiency of the visual search [74], [75], this result demonstrates that instrument layout B supports cognitive processing but is not suitable for an efficient visual search. This is consistent with the results for the RT, in that RT was significantly longer for instrument layout B for perceptual-level tasks but not for cognitive-level tasks.

IV. Simulations on RT and Eye Movement Using the Computational Cognitive Model

From our experimental study, we observed different eye fixation characteristics for perceptual- and cognitive-level tasks. Because researchers have developed computational models that can be embedded in the actual system such as adaptive workload management systems [25] and night vision enhancement systems [23], we expect that the integration of experimental results with computational models will increase the potential for practical application.

A. Method

1) Modeling Perceptual and Cognitive LOP: QN-MHP is a simulation model represented as a queueing network structure. In this study, the model was structured using Arena 14.0. Entities in the QN-MHP are assumed to flow through a visual system according to a Poisson distribution until the human achieves the set tasks [21]. Once the entities have obtained the required information, they proceed to servers according to an exponential distribution with first-come, first-serve scheduling, thus incurring a time delay if the server is busy. The processing time in the perceptual subnetwork follows the distribution of ppt $\sim$ E(0.042, 0.025) with a capacity of 4, while the cognitive processing time distribution is cpt $\sim$ E(0.018, 0.006) with an infinite capacity. Motor processing time follows the mpt distribution $\sim$ E(0.024, 0.010), with capacity of 1. Foveal vision is set to two degrees of the visual field, while the field of the parafovea is extended from

TABLE I
RESULTS OF THE MANN–WHITNEY TEST ON THE RT FOR DIFFERENT INSTRUMENT LAYOUTS AND LOPS

		Perceptual				Cognitive				U	r
		A	B	C	D	A	B	C	D		
A	U	2.467 (1.382)				1.671 (0.852)				3892.0**	0.409
	r										
B	U	7512.5*	2.700 (1.120)			5172.0	1.701 (0.925)			2918.5**	0.588
	r	0.154				0.015					
C	U	8509.5	5948.0*	2.396 (1.002)		4797.5	6275.0	1.495 (0.674)		3087.5**	0.556
	r	0.017	0.153			0.094	0.121				
D	U	8553.5	6268.0	6902.5	2.530 (1.371)	5259.5	6872.5	6463.5	1.739 (1.046)	4572.5**	0.354
	r	0.029	0.122	0.008		0.007	0.037	0.072			

Note: Entries on the diagonal indicate mean values with standard deviation given in parentheses. $^{*}p$-value < 0.05, $^{**}p$-value < 0.001.

TABLE II
RESULTS OF THE MANN–WHITNEY TEST COMPARING THE NUMBER OF EYE FIXATIONS AND FIXATION DURATION FOR DIFFERENT INSTRUMENT LAYOUTS AND LOPS

			Perceptual				Cognitive				U	r
			A	B	C	D	A	B	C	D		
Fixation duration	A	U	1.081 (0.644)				0.718 (0.307)				4481.0**	0.324
		r										
	B	U	6226.0**	0.762 (0.403)			3395.0**	0.548 (0.284)			4288.0**	0.404
		r	0.304				0.364					
	C	U	8320.0	4582.0**	1.134 (0.643)		4832.0	5671.0**	0.660 (0.389)		3649.5**	0.485
		r	0.061	0.358			0.088	0.206				
	D	U	7402.0*	3706.0**	6345.0	1.223 (0.698)	4008.0**	3521.0**	4508.0**	0.997 (0.669)	5235.0**	0.235
		r	0.143	0.468	0.081		0.237	0.503	0.358			
Number of fixations	A	U	4.711 (2.762)				3.876 (2.151)				5137.0**	0.225
		r										
	B	U	5596.0**	6.333 (2.797)			3446.0**	5.175 (2.519)			5299.5**	0.264
		r	0.374				0.355					
	C	U	8462.0	4722.5**	4.722 (2.260)		4939.5	4968.0**	3.916 (1.790)		5564.0**	0.214
		r	0.046	0.339			0.051	0.292				
	D	U	7972.0	4120.0**	6133.0	4.431 (2.587)	4758.0	4101.0**	5944.5*	3.509 (1.943)	5460.0**	0.202
		r	0.078	0.408	0.111		0.094	0.421	0.139			

Note: Entries on the diagonal indicate mean values with standard deviation given in parentheses. $^{*}p$-value < 0.05, $^{**}p$-value < 0.001.

6° to 14° in increments of 4° because the average parafoveal area is 10°.

2) Simulation Replication Time: When a stimulus is provided, a human first encodes the saliency across the stimuli based on a bottom–up process, and then, selects where to focus based on a top–down process volitionally [24], [76]–[79]. As a consequence of this interactive mechanism, a human fixates on a single winning location. While preattentive selection for visual search tasks has been studied in the context of visual search [80]–[82], it is not the aim of our research to investigate the mechanism of first-time fixation. Therefore, we empirically evaluated the probability that first-time fixation is in the area of interest (AOI) and applied these results in the simulation. The method for incorporating the probabilities of the experimental results in the computational model is not new [35], [62]. For instrument layouts A–D, participants first fixated on the AOI with probabilities 73%, 86%, 65%, and 72%, respectively. For the case where first-time fixation remains in the AOI, the basic number of iterations was set to 1000 times, while the simulation iterated 370, 170, 540, and 400 times, respectively, for the other cases.

Our model assumes a hunt-feature rule whereby subjects search targets using one of the selected features [19], [23]. If the goal is associated with a certain feature and there is an object with that feature within the preattended visual information, the eye moves to the object. Meanwhile, head movement was not considered, as tasks only required small eye shifts [83], [84].

3) Modeling Perceptual- and Cognitive-Level Tasks: The perceptual- and cognitive-level tasks are analyzed using GOMS-style hierarchical tasks (see Fig. 3), which is the task analysis technique adopted in the QN-MHP. In the model, rules based on a top–down process are assumed, since we observed sufficiently trained drivers. First, if participants did not focus on the AOI at their initial fixation, we assumed that the next fixation would be located in the AOI. Second, we assumed that the end point of the needle (needle point) is a primary hunting feature; therefore, the eyes move to the tip of a needle when located in a parafoveal

<Method to accomplish a goal of recognizing the speed>

Step 1. GLANCE a stimulus
Step 2. Retrieve information from the stimulus.
Step 3. Pose attention window (AOI)
Step 4. Select point within AOI.
Step 5. Decide: If Focal information is relevant with 'Needle Point' or 'Marker', go to step 11
Else: go to step 6
Step 6. If Parafoveal information is relevant with Hunt feature 'Needle Point' or 'Marker', go to step 12
Else: go to step 7
Step 7. If Primary Hunt feature is relevant with Hunt feature 'Location', go to step 9
Else: go to step 8
Step 8. Loop until Find Hunt feature 'Location', go to step 5
Step 9. Decide: If focal information is relevant with Hunt feature 'Needle', go to step 10
Else: Loop until find Hunt feature 'Needle', go to step 5
Step 10. Loop until Find Hunt feature 'Needle Point', go to step 5
Step 11. If read both side of Markers and Needle Point, go to step 13
Else: go to step 6
Step 12. Randomly Move eye to one of Markers and Needle Point among Parafoveal information (not have been read), go to step 5
Step 13. Retrieve the 'Goal selection' and related 'Goal procedure', go to step 5
Step 14. Calculate distance between cursor and markers
Step 15. Report goal accomplished

<Method to accomplish a goal of decision making on the speed>

Step 1. GLANCE a stimulus
Step 2. Retrieve information from the stimulus.
Step 3. Pose attention window (AOI)
Step 4. Select point within AOI.
Step 5. Decide: If Focal information is relevant with 'Needle Point', go to step 11
Else: go to step 6
Step 6. If Parafoveal information is relevant with Hunt feature 'Needle Point', go to step 10
Else: go to step 7
Step 7. If Primary Hunt feature is relevant with Hunt feature 'Location', go to step 9
Else: go to step 8
Step 8. Loop until Find Hunt feature 'Location', go to step 5
Step 9. Decide: If focal information is relevant with Hunt feature 'Needle', go to step 10
Else: Loop until find Hunt feature 'Needle', go to step 5
Step 10. Loop until Find Hunt feature 'Needle Point', go to step 5
Step 11. If Focal information is relevant to 'Needle Point', and Parafoveal information is relevant to 'Marker', go to step 12
Step 12. Retrieve the 'Goal selection' and related 'Goal procedure', go to step 13
Step 13. Compare focal information and 'Goal procedure'
Step 14. Report goal accomplished

Fig. 3. Pseudo-code for perceptual- and cognitive-level tasks in the QN-MHP.

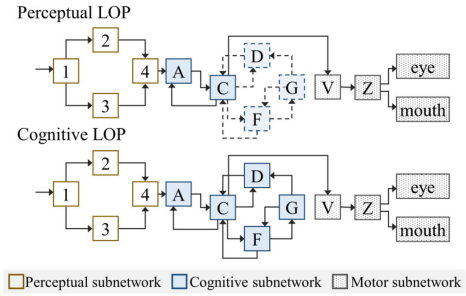


Fig. 4. Structure of the queueing network cognitive architecture. Note: 1: common visual processing; 2: visual recognition; 3: visual location; 4: visual recognition and location integration; A: visuospatial sketchpad; C: central executive; D: long-term procedural memory; F: complex cognitive function; G: goal initiation; and V: sensory-motor integration.

area. Third, the shape of the needle is also a hunting feature, and therefore, participants look at the needle when moving their eyes into the AOI. In addition, we assumed that a perceptual-level task requires fine needle-point images and nearby markers in the focal area, whereas a cognitive-level task only requires robust images of markers based on the experimental results.

From the task analysis, servers D, G, and F were excluded for perceptual-level tasks (see Fig. 4) because this type of task does not require a higher LOP to arrive at a determination. Server D retrieves the task list of the current primary goal from the long-term working memory, while server G sets the current primary goal while focusing attention on the goals related to any event. Server F usually supports higher level cognitive tasks, including mathematical computation and image manipulation, which is closely associated with server D [22], [85].

In structuring the model, instrument images were first coded into 13×16 data arrays to convey the location, shape, and edge values of the stimulus. The size of one block was determined by the attended vision area (2°). When an entity received information, it moved throughout the QN-MHP servers. The processing logics of entities were stated in the form of the If–Then rules (see Fig. 3), and the processing time of servers followed exponential distributions. One simulation session terminated when the eye located to the predesignated coordinates. The termination time was recorded as the RT, and the values of location parameters were recorded as eye movements.

TABLE III
MEAN RT RESULTS WITH STANDARD DEVIATION FOR SIMULATIONS (6°, 10°, 14°) AND EXPERIMENT

		6°	10°	14°	experiment
A	perceptual	2.157 (1.750)	2.437 (1.873)	2.289 (1.996)	2.467 (1.382)
	cognitive	1.959 (1.793)	2.110 (2.010)	2.113 (1.932)	1.671 (0.852)
B	perceptual	2.664 (1.909)	2.916 (1.924)	2.944 (1.942)	2.700 (1.120)
	cognitive	1.609 (1.535)	1.595 (0.562)	1.346 (1.159)	1.701 (0.925)
C	perceptual	2.257 (1.975)	2.159 (1.828)	2.271 (1.916)	2.396 (1.002)
	cognitive	2.096 (1.942)	2.151 (2.080)	2.113 (1.960)	1.495 (0.674)
D	perceptual	2.405 (1.765)	2.352 (1.713)	2.320 (1.697)	2.530 (1.371)
	cognitive	1.780 (1.765)	1.895 (1.815)	1.779 (1.668)	1.739 (1.046)

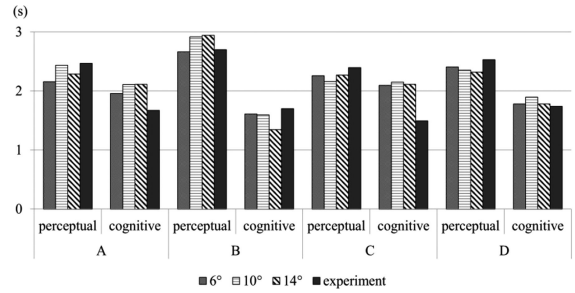


Fig. 5. Mean RT values for simulations (6°, 10°, 14°) and experiment.

B. Results

A computational cognitive model was developed to simulate the RT and eye fixation characteristics for each LOP. The simulation and experimental results are compared.

1) Reaction Time (RT): The mean RT values obtained from simulation and human experiment are shown in Table III and Fig. 5. Although the mean values are mostly similar, there were

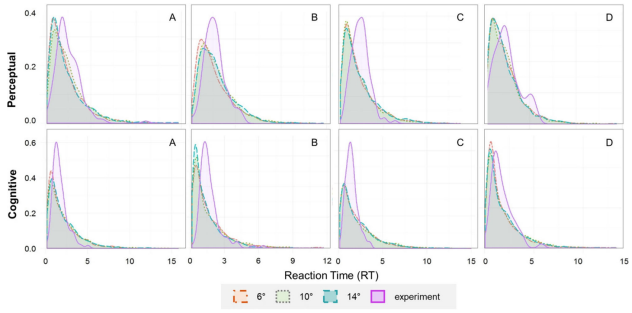


Fig. 6. RT density distributions for simulations (6°, 10°, 14°) and experiment.

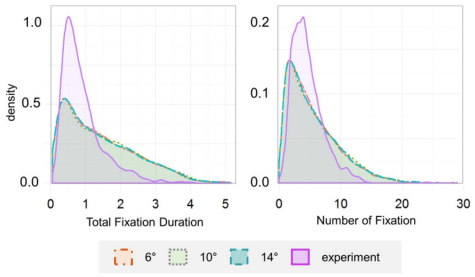


Fig. 7. Density distributions of mean fixation duration and number of fixations for simulations and experiment.

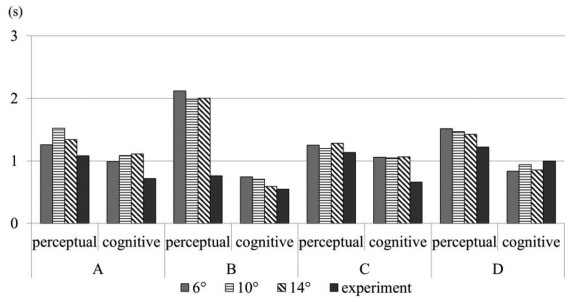


Fig. 8. Mean values of total fixation duration.

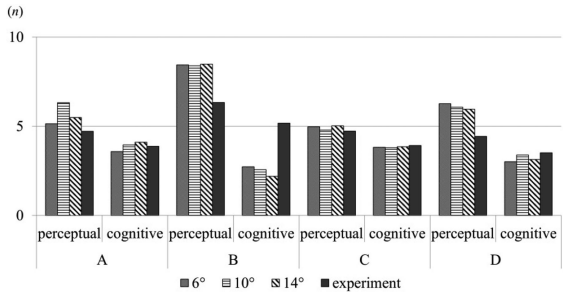


Fig. 9. Mean number of eye fixations.

significant differences between the simulation and experimental results ($p < 0.001$). For perceptual- and cognitive-level tasks, the simulation results predicted RT values, respectively, 0.112 s higher and 0.249 s lower than the experimental results. These results are due to the differences in the distributions for the simulation and experiment; as shown in Fig. 6, the simulation RT values have longer tails than those of the experiment. In addition, contrary to our assumption, no effect of the parafoveal area was identifiable. This is due to the top–down processes used in the simulation. That is, the If–Then rules adopted in the simulation model may have minimized the effect of the bottom–up process.

Despite these limitations, Fig. 5 shows that the simulation results are mostly similar to the experimental ones. Similar to the findings of the experimental study, instrument layout B yields the longest RT of all the instrument layouts, while A and C show relatively lower RT values. Moreover, the experimental study reveals higher RTs (by 0.510 s) for the perceptual LOP than the cognitive LOP, while the corresponding simulation RTs were 0.871 s higher. In summary, we were able to extract the RT trends of actual humans from the simulation results depending on the instrument layout types and LOP.

2) Eye Fixation Characteristics: Eye fixation was defined as a focus of active interest that lasts at least 60 ms. The results of the experimental study suggest a longer duration and higher frequency of fixations when conducting a perceptual-level task ($p < 0.05$). As illustrated in Fig. 7, the distributions of the fixation frequency and duration obtained from the QN-MHP have longer tails than those for human subjects. This phenomenon may be because all participants in the experiment were in their 20 s, whereas the innate logic in the QN-MHP assumes a wide range of ages. However, age influences eye fixation characteristics such that older adults are known to exhibit higher fixation durations for some visual search tasks [86]. To reduce the effect of the longer tail, we eliminated outliers in the top 10% of the simulation data.

The simulations show fixation durations for the perceptual LOP that were longer than those of the cognitive LOP from 0.150 to 1.413 s ($p < 0.05$), corresponding with those of the experimental results ranging from 0.214 to 0.474 s. The mean values of total eye fixation duration in Fig. 8 represent some marked differences between the simulation and experimental results. For the perceptual LOP, the simulation results yielded much higher values for instrument layout B, whereas the experimental study yielded the shortest duration.

Fig. 9 depicts the number of fixations, showing a similar trend for the simulation and experimental results except for the instrument layout B. The mean values and standard deviations of the duration and number of fixations are listed in Table IV.

V. Discussion and Conclusion

This paper presented an experimental and simulation study for a visual search task on car instruments to verify whether different human performances are guided by different LOPs. Overall, the simulation model showed results consistent with those of the experiment; the RT for cognitive-level tasks was shorter than that for perceptual-level tasks. Moreover, the eyes fixated for a shorter duration and with lower frequency.

Such results correspond with those of classic studies on the visual search in terms of the relationship between eye fixation characteristics and cognitive processes. The visual routine has been defined [87], confirming that a human uses different eye movement strategies in a top–down process. If a task requires a higher level of memory performance, total fixation duration increases [88], [89]. More recently, Grill-Spector and Kanwisher [90] revealed that a longer fixation duration is necessary to

TABLE IV
MEANS AND STANDARD DEVIATIONS OF FIXATION DURATION AND NUMBER OF FIXATIONS

		Perceptual				Cognitive			
		6°	10°	14°	exp	6°	10°	14°	exp
Fixation duration	A	1.257 (0.978)	1.521 (1.047)	1.342 (1.054)	1.081 (0.644)	0.989 (0.984)	1.083 (1.041)	1.112 (1.013)	0.718 (0.307)
	B	2.120 (1.205)	1.983 (1.159)	2.002 (1.191)	0.762 (0.403)	0.744 (0.802)	0.709 (0796)	0.589 (0.663)	0.548 (0.284)
	C	1.252 (1.007)	1.200 (0.975)	1.282 (1.033)	1.134 (0.643)	1.058 (1.019)	1.049 (1.075)	1.066 (1.046)	0.660 (0.389)
	D	1.516 (1.002)	1.468 (0.985)	1.426 (0.969)	1.223 (0.698)	0.835 (0.877)	0.943 (0.923)	0.854 (0.911)	0.997 (0.669)
Number of fixations	A	5.135 (3.952)	6.315 (4.397)	5.495 (4.249)	4.711 (2.762)	3.577 (3.692)	3.952 (3.981)	4.104 (3.834)	3.876 (2.151)
	B	8.429 (4.900)	8.394 (4.799)	8.483 (5.131)	6.333 (2.797)	2.729 (3.068)	2.570 (2.867)	2.186 (2.531)	5.175 (2.519)
	C	4.970 (3.928)	4.780 (3.951)	5.022 (4.005)	4.722 (2.260)	3.818 (3.803)	3.791 (3.972)	3.850 (3.894)	3.916 (1.790)
	D	6.260 (4.118)	6.071 (4.050)	5.961 (3.928)	4.431 (2.587)	3.011 (3.222)	3.380 (3.480)	3.132 (3.523)	3.508 (1.943)

ascertain finer details of an image, as shown when conducting perceptual-level tasks. The results of this paper are meaningful in that similar results with a computational model were demonstrated. The findings from this paper are specified as follows.

First, we examined three different field-of-view degrees of the parafoveal area in the QN-MHP simulation modeling: 6°, 10°, and 14°. We hypothesized that a wider parafoveal area would be advantageous in identifying hunt features resulting in decreased RT. However, our results showed no dependence on the degree of the parafoveal area. This means that the goal-driven top–down process is the dominating influence on the eye movement, which is demonstrated as the production rules in the simulation. However, note that this result is limited to our experimental condition, as the effect of the parafoveal area would have been diminished because there were few distractors and the image size was small.

Second, we could represent the mental status of drivers by manipulating the route of the entities. Although QN-MHP studies have either selected relevant servers or proposed parameters for a specific task [91], no study has manipulated routes to compare their differences. This paper adopted different routes for each LOP, and subsequently, demonstrated similar trends to those observed in the experimental results. Although there were significant differences between the distributions of experiments and simulation, similar trends were identified for eye movements in association with the stimuli characteristics. The number of fixations and fixation duration for the instrument layout B increased in both the experimental and simulation results. These results may have been affected by the simulation settings because we set the minimum unit of the eye movement to 2°, which means that small saccades of the human eye cannot be reproduced. In addition, the color of the needles may have affected this result. The RT results of the simulations and experiment showed the biggest differences for the cognitive-level tasks on instrument layout types A and C, whose needles were red, while the others were white. Considering that visibility (saliency) influences visual search behavior, we assume that participants perceived the red color as a target feature, thereby reducing the RT.

Third, we examined the influence of perceptual- and cognitive-level tasks in a controlled environment. If a driver's LOP could be identified in real time through eye movements, it will be possible to provide appropriate interfaces to support the driver's perception or decision making; for example, a phonic instruction or a digital numeric value could be additionally provided for perceptual-level tasks. As such, this study enables a driver's temporary goals to be perceived to determine which service should be provided in a driver assistance system. However, it is necessary to implement a simulation model that considers the head movement of drivers in further research, since, in addition to eye movement characteristics, this directly affects the RT. Moreover, the environmental variables of real driving contexts should also be considered because driving was not carried out as the primary task in this study to eliminate the effect of the environmental factors.

In summary, in a series of experiments requiring both top–down and bottom–up processes to achieve a goal [92], we examined RT and eye fixation characteristics for two different levels of tasks through experimentation, and explored the mechanism using a QN-MHP simulation model. Although this is not the first time a computational model involving this approach has been constructed [19], [21], [59], no studies have observed the LOP by manipulating an entity's routes to infer the cognitive process. In terms of application, this study revealed the possibility of using human-generated data to evaluate the innate purposes of drivers. Based on the identified LOP, subsidiary information could be provided on a head-up display, or the cognitive workload of drivers could be calculated to prevent accidents. Rather than calculating the internal states for specific contexts, the LOP approach could provide simpler and more robust services. In future research, a more practical application method should be considered in an actual driving environment.

REFERENCES

[1] Ford Motor Company, "Ford develops heart rate monitoring seat; adds new element to company's in-car health and wellness research portfolio," in *PRNewswire*, 2011. [Online]. Available: http://www.prnewswire.com/news-releases/ford-develops-heart-rate-monitoring-seat-adds-new-element-to-companys-in-car-health-and-wellness-research-portfolio-122483828.html. Accessed: Feb. 20, 2015.

[2] J. Hernandez, D. McDuff, X. Benavides, J. Amores, P. Maes, and R. Picard, "AutoEmotive: Bringing empathy to the driving experience to manage stress," in *Proc. Companion Publication Designing Interactive Syst.*, 2014, pp. 53–56.

[3] A. Heide and K. Henning, "The "cognitive car": A roadmap for research issues in the automotive sector," *Annu. Rev. Control*, vol. 30, pp. 197–203, 2006.

[4] C. M. Wickens, M. E. Toplak, and D. L. Wiesenthal, "Cognitive failures as predictors of driving errors, lapses, and violations," *Accident Anal. Prevention*, vol. 40, pp. 1223–1233, 2008.

[5] T. Inagaki, "Smart collaboration between humans and machines based on mutual understanding," *Annu. Rev. Control*, vol. 32, pp. 253–261, 2008.

[6] M. Sivak, "The information that drivers use: Is it indeed 90% visual?" *Perception*, vol. 25, pp. 1081–1089, 1996.

[7] D. Windridge, A. Shaukat, and E. Hollnagel, "Characterizing driver intention via hierarchical perception–action modeling," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 1, pp. 17–31, Jan. 2013.

[8] T. Kujala, J. Mäkelä, I. Kotilainen, and T. Tokkonen, "The attentional demand of automobile driving revisited occlusion distance as a function of task-relevant event density in realistic driving scenarios," *Human Factors*, vol. 58, pp. 163–180, 2016.

[9] Y. Yamani, W. J. Horrey, Y. Liang, and D. L. Fisher, "Sequential in-vehicle glance distributions an alternative approach for analyzing glance data," *Human Factors*, vol. 57, pp. 567–572, 2015.

[10] B. Metz, N. Schömig, and H. P. Krüger, "Attention during visual secondary tasks in driving: Adaptation to the demands of the driving task," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 14, pp. 369–380, 2011.

[11] *Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices*, Dept. of Transportation (DOT), National Highway Traffic Safety Administration, Washington, DC, USA, 2012.

[12] Y. Zhang, D. B. Kaber, M. Rogers, Y. Liang, and S. Gangakhedkar, "The effects of visual and cognitive distractions on operational and tactical driving behaviors," *Human Factors*, vol. 56, pp. 592–604, 2013.

[13] L. Li, D. Wen, N.-N. Zheng, and L.-C. Shen, "Cognitive cars: A new frontier for ADAS research," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 395–407, Mar. 2012.

[14] B. Kapitaniak, M. Walczak, M. Kosobudzki, Z. Jozwiak, and A. Bortkiewicz, "Application of eye-tracking in the testing of drivers: A review of research," *Int. J. Occupat. Med. Environ. Health*, vol. 28, pp. 941–954, 2015.

[15] A. Pentland and A. Liu, "Modeling and prediction of human behavior," *Neural Comput.*, vol. 11, pp. 229–242, 1999.

[16] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 340–350, Jun. 2007.

[17] C. D. Wickens, J. Goh, J. Helleberg, W. J. Horrey, and D. A. Talleur, "Attentional models of multitask pilot performance using advanced display technology," *Human Factors*, vol. 45, pp. 360–380, 2003.

[18] W. J. Horrey, C. D. Wickens, and K. P. Consalus, "Modeling drivers' visual attention allocation while interacting with in-vehicle technologies," *J. Exp. Psychol. Appl.*, vol. 12, pp. 67–78, 2006.

[19] J. R. Anderson, M. Matessa, and C. Lebiere, "ACT-R: A theory of higher level cognition and its relation to visual attention," *Human-Comput. Interact.*, vol. 12, pp. 439–462, 1997.

[20] D. D. Salvucci, "Modeling driver behavior in a cognitive architecture," *Human Factors*, vol. 48, pp. 362–380, 2006.

[21] Y. Liu, "Queueing network modeling of elementary mental processes," *Psychological Rev.*, vol. 103, pp. 116–136, 1996.

[22] Y. Liu, R. Feyen, and O. Tsimhoni, "Queueing network-model human processor (QN-MHP): A computational architecture for multitask performance in human-machine systems," *ACM Trans. Comput.-Human Interact.*, vol. 13, pp. 37–70, 2006.

[23] J. H. Lim, Y. Liu, and O. Tsimhoni, "Investigation of driver performance with night-vision and pedestrian-detection systems—Part 2: Queuing network human performance modeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 765–772, Dec. 2010.

[24] J. H. Lim and Y. Liu, "Modeling the influences of cyclic top-down and bottom-up processes for reinforcement learning in eye movements," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 4, pp. 706–714, Jul. 2009.

[25] C. Wu, O. Tsimhoni, and Y. Liu, "Development of an adaptive workload management system using the queueing network-model human processor (QN-MHP)," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 463–475, Sep. 2008.

[26] F. I. Craik and R. S. Lockhart, "Levels of processing: A framework for memory research," *J. Verbal Learn. Verbal Behav.*, vol. 11, pp. 671–684, 1972.

[27] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye Movements and Vision*. Boston, MA, USA: Springer, 1967, ch. 7, pp. 171–211.

[28] F. I. Craik and E. Tulving, "Depth of processing and the retention of words in episodic memory," *J. Exp. Psychol., Gen.*, vol. 104, pp. 268–294, 1975.

[29] M. Seeck *et al.*, "Evidence for rapid face recognition from human scalp and intracranial electrodes," *Neuroreport*, vol. 8, pp. 2749–2754, 1997.

[30] Y. Mouchetant-Rostaing, M. H. Giard, S. Bentin, P. E. Aguera, and J. Pernier, "Neurophysiological correlates of face gender processing in humans," *Eur. J. Neurosci.*, vol. 12, pp. 303–310, 2000.

[31] R. Vanrullen and S. J. Thorpe, "The time course of visual processing: From early perception to decision-making," *J. Cogn. Neurosci.*, vol. 13, pp. 454–461, 2001.

[32] R. S. Lockhart, F. I. Craik, and L. Jacoby, "Depth of processing, recognition and recall," in *Recall and Recognition*. Oxford, U.K.: Wiley, 1976.

[33] B. M. Velichkovsky, "Heterarchy of cognition: The depths and the highs of a framework for memory research," *Memory*, vol. 10, pp. 405–419, 2002.

[34] F. I. Craik and M. J. Watkins, "The role of rehearsal in short-term memory," *J. Verbal Learn. Verbal Behav.*, vol. 12, pp. 599–607, 1973.

[35] J. S. Greenstein and W. B. Rouse, "A model of human decisionmaking in multiple process monitoring situations," *IEEE Trans. Syst., Man, Cybern.*, vol. 12, no. 2, pp. 182–193, Mar. 1982.

[36] J. H. Kim, J. H. Lim, C. I. Jo, and K. Kim, "Utilization of visual information perception characteristics to improve classification accuracy of driver's visual search intention for intelligent vehicle," *Int. J. Human-Comput. Interact.*, vol. 31, pp. 717–729, 2015.

[37] J. A. Michon, "A critical view of driver behavior models: What do we know, what should we do?" *Human Behav. Traffic Safety*, L. Evans and R. Schwing Eds., Human behavior and traffic safety, New York: Plenum press, 1984, pp. 485–525.

[38] C. D. Wickens, "Effort in human factors performance and decision making," *Human Factors*, vol. 56, pp. 1329–1336, 2014.

[39] P. Graf and D. L. Schacter, "Implicit and explicit memory for new associations in normal and amnesic subjects," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 11, pp. 501–518, 1985.

[40] B. R. Newell and S. Andrews, "Levels of processing effects on implicit and explicit memory tasks: Using question position to investigate the lexical-processing hypothesis," *Exp. Psychol.*, vol. 51, pp. 132–144, 2004.

[41] D. Prabu and E. Hirshman, "Dual-mode presentation and its effect on implicit and explicit memory," *Amer. J. Psychol.*, vol. 111, pp. 77–87, 1998.

[42] D. L. Schacter and P. Graf, "Effects of elaborative processing on implicit and explicit memory for new associations," *J. Exp. Psychol. Learn., Memory, Cognition*, vol. 12, pp. 432–444, 1986.

[43] M. Jelicic and B. Bonke, "Level of processing affects performance on explicit and implicit memory tasks," *Perceptual Motor Skills*, vol. 72, pp. 1263–1266, 1991.

[44] S. B. Hamann, "Level-of-processing effects in conceptually driven implicit tasks," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 16, pp. 970–977, 1990.

[45] J. Theeuwes, P. Atchley, and A. F. Kramer, "On the time course of top-down and bottom-up control of visual attention," in *Control of Cognitive Processes: Attention and Performance XVIII*. Cambridge, MA, USA: MIT Press, 2000, pp. 105–124.

[46] J. M. Wolfe, S. J. Butcher, C. Lee, and M. Hyle, "Changing your mind: on the contributions of top-down and bottom-up guidance in visual search for feature singletons," *J. Exp. Psychol., Human Perception Perform.*, vol. 29, pp. 483–502, 2003.

[47] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 2000.

[48] H. Deubel and W. X. Schneider, "Saccade target selection and object recognition: Evidence for a common attentional mechanism," *Vis. Res.*, vol. 36, pp. 1827–1837, 1996.

[49] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception Psychophys.*, vol. 57, pp. 787–795, 1995.

[50] E. Kowler, E. Anderson, B. Dosher, and E. Blaser, "The role of attention in the programming of saccades," *Vis. Res.*, vol. 35, pp. 1897–1916, 1995.

[51] V. Faure, R. Lobjois, and N. Benguigui, "The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 40, pp. 78–90, 2016.

[52] M. S. Young, J. M. Mahfoud, N. A. Stanton, P. M. Salmon, D. P. Jenkins, and G. H. Walker, "Conflicts of interest: The implications of roadside advertising for driver attention," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 12, pp. 381–388, 2009.

[53] C. W. Myers and W. D. Gray, "Visual scan adaptation during repeated visual search," *J. Vis.*, vol. 10, no. 8, pp. 4-1–4-14, 2010.
[54] U. Neisser, "Anticipations, images, and introspection," *Cognition*, vol. 6, pp. 169–174, 1978.
[55] B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups," *Human Factors*, vol. 54, pp. 454–468, 2012.
[56] M. A. Recarte and L. M. Nunes, "Mental workload while driving: Effects on visual search, discrimination, and decision making," *J. Exp. Psychol. Appl.*, vol. 9, pp. 119–137, 2003.
[57] A. Newell, *Unified Theories of Cognition*. Cambridge, MA, USA: Harvard Univ. Press, 1994.
[58] S. Card, T. Moran, and A. Newell, "The model human processor—An engineering model of human performance," in *Handbook of Perception and Human Performance*, vol. 2. Oxford, U.K.: Wiley, 1986, pp. 1–45.
[59] D. E. Kieras and D. E. Meyer, "An overview of the EPIC architecture for cognition and performance with application to human-computer interaction," *Human-Comput. Interact.*, vol. 12, pp. 391–438, 1997.
[60] Y. Liu, "QN-ACES: Integrating queueing network and ACT-R, CAPS, EPIC, and Soar architectures for multitask cognitive modeling," *Int. J. Human-Comput. Interact.*, vol. 25, pp. 554–581, 2009.
[61] B. E. John and D. E. Kieras, "Using GOMS for user interface design and evaluation: Which technique?" *ACM Trans. Comput.-Human Interact.*, vol. 3, pp. 287–319, 1996.
[62] M. D. Byrne, "ACT-R/PM and menu selection: Applying a cognitive architecture to HCI," *Int. J. Human-Comput. Stud.*, vol. 55, pp. 41–84, 2001.
[63] T. Kujala and D. D. Salvucci, "Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control," *Int. J. Human-Comput. Stud.*, vol. 79, pp. 66–78, 2015.
[64] M. Nijboer, J. Borst, H. van Rijn, and N. Taatgen, "Contrasting single and multi-component working-memory systems in dual tasking," *Cogn. Psychol.*, vol. 86, pp. 1–26, 2016.
[65] C. Changxu and Y. Liu, "Queuing network modeling of the psychological refractory period (PRP)," *Psychological Rev.*, vol. 115, no. 4, pp. 913–954, 2008.
[66] D. E. Meyer and D. E. Kieras, "A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms," *Psychological Rev.*, vol. 104, pp. 3–65, 1997.
[67] O. Tsimhoni and Y. Liu, "Modeling steering using the queueing network—model human processor (QN-MHP)," *in Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 2003, pp. 1875–1879.
[68] R. Feyen and Y. Liu, "Modeling task performance using the queuing network-model human processor (QN-MHP)," in *Proc. 4th Int. Conf. Cogn. Model.*, 2001, pp. 73–78.
[69] J. H. Lim and Y. Liu, "A queuing network model for visual search and menu selection," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 2004, pp. 1846–1850.
[70] C. Wu and Y. Liu, "Queuing network modeling of driver workload and performance," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 528–537, Sep. 2007.
[71] L. Bi, C. Wang, X. Yang, M. Wang, and Y. Liu, "Detecting driver normal and emergency lane-changing intentions with queuing network-based driver models," *Int. J. Human-Comput. Interact.*, vol. 31, pp. 139–145, 2015.
[72] *Road* Vehicles-*Ergonomic* Aspects *of* Transport Information *and* Control Systems—*Occlusion* Method *to* Assess Visual Demand Due *to the* Use *of* In-Vehicle Systems, ISO 16673:2007, 2007.
[73] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cogn. Psychol.*, vol. 8, pp. 441–480, 1976.
[74] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *Int. J. Ind. Ergonom.*, vol. 24, pp. 631–645, 1999.
[75] A. Poole and L. J. Ball, "Eye tracking in HCI and usability research," *Encyclopedia Human Comput. Interact.*, vol. 1, pp. 211–219, 2006.
[76] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. Dordrecht, Netherlands: Springer, 1987, pp. 115–141.
[77] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, pp. 194–203, 2001.
[78] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[79] G. Zelinsky, W. Zhang, B. Yu, X. Chen, and D. Samaras, "The role of top-down and bottom-up processes in guiding eye movements during visual search," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1569–1576, 2005.
[80] A. Treisman, "Preattentive processing in vision," *Comput. Vis., Graph., Image Process.*, vol. 31, pp. 156–177, 1985.
[81] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bull. Rev.*, vol. 1, pp. 202–238, 1994.
[82] J. E. Hoffman, "A two-stage model of visual search," *Perception Psychophys.*, vol. 25, pp. 319–327, 1979.
[83] G. R. Barnes, "Vestibulo-ocular function during co-ordinated head and eye movements to acquire visual targets," *J. Physiol.*, vol. 287, pp. 127–147, 1979.
[84] J. S. Stahl, "Amplitude of human head movements associated with horizontal saccades," *Exp. Brain Res.*, vol. 126, no. 1, pp. 41–54, 1999.
[85] R. G. Feyen, "Modeling human performance using the queuing network-model human processor (QN-MHP)," Ph.D. dissertation, Univ. of Michigan, Ann Arbor, MI, USA, 2002.
[86] L. C. McPhee, C. T. Scialfa, W. M. Dennis, G. Ho, and J. K. Caird, "Age differences in visual search for traffic signs during a simulated conversation," *Human Factors*, vol. 46, pp. 674–685, 2004.
[87] S. Ullman, "Visual routines," *Cognition*, vol. 18, pp. 97–159, 1984.
[88] G. R. Loftus and N. H. Mackworth, "Cognitive determinants of fixation location during picture viewing," *J. Exp. Psychol. Human Perception Perform.*, vol. 4, pp. 565–572, 1978.
[89] M. M. Hayhoe, D. G. Bensinger, and D. H. Ballard, "Task constraints in visual working memory," *Vis. Res.*, vol. 38, pp. 125–137, 1998.
[90] K. Grill-Spector and N. Kanwisher, "Visual recognition as soon as you know it is there, you know what it is," *Psychological Sci.*, vol. 16, pp. 152–160, 2005.
[91] C. Wu, "Queueing network modeling of human performance and mental workload in perceptual-motor tasks," Ph.D. dissertation, Univ. of Michigan, Ann Arbor, MI, USA, 2007.
[92] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vis. Res.*, vol. 40, pp. 1469–1487, 2000.
[93] E. E. Cureton, "Rank-biserial correlation," *Psychometrika*, vol. 21, no. 3, pp. 287–290, 1956.

Ye Lim Rhie received the B.S. degree in industrial engineering from Hongik University, Seoul, Korea, in 2012, and the Ph.D. degree from the Department of Industrial Engineering, Seoul National University, Seoul, in 2017.

She is a Senior Researcher with the Agency for Defense Development, Daejeon, Korea. Her research interests include human–computer interaction and modeling and simulation.

Ji Hyoun Lim (M'15) received the B.S.E degree in industrial engineering with minor in psychology from Seoul National University, Seoul, Korea, in 2000, and the M.S.E. and Ph.D. degrees in industrial and operations engineering from the University of Michigan, Ann Arbor, MI, USA, in 2003 and 2007, respectively.

She is a Human Factors Engineer in San Francisco/Bay area, CA, USA. Her research interests include computational cognitive modeling, semantic network analysis, and human automation interaction.

Myung Hwan Yun (M'17) received the B.S. and M.S. degrees in industrial engineering from Seoul National University, Seoul, Korea, and the Ph.D. degree in industrial and manufacturing engineering from Penn State University, University Park, PA, USA.

He is a Professor with the Department of Industrial Engineering and Institute for Industrial Systems Innovation, Seoul National University. His research interests include human factors in product design and affective product design.

A Formal Approach to Connectibility Affordances

Andrew J. Abbate, *Member, IEEE*, and Ellen J. Bass, *Senior Member, IEEE*

***Abstract*—Connectibility affordances, or opportunities for a user to establish input/output cable connections, can be critical to the safety and usability of complex systems. To support model-based analyses, this research introduces a formal approach: Connectibility Affordance VErification, Modeling, and ENumeration (CAVEMEN). CAVEMEN is applicable to a human-environment system encompassing physical entities with specified properties, a user with specified motor abilities, and connectibility affordance instances involving different combinations of source-target connections. The modeling technique leverages object-oriented principles to define one instance of a connectibility affordance with respect to one unique combination of connection sources and targets. An inspection technique supports the enumeration of connectibility affordance instances that are desired (supporting a correct connection) and undesired (supporting an incorrect connection). A model checking technique aids in verifying accuracy, meaning the user can actualize desired affordance instances, and robustness, meaning undesired affordance instances never emerge. An XML-based grammar, a model checking syntax translation tool, and a linear temporal logic specification of accuracy and robustness support the analyses. We demonstrate CAVEMEN with a pacemaker system case study. The inspection aids in identifying nine desired and 18 undesired instances of chamber-port connectibility. The trace evaluation shows that accuracy and robustness depends on whether an entity property of interest can change in the environmental context. These results indicate that CAVEMEN shows promise for analyzing connectibility affordances of a safety-critical system.**

***Index Terms*—Affordance, formal methods, human performance modeling, interface evaluation, model-based design, usability.**

I. Introduction

A COMPLEX system can incorporate connectible hardware that a user (e.g., installer and operator) needs to physically manipulate, such as cables, output connectors, and input sockets. To inform the design of connectible hardware, human factors engineering (HFE) researchers and standards organizations have developed measures that should be tested early in the design process. For example, connectible hardware can be considered *accurate* if the user can manipulate it in ways that establish operational configurations, and *robust* if the system helps to prevent the user from establishing incorrect configurations [1], [2].

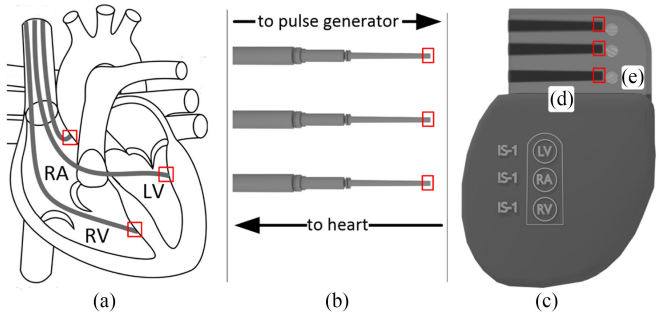


Fig. 1. Graphical rendering of a pacemaker system. Red boxes are added for reference in text. (a) Patient's heart showing segments of each lead. Lead distal tips are implanted in three chambers: LV stands for "left ventricle." RA stands for "right atrium." RV stands for "right ventricle." (b) Lead proximal tips. Arrows are added to identify connection targets of both lead tips (middle segments not shown). (c) Pulse generator. (d) Input ports. (e) Loosened set screw.

Consider the example of a surgically implanted pacemaker (see Fig. 1). Such a system provides heart failure patients with life-sustaining therapy via a programmed pulse generator. The pulse generator detects cardiac anomalies and delivers corrective electrical pulses to three heart chambers: the left ventricle (LV), which pumps oxygenated blood through the body; the right atrium (RA), which receives deoxygenated blood from the body; and the right ventricle (RV), which pumps deoxygenated blood to the lungs. The pulses are delivered via three identical leads having a distal tip that is implanted in a heart chamber and a proximal tip that is connected to a corresponding pulse-generator port [labeled on the pulse generator from top to bottom in Fig. 1(c)]. We characterize an operational configuration as follows: each lead distal tip is contained within the interior of exactly one heart chamber, the interior of a heart chamber is covering the front surface of exactly one lead distal tip, and the front surface of a proximal tip [see red boxes in Fig. 1(b)] and the back surface of a pulse-generator port [see red boxes in Fig. 1(d)] are covering each other. Each pulse-generator port has a set screw that must be loosened for a lead proximal tip to be fully inserted [see Fig. 1(e)].

Suppose the configuration in Fig. 1 emerges during an implantation surgery: each lead distal tip is implanted in a target heart chamber, each lead proximal tip is disconnected from a target pulse-generator port [see Fig. 1(b) and (d)], and each port's set screw is loosened [see Fig. 1(e)]. To establish a chamber-port connection, the surgeon needs to align a proximal tip with a target pulse-generator port and move the pulse generator toward the proximal tip with sufficient force to establish the connection. Three instances of such an action support an accurate operational configuration via three correct connections: LV chamber to LV

Manuscript received April 10, 2018; revised August 30, 2018; accepted November 17, 2018. Date of publication January 31, 2019; date of current version November 21, 2019. The work of A. J. Abbate was supported by the 2015–2016 U.S. Department of Education GAANN Interdisciplinary Collaboration and Research Enterprise for Healthcare fellowship (Grant P200A150023). (*Corresponding author: Andrew J. Abbate.*)

A. J. Abbate was with Drexel University, Philadelphia, PA 19104 USA. He is now with the Pacific Science & Engineering Group, San Diego, CA 92121 USA (e-mail: andrewabbate@pacific-science.com).

E. J. Bass is with the College of Nursing and Health Professions and the College of Computing and Informatics, Drexel University, Philadelphia, PA 19104 USA (e-mail: ellen.j.bass@drexel.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2886265

port, RA chamber to RA port, and RV chamber to RV port. However, unsafe configurations for the patient, such as one in which the LV chamber is connected to the RV port, are also possible (in part because there are three of the same lead, rather than a physically different lead for each chamber-port connection). Thus, while utilizing three of the same lead could have logistical benefits (e.g., ease of manufacturing), such a design also enables incorrect chamber-port connections. This reflects a robustness problem—one that emerges in existing pacemaker systems.[1]

A. Connectibility Affordances

The pacemaker system example of chamber-port connectibility could be characterized as an *affordance* [3]—an opportunity for a user to execute an action in a human-environment system (HES) [4].

In this research, an HES encompasses a three-dimensional (3-D) spatial area, a set of physical objects (referred to as *entities*), a user, and contextual factors that shape human-device interaction (e.g., automation). The user has motor abilities to physically manipulate the entities, whereas the entities have properties that physically constrain or support physical manipulations. An affordance emerges when specified properties of the entities and motor abilities of the user co-occur. If an affordance emerges, then the user can actualize it by executing one corresponding motor action.

Our interpretation of affordance is one of several that have proven useful in model-based analyses of human-interactive systems. Extant models commonly specify an affordance with respect to one human operator, one environment, and one particular set of physical entities therein. However, as demonstrated with the pacemaker system example, an affordance can involve many different combinations of physically equivalent entities. Such an affordance can be defined in an object-oriented way, where each unique combination of physically equivalent entities supports one unique instance of an object-oriented affordance. This research focuses on object-oriented connectibility affordances.

To characterize object-oriented connectibility affordances, consider a system in which there are many duplicates of the same entities and thus many ways of connecting them. Every unique connection configuration can reflect a different instance of the same connectibility affordance. The number of connection configurations—and thus the number of affordance instances—is defined formally in (1), where A is the number of instances, E is the number of duplicate entities that can establish the connection, $sources_1, \ldots, sources_n$ is the number of connection sources on the entity, and $targets_1, \ldots, targets_n$ is the number of connection targets for each source

$$A = E(sources_1 \times targets_1 \times \cdots \times sources_n \times targets_n). \quad (1)$$

Currently, it could be difficult for analysts to enumerate and evaluate object-oriented connectibility affordances with respect to many duplicate entities in combination. The problem space is further complicated by emergent behaviors in the environmental context, including user behaviors, such as physically manipulating multiple entities in parallel, and system behaviors, such as automated sensing and actuation. Thus, analysts could benefit from an improved approach. Such an approach should facilitate the enumeration of desired/undesired affordance instances with respect to duplicate entities in combination. It should also incorporate unambiguous methods and measures for verifying accuracy and robustness. The measures should be applicable to desired/undesired instances of the same connectibility affordance, whereas the methods should be applicable to emergent behaviors in the environmental context.

B. Research Contributions

To support improved analyses of object-oriented connectibility affordances, this research introduces a formal, model-based approach: Connectibility Affordance VErification, Modeling, and ENumeration (CAVEMEN). CAVEMEN extends our earlier research [5], which employs model checking—a highly automated technique for "proving" system properties [6]—to support the identification of potential affordance problems. Extensions enable the enumeration of desired/undesired connectibility affordance instances with respect to many duplicate entities in combination, verification of accuracy and robustness, and modeling of emergent behaviors in different environmental contexts.

Two tools and two analysis techniques support these extensions. Tools include CAVEMEN-XML, a custom encoding language for specifying an HES and emergent affordances therein, and an automated translator, which parses an instantiated CAVEMEN-XML representation and generates model checking syntax.[2] Analysis techniques include inspection, a technique for the analyst to enumerate and characterize desired/undesired instances of the same connectibility affordance, and trace evaluation, a technique for verifying a specification of accuracy and robustness via model checking. We introduce an implementation of the trace evaluation technique for analyzing emergent behaviors in two environmental contexts: one in which a condition for the affordance to emerge cannot change, and one in which the same condition can change. We demonstrate an application of CAVEMEN with a pacemaker system case study.

II. Background

A. Affordance Modeling

Researchers at the intersection of ecological psychology and computer science have developed a variety of model-based approaches to affordance. As mentioned, CAVEMEN differs from extant approaches because it defines affordances in an object-oriented way, where one instance of an affordance involves a unique combination of duplicate environment entities. In contrast, extant approaches commonly model one affordance at a time with respect one fixed set of entities. These approaches commonly enable a normative, task-analytic approach to affordances, where human perception and cognition partially control what affordances emerge and what actions are taken. CAVEMEN offers a different perspective—one that exclusively addresses physical artifacts, but inclusively addresses both desired

[1] See for example https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/detail.cfm?mdrfoi__id=6920003&pc=DTB

[2] The XML schema, translation tool, and case study models described in this article are available for download at https://github.com/andrew-j-abbate/CAVEMEN

and undesired affordances. We next review a subset of the affordance modeling literature that reflects these differences.

Wells [7] integrates the concept of affordance with Turing's theory of computation [8]. An affordance model captures one normative, temporal ordering of human actions within a Turing machine representation. Symbolic variables represent a human operator, actions that can be executed, and properties of the environment. The model enables simulation of a human operator progressing toward one goal state of the environment by interacting with a specified set of entities.

Turvey [9] models affordances in the context of a particular task or objective. Environment properties and human properties operate as inputs to a "juxtaposition function" that defines the intersection of affordances and the human operator's goal-driven intentions. The model assumes that an affordance only emerges if it supports an intended behavior; and if an affordance emerges, then it is always actualized.

According to Stoffregen [4], one potential issue with Turvey's model is that it does not account for a human operator choosing against actualizing an affordance. He addresses this issue by removing the juxtaposition function and replacing it with a psychological choice function. Together, the affordance model and psychological choice function specify that an affordance can emerge even if it does not support an intended behavior, and the human operator will choose to actualize the affordance only if it supports an intended behavior.

In [10], Rothrock *et al.* integrate the models described in [9] and [7] with infrastructure that abstracts human perception. Three types of sensory functions represent what audio, visual, and haptic properties of the environment the human operator can perceive. An affordance emerges if the human operator can perceive relevant environment properties, whereas the human operator actualizes the affordance if it supports progress toward a goal [11].

In [12], Lenarčič and Winter leverage situation theory [13] to define a hierarchical model of affordance. Here, an affordance is composed of two hierarchical elements: one situation and one human operator. The situation is defined by a set of environmental conditions, and the human operator is defined by a set of cognitively/physically feasible abilities [12]. The model captures HES dynamics as the human operator executes goal-oriented actions.

B. Model Checking

Model checking commonly involves three sequential steps for the analyst: encoding a *formal model* of the target system, which can abstract a broad range of temporally evolving behaviors in terms of valued-variable states and next-state transitions [14]; encoding a specification that unambiguously characterizes a behavior of interest with respect to the formal model, typically using the semantics of a temporal logic, such as linear temporal logic (LTL) [15]; and invoking a model checker that searches the formal model for specification violations. If the model checker finds a specification violation, then it returns one trace of sequentially ordered states and transitions through the formal model leading up to the violation. If the specification characterizes a desired behavior, then the trace is called a *counterexample*. If the specification characterizes an undesired behavior, then the trace is called a *witness*.

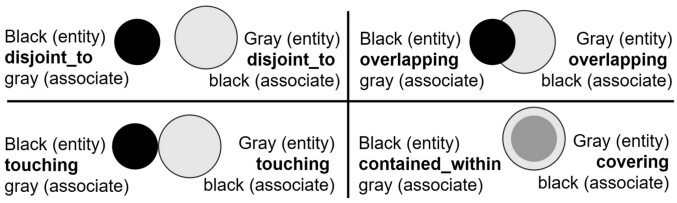


Fig. 2. Topology-keyword semantics in two dimensions. Topology keywords are in boldface text.

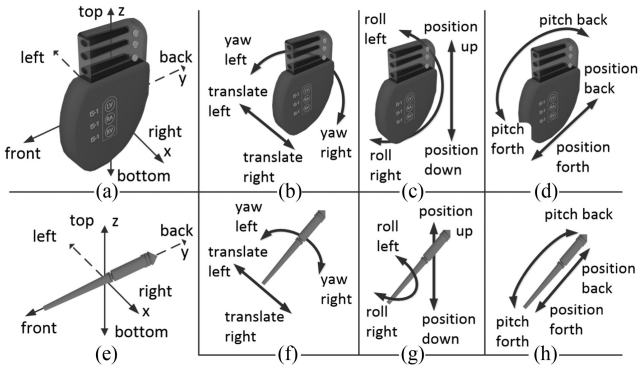


Fig. 3. (a) Surfaces of a pacemaker pulse generator and its ports. (b)–(d) Pulse generator movements. (e) Surfaces of lead proximal tips. (f)–(h) Lead proximal tip movements.

C. Analyzing Affordances via Model Checking

Our earlier approach [5] enables model checking analyses of affordances with respect to an HES encompassing a 3-D spatial environment, physical entities, and a human operator. We accomplish this using Stoffregen's affordance formalism [4], standard engineering terminology [16]–[18], and the Symbolic Analysis Laboratory (SAL) model checking system [19]. We describe these techniques ahead to aid in understanding CAVEMEN extensions.

1) Environment Model: The environment is modeled using one hierarchical variable *X*. Lower level variables represent the entities therein and their relevant physical properties. For reduced ambiguity, the analyst specifies each entity's part-whole composition, where parts are permanently attached fixtures, each of which has its own part-whole composition and properties.

Properties can be binary (true/false), numerical, or spatial relationship. Binary and numerical properties are modeled using variable names that aid in identifying what property is being represented. Spatial-relationship properties incorporate one topology keyword derived from [16], one direction keyword derived from [17], and one other entity (referred to as an "associate") with respect to which the spatial relationship is defined.

Five topology keywords (disjoint to, touching, overlapping, contained within, and covering) enable the analyst to specify 2-D connectedness of one entity with respect to one associate (see Fig. 2). Six direction keywords (top, bottom, left, right, back, and front) define unchanging directional surfaces of the associate [see for example the surfaces in Fig. 3(a) and (e)], which the analyst can either interpret as interior surfaces or exterior surfaces, depending on the application. Each topology-direction pair is mutually exclusive; for example, an entity can-

not be specified as simultaneously touching and overlapping the front surface of an associate. These semantics enable specification of 3-D connectedness between *any* surface(s) of an entity with respect to one *particular* surface of an associate. For reduced ambiguity, the analyst should encode an accompanying spatial-relationship property that specifies 3-D connectedness between the associate and one surface of the entity.

When instantiating a formal model, the analyst assigns each property one of the following behaviors: an unchanging initial-state value, which is used if the property never changes; a particular next-state value, which is used if the property only changes due to the user actualizing an affordance; or one of many possible next-state values, which is used if the property can change due to other events in the environmental context. An unchanging property can only operate as an affordance input, whereas a transitioning property can operate as an affordance input, output, or both (affordance inputs and outputs are explained in Section II-C3).

2) User Model: The user is modeled using one variable *Z*, which leverages the same hierarchical representation of the entities as the environment variable *X*. However, instead of representing physical properties of the entities, lower level variables of *Z* represent motor abilities to physically manipulate the entities. The model assumes that the user maintains one stationary position in the spatial environment and that all entities can be moved in parallel.

Each motor ability is uniquely defined with respect to each entity. Six degrees of freedom (6DoF) [18] keywords describe how the user can position, translate, and rotate an entity about its origin. The direction of each movement is specifiable along the *x*, *y*, and *z* axes with respect to the surfaces that direction keywords represent for spatial-relationship properties [see Fig. 3(b)–(d), where directions correspond to surfaces in Fig. 3(a)]. The analyst assigns each keyword a numeric value that represents the maximum force magnitude with which a movement can be executed. Specifying force magnitude is useful if actualizing an affordance requires the user to overcome some opposing force imposed by the environment, such as air pressure or friction. If these force magnitudes are unknown, then the analyst should assign each 6DoF keyword a value of 1 or 0, meaning the movement is possible or impossible for the user respectively.

3) Affordance Model: Affordances are modeled as input/output functions. Inputs are conditions that the entity properties and motor abilities must concurrently satisfy for the affordance to emerge. Outputs are conditions that entity properties must satisfy as an immediate consequence of the user actualizing the affordance. A model checking instantiation of an extant affordance formalism [4] supports these semantics

$$possesses(affordance)(Z, X). \tag{2}$$

Equation (2) is a nested Boolean function of affordances (*affordance*), the user (*Z*), and the environment (*X*). *affordance* is an enumerated list of one or more named affordance variables, where each variable name aids in identifying the affordance being modeled. The outer function takes one affordance variable as an input, whereas the inner function takes the user *Z* and the environment *X* as inputs. Using all three inputs, (2) returns *true* for all affordances whose input conditions are satisfied by lower level variables of *Z* and *X*. This is accomplished for each modeled affordance by encoding a conjunction of Boolean expressions, one for each entity-property input and one for each motor-ability input. The conjunctions for entity-property inputs should be encoded first, followed by the conjunctions for motor-ability inputs. This is because motor-ability inputs specify what movements are possibly needed in any HES configuration satisfying the entity-property inputs, such as one in which the entities are facing different directions and need to be aligned in order to establish a connection.

After instantiating all conjunctions, the analyst can utilize (2) to encode one next-state transition for each modeled affordance, where next-state values are affordance-output conditions. Multiple next-state transitions can be enabled in any state of the formal model (i.e., multiple affordances can emerge in parallel), but exactly one transition can execute (i.e., one affordance can be actualized). CAVEMEN inherits this feature of our prior approach because it improves physical validity. For example, a connection source could be simultaneously connectible to multiple connection targets, each of which occupies a different spatial location. Since one connection source cannot simultaneously occupy multiple locations, exactly one instance of the affordance can realistically be actualized at a time.

III. CAVEMEN Extensions

Here, we identify how CAVEMEN extends our earlier approach with respect to object-oriented connectibility affordances. Where applicable, we identify extant tools and techniques that inspire the extensions. A case study demonstration appears in Section V.

1) Characterization of Desired/Undesired Connectibility Affordance Instances: As mentioned, one knowledge gap is the difficulty for an analyst to characterize all instances of a connectibility affordance with respect to many duplicate entities in combination. The CAVEMEN-XML language and translator address this gap.

Using CAVEMEN-XML, the analyst can specify each entity once, including what quantity of the entity (one or more) is applicable to the affordance(s) of interest. Natural language keywords specify each entity's numerically quantifiable, binary, and spatial-relationship properties that are relevant, including if/how these properties can transition in a formal model. Our custom translator instantiates model checking syntax representing each duplicate entity, including all lower level entities, properties, and next-state transitions. This extension enables the translator to generate all instances of the same affordance with respect to duplicate entities. Two additional extensions facilitate characterization and enumeration of the affordance instances: specification of the inputs and outputs in a general way with respect to duplicate entities, and generation of model checking syntax for each instance with respect to correct combinations of duplicate entities.

To support the first extension, we leverage a technique that has proven useful in the enhanced operator function model (EOFM) task-analytic framework [20], [21]. In EOFM, a custom, XML-based grammar (EOFM-XML) incorporates keywords for specifying cardinal and temporal orderings of actions/activities that can be executed in a goal-driven task, where cardinal orderings include all activities/actions, at least one activity/action, or exactly one activity/action. An automated translation tool generates model checking syntax representing all of the ways to

execute the same goal-driven task(s) with respect to an instantiated EOFM-XML representation. For connectibility affordance applications, natural language keywords of CAVEMEN-XML enable the analyst to specify what combination of duplicate entities must satisfy the input or output condition: all, at least some number, or exactly some number. Each condition is otherwise specified as in our earlier approach: inputs are conditions that must be satisfied for the affordance to emerge and outputs are conditions that become satisfied as a consequence of actualizing the affordance.

To support the second extension mentioned above, the translator uses custom algorithms that generate model checking syntax for all instances of the same connectibility affordance. The duplicate entities and affordance input/output combinations specified in a CAVEMEN-XML representation have corresponding model checking syntax that uniquely identifies each instance of the same entity, property, motor ability, and affordance.

2) Enumeration of Desired/Undesired Connectibility Affordance Instances: One purpose of the extensions mentioned earlier is to support the enumeration of desired/undesired connectibility affordance instances. This is accomplished in CAVEMEN via an inspection technique: the analyst enumerates desired and undesired instances of the same connectibility affordance with respect to translated model checking syntax. A desired instance has inputs supporting a correct connection, whereas an undesired instance has inputs supporting an incorrect connection. An ideal result is no undesired instances, whereas a minimally acceptable result is at least one desired instance. The inspection result informs additional formal model syntax that is needed to support verification of accuracy and robustness.

3) Specification and Verification of Accuracy and Robustness: Currently, model checking of connectibility affordances is constrained to the model checking syntax that the analyst can manually encode and the specifications that the analyst can identify. To support the application of other model checking approaches, researchers have developed generalizable specifications and tools that facilitate the encoding processes [22]. In this vein, CAVEMEN incorporates support for connectibility affordance applications.

To inform generalizable specifications, we introduce an LTL specification of accuracy and robustness (3), where the first line is accuracy and the second line is robustness. Accuracy means that a specified set of desired instances will eventually be actualized, and robustness means that no undesired instances will ever emerge. "F" is a temporal operator meaning "eventually," "G" is a temporal operator meaning "always," and "∀" is a variable quantifier meaning "for all"

$$\left(\begin{array}{l} \mathsf{F}(\forall d : desired \,|\, actualized[d]) \wedge \\ \quad \mathsf{G}(\forall u : undesired \,|\, \neg(possesses(u)(Z, X)) \end{array}\right). \tag{3}$$

The variables *d* and *u* come from the enumerated lists *desired* and *undesired*, respectively, both of which the analyst instantiates based on the inspection result: *desired* includes desired affordance instances of interest, and *undesired* includes all undesired instances. The translator automatically instantiates *actualized*, which is an indexed array of Boolean-valued affordance instances. The translator assigns each index (representing one affordance instance) an initial value of *false* and an irreversible next state of *true* once the affordance is actualized. These semantics enable the formal model to keep track of which affordance instances have been actualized.

The translator instantiates two forms of the specification that support different kinds of trace evaluations: one positive form, which is encoded as shown in (3), and one negative form, which has "¬" (meaning "not") outside the large, left-hand parenthesis shown in (3). The positive form is verified to search for counterexamples. If a counterexample is returned, then it will show one trace through the formal model in which one or more undesired affordance instances emerge, a desired instance is not actualized, or both (an undesired result). No counterexamples means that there are no violations of accuracy and robustness (a desired result). The negative form is verified to search for witnesses. If a witness is returned, then it will show a trace through the model in which all specified desired instances are actualized without an undesired instance ever emerging (a desired result). No witnesses means that the system is not both accurate and robust (an undesired result). Specifications are verified using the SAL infinite bounded model checker (SAL-INF-BMC), as its algorithms are optimized for generating traces [19] and avoiding the problem of state-space explosion [23].

4) Modeling of Emergent Behaviors in the Environmental Context: Two CAVEMEN extensions support improved analyses of emergent behaviors in the environmental context: formal modeling of emergent user motor abilities to move subsets of the entities in parallel, and evaluating accuracy and robustness with respect to different environmental contexts.

As mentioned, our earlier approach assumes that the user can always move all entities in parallel. For object-oriented connectibility affordances, one limitation of such an assumption is that the combination of entities that can be moved in parallel is critical to what instance of the affordance emerges. Consider the pacemaker system example of chamber-port connectibility. If one lead distal tip is implanted within each chamber and all lead proximal tips are disconnected from all pulse-generator ports (see Fig. 1), then which instance of chamber-port connectibility emerges depends in part on which lead proximal tip the surgeon can move in parallel with the pulse generator (such as by gripping either entity with each hand).

Robotics researchers have developed ways of formally modeling these kinds of parallel movement abilities by specifying that a robot has two arms and two hands, both of which can be utilized at the same time to move entities [24]. While such a technique has proven useful in robotics applications, modeling humans in the same way could be inappropriate per international accessibility standards [25]. Thus, to support the first extension mentioned before, CAVEMEN incorporates an alternative technique. Using natural language keywords of CAVEMEN-XML, the analyst can specify which entities (and how many duplicates) the user can move in parallel. The translator generates model checking syntax enabling exactly one such set of movements in every state of the formal model.

To support the second extension mentioned above, we leverage a trace evaluation technique that has proven useful in task-analytic applications [26]: comparing the model checking results of two formal models with respect to a one-line change of intermediate-language syntax. To employ this technique in CAVEMEN, the analyst identifies an entity property of interest

that operates as an affordance input. In one CAVEMEN-XML representation, the analyst specifies that this property has an unchanging initial state. In a second CAVEMEN-XML representation, the analyst specifies that this same property can transition nondeterministically, which abstracts a different environmental context. The analyst then verifies both translated formal models with respect to the same instantiation of accuracy and robustness. This technique aids in identifying whether emergent behaviors can affect accuracy and robustness.

IV. CAVEMEN-XML

CAVEMEN-XML enables the analyst to specify a static representation of an HES and connectibility affordances. It employs eXtensible Markup Language (XML) [27], an international standard for representing structured data. An XML document is defined within a single *root* element. Lower level elements are called *children*. All children of the root element can have valued attributes, text content, and children.

In CAVEMEN, an XML Schema Document (XSD) [28] grammar enforces CAVEMEN-XML syntax that is needed for the translator to generate model checking syntax. The translator, which can be called from a desktop web browser, validates CAVEMEN-XML syntax against the grammar. We next describe the syntax, structure, and semantics of CAVEMEN-XML. Elements are in boldface text, attributes are in italic text, and attribute values are in quotes. Subsection titles are children of the root element, **hes**.

A. Environment

A CAVEMEN-XML document has one **environment** element, which represents a 3-D spatial area. One or more child **entity** elements represent the physical objects therein.

Each **entity** element has two attributes: *name* and *quantity*. The *name* attribute value should be an alphabetic string that aids in identifying what entity is being specified. To support formal model translation, heterarchical entities must have unique *name* attribute values. The *quantity* attribute value is a positive integer that specifies how many of the entities are applicable to the affordance(s) of interest. One or more **property** child elements specify the entity's numerical, binary, and spatial-relationship properties.

Each **property** element has three attributes: *evolution*, *type*, and *name*. The *evolution* attribute can have one of three values: "human" if a property only transitions due to the user actualizing an affordance, "environment" if the property can transition due to other events in the environmental context (or also due to the user actualizing affordances), or "none" if the property never transitions. The *type* attribute can have one of three values: "Boolean" if the property is binary, "numerical" if the property is numerically quantifiable, or "spatial" if the property is a spatial relationship. Text content, which is only applicable for spatial-relationship properties, must reference one or more dot-separated *name* attribute values of one or more **entity** elements, where each dot specifies one step down in the **entity** element hierarchy. This convention is needed to specify reciprocal spatial relationships between the entity, surface(s) of the associate, and vice versa.

TABLE I
Quantity-Operator VALUE SEMANTICS

"all"	"at_least-*x*"	"exactly-*x*"
All entity instances must satisfy the condition	*x* or more entity instances must satisfy the condition	*x* entity instances must satisfy the condition, while all others must not

B. Human

A CAVEMEN-XML document has one **human** element, which represents the user. Child elements are **ability** and **ability-set**.

Each **ability** element has two attributes: *name* and *entity*. The *name* attribute value is a unique alphabetic string that aids in identifying what ability is being modeled. The **entity** attribute identifies what entity can be moved. Its value references one or more dot-separated *name* attribute values of **entity** elements.

For each child element of **ability**, attributes identify what 6DoF movements the user can execute. Each value is a positive number that specifies the maximum force magnitude with which the user can execute the movement. If force magnitudes are unknown, then each value should be "1" to specify that the movement is possible. Omitting an attribute is equivalent to specifying that the movement is impossible for the user.

Each **ability-set** element specifies parallel movements that can be executed in a mutually exclusive way with respect to all other movements. The *name* attribute value is a unique alphabetic string that aids in identifying what set of movements is being modeled. One or more **ability-ref** child elements specify what movements can be executed. Each **ability-ref** element has two attributes: *name* and *quantity*. The *name* attribute references an **ability** element by its *name* attribute. The *quantity* attribute value is a positive integer that specifies how many of the referent entity (i.e., the *entity* attribute value of the referenced **ability** element) can be moved in parallel.

C. Affordance

One or more **affordance** elements represent input/output affordances. Each **affordance** element has one valued attribute, *name*, which is a unique alphabetic string that aids in identifying what affordance is being modeled. Child elements are **environment-input**, **human-input**, and **environment-output**. We next describe the attributes and text content of these child elements.

Each **environment-input** element specifies an entity-property condition that must be satisfied for the affordance to emerge. It has three attributes: *name*, *quantity-operator*, and *equality-operator*. The *name* attribute value references the *name* attribute value of a **property** element. Together, the text content and *equality-operator* attribute specifies the condition with respect to the referent property (see Table II). If the referent property's *type* attribute value is "numerical" or "Boolean," then text content can be a logical expression (i.e., one that is either true or false). If the referent property's *type* is "spatial," then text content can be one topology keyword and one optional direction keyword. Omitting a direction keyword is equivalent to specifying that all six direction keywords must be assigned the same topology-keyword value. For any referent property, the

TABLE II
Equality-Operator VALUE SEMANTICS

Value	Applicable element(s)	Condition
"equal-to"	**environment-input**, **environment-output**, **human-input** children	The property or ability must be equal to the specified text content
"not-equal-to"	**environment-input**, **environment-output**, **human-input** children	The property or ability must not be equal to the specified text content
"less-than-or-equal-to"	Numerical-type **environment-input/ environment-output**, **human-input** children	The property or ability must be less than or equal to the specified text content
"greater-than-or-equal-to"	Numerical-type **environment-input/ environment-output**, **human-input** children	The property or ability must be greater than or equal to the specified text content
"less-than"	Numerical-type **environment-input/ environment-output**, **human-input** children	The property or ability must be less than the specified text content
"greater-than"	Numerical-type **environment-input/ environment-output**, **human-input** children	The property or ability must be greater than the specified text content

quantity-operator attribute specifies how many duplicate entities, each of which has an instance of the same referent property, must satisfy the condition (see Table I).

Each **human-input** element specifies a motor-ability condition that must be satisfied for the affordance to emerge. It has two attributes: *name* and *quantity-operator*. The *name* attribute value references the *name* attribute value of an **ability** element. Optional child elements specify the movements that are possibly needed in order to actualize the affordance. Together, the text content and *equality-operator* attribute of each child element specifies one condition, as described in Table II. Omitting child elements is equivalent to specifying that all movements of the entity identified in the referent **ability** element are needed to actualize the affordance, which is either all maximum force magnitudes or "1" (meaning the movement must be possible). The *quantity-operator* attribute specifies how many duplicate entities, each of which has an instance of the same ability, must satisfy the condition (see Table I).

Each **environment-output** child element specifies what entity-property changes occur as a consequence of actualizing the affordance. Its attributes and text context are encoded in the same way as described for **environment-input** elements mentioned earlier.

V. CASE STUDY

Here, we apply the CAVEMEN approach in a pacemaker system case study of chamber-port connectibility. We encode a CAVEMEN-XML representation representing the pacemaker system described in Section I. We model one heart having three applicable chambers, one pulse generator having three applicable ports, one set screw for each port, three leads, one proximal tip and one distal tip for each lead, and one surgeon (i.e., the user) who can move the pulse generator and any one lead proximal tip in parallel. We then translate the CAVEMEN-XML representation to SAL in order to conduct inspection and trace evaluation analyses.

Based on (1), we expect the translator to instantiate 27 instances of the chamber-port connectibility as there are three leads, one distal connection source having three targets, and one proximal connection source having three targets ($27 = 3(1 \times 3 \times 1 \times 3)$). To support the inspection, we define a desired instance as one supporting a correct chamber-port connection: LV chamber to LV port, RA chamber to RA port, and RV chamber to RV port. The translator appends integers to CAVEMEN-XML syntax in order to uniquely define duplicate entities; thus, for the purpose of this case study, we refer to the chambers and pulse-generator ports as follows: "1" corresponds to "LV," "2" corresponds to "RA," and "3" corresponds to "RV."

To support the trace evaluation, we verify the negative form of (3) to search for witnesses—traces through the formal model showing that it is possible for the system to satisfy the accuracy and robustness specification.

As mentioned in Section I, the case study pacemaker system can be considered accurate with respect to chamber-port connectibility because the surgeon can actualize three desired instances—one for each lead. In the parlance of CAVEMEN, we identify exactly one set of initial environment-input conditions in which the system can be considered accurate: all three set screws are loosened, all three lead distal tips are implanted correctly, and all three lead proximal tips are disconnected from pulse-generator ports (see the configuration shown in Fig. 1).

We hypothesize that these environment-input conditions cannot support both accuracy and robustness with respect the chamber-port connectibility, unless different conditions emerge. One such set of conditions includes which pulse-generator port set screws are loosened, which lead proximal tip the surgeon can move, and in which heart chamber the same lead's distal tip is implanted. To support accuracy, the set screw of the correct target port must be loosened (enabling the correct connection). To support robustness, set screws of incorrect target ports must be tightened (disabling incorrect connections). If the states of all three set screws can change, then three desired instances of chamber-port connectibility can be actualized without an undesired instance ever emerging. If states of the set screws are unchanging, then it should be impossible for the system to be both accurate and robust.

To test this hypothesis, we encode two CAVEMEN-XML representations: one in which the sets screws have unchanging initial-state values (the first model) and one in which their values can transition nondeterministically (the second model). We then translate, verify, and compare model checking results of these two models with respect to the negative form of the accuracy and robustness specification. We expect the first model to return no witnesses (an undesired result) and the second model to return a witness (a desired result).

A. CAVEMEN-XML Representations

Both CAVEMEN-XML representations are 63 lines. Three top-level **entity** elements (direct children of **environment**) include "Heart" (see Fig. 4c), which represents the patient's heart; "PulseGenerator" (see Fig. 4j), which represents the pulse generator; and "Lead" (see Fig. 4m), which represents three leads. "Heart" has one child entity named "Chamber" (see Fig. 4b),

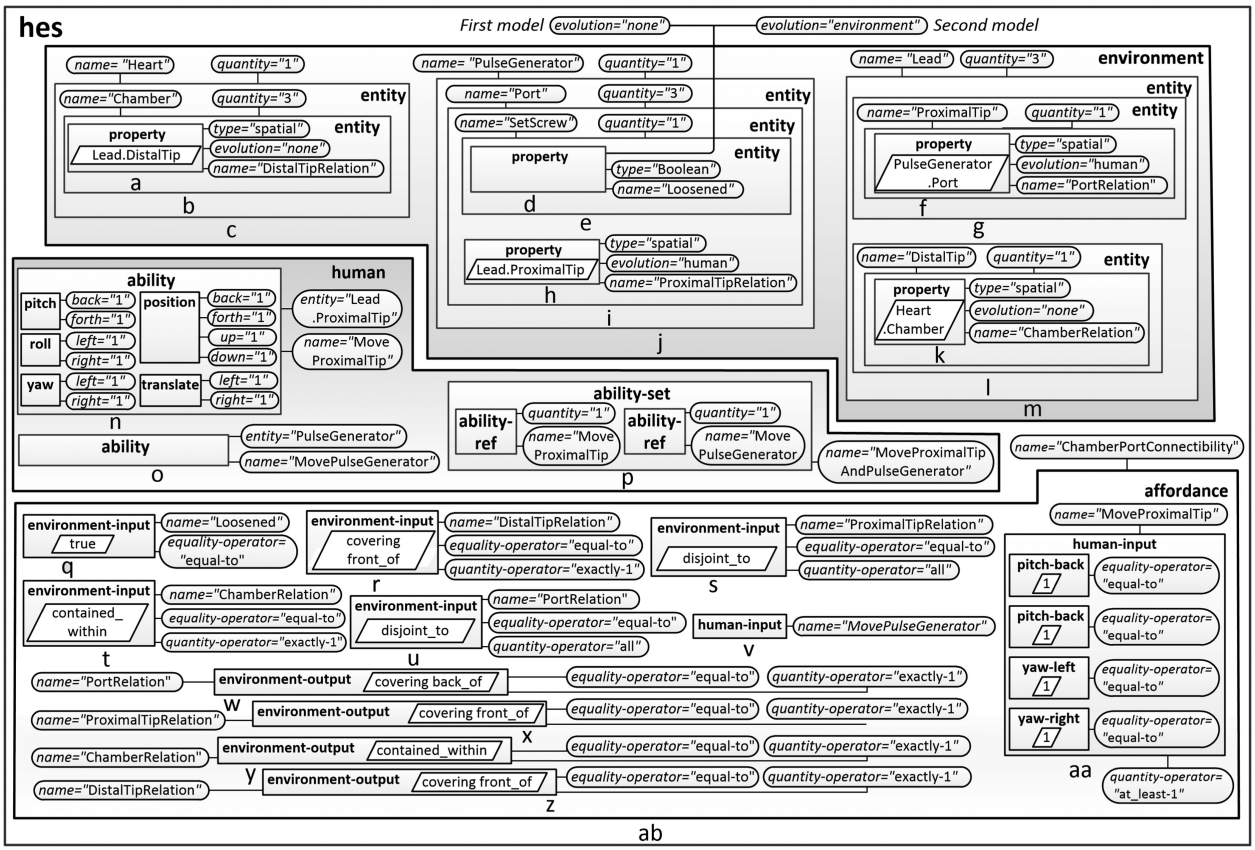


Fig. 4. CAVEMEN-XML representation of the pacemaker system HES. Elements, attributes, attribute values, and text content are represented using the conventions of Section IV. Letters are added for reference in text of Section V-A. Shading aids in differentiating child elements of **hes**: **environment** (white to gray), **human** (gray to white), and **affordance** (none).

which represents the three applicable chambers (LV, RV, and RA). "PulseGenerator" has one child entity named "Port" (see Fig. 4i), which represents the pulse generator's three ports. "Port" has one child entity named "SetScrew" (see Fig. 4e), which represents the set screw of each port. "Lead" has two child entities: "ProximalTip" (see Fig. 4g), which represents each lead's proximal tip; and "DistalTip," which represents each lead's distal tip (see Fig. 4l).

There are five applicable properties for the modeled entities. In the first model, three properties are specified as unchanging (*evolution* = "none"): one spatial property of "Chamber" named "DistalTipRelation," which represents the spatial relationship between a heart chamber and a lead distal tip (see Fig. 4a); one Boolean property of "SetScrew" named "Loosened," which represents whether a set screw is loosened (see Fig. 4d); and one spatial property of "DistalTip" named "ChamberRelation," which represents the spatial relationship between a lead distal tip and a heart chamber (see Fig. 4k). In the second model, "Loosened" is specified as changing independently of chamber-port connectibility (*evolution* = "environment").

In both models, two properties are specified as changing due to the surgeon actualizing an affordance (*evolution* = "human"): one spatial property of "Port" named "ProximalTipRelation," which represents the spatial relationship between a pulse generator port and a lead distal tip (see Fig. 4h); and one spatial property of "ProximalTip" named "PortRelation," which represents the spatial relationship between a lead distal tip and a pulse generator port (see Fig. 4f).

Both models include a **human** element having three child elements that represent the surgeon's motor abilities. One **ability** element named "MoveProximalTip" specifies that the surgeon can move a lead proximal tip along all axes and directions (see Fig. 4n). Another ability element named "MovePulseGenerator" specifies the same movements for the pulse generator (see Fig. 4o, child elements not depicted are identical to those of Fig. 4n). One **ability-set** element named "MoveProximalTipAndPulseGenerator" specifies that the surgeon can move the pulse generator and one lead proximal tip in parallel (see Fig. 4p). We do not model the surgeon's ability to move two lead proximal tips in parallel because such an ability is unneeded to actualize chamber-port connectibility.

One **affordance** element named "ChamberPortConnectibility" represents the affordance of interest (see Fig. 4ab). Five **environment-input** elements specify that "ChamberPortConnectibility" emerges if all of the following conditions hold: at least one pulse-generator port's set screw is loosened (see Fig. 4q), exactly one lead distal tip is contained within a heart chamber (see Fig. 4t), exactly one heart chamber is covering the front surface of a lead distal tip (see Fig. 4r), all pulse generator ports are disjoint to a lead proximal tip (see Fig. 4s), and all lead proximal tips are disjoint to a pulse-generator port (see Fig. 4u). These conditions capture many possible HES configurations, including the one shown in Fig. 1.

Two **human-input** elements specify that the following conditions must hold for "ChamberPortConnectibility" to emerge: the surgeon can position, translate, pitch, yaw, and roll the

pulse generator in all directions (see Fig. 4v); and the surgeon can pitch and yaw at least one lead proximal tip in all directions (see Fig. 4aa). These conditions capture the movements that are potentially needed for the surgeon to actualize chamber-port connectibility in any HES configuration satisfying the **environment-input** conditions.

Four **environment-output** elements specify that the following conditions must hold as a consequence of actualizing "ChamberPortConnectibility": exactly one lead proximal tip is covering the back surface of a pulse generator port (see Fig. 4w), exactly one pulse-generator port is covering the front surface of a lead proximal tip (see Fig. 4x), exactly one lead distal tip is contained within a target heart chamber (see Fig. 4y), and exactly one heart chamber is covering the front surface of a lead distal tip (see Fig. 4z).

B. Inspection

We invoked the translation tool to generate SAL models for the two CAVEMEN-XML representations. A total of 2662 lines of SAL code were generated for the first model. A total of 2668 lines of SAL code were generated for the second model. Both translated models represented 27 instances of the affordance, where nine are desired and 18 are undesired. The nine desired instances support a correct chamber-port connection, whereas the 18 undesired instances support an incorrect chamber-port connection.

To aid in understanding the logic of the translation tool, all input conditions for one of nine desired instances are shown in (4). References to Fig. 4 specify what line(s) of (4) correspond to **environment-input** elements of the CAVEMEN-XML representation. Horizontal lines demarcate sets of input conditions that correspond to the same Fig. 4 letter

$$
\begin{array}{l}
X.PulseGenerator.Port_1.SetScrew_{Loosened} = true \ \wedge \text{(Fig. 4q)} \\
\hline
\forall d : \{front\_of,\ back\_of,\ top\_of,\ bottom\_of,\ left\_of,\ right\_of\} \mid \\
X.Lead_1.DistalTip[Heart.Chamber_1][d] = contained\_within \ \wedge \\
\neg(X.Lead_2.DistalTip[Heart.Chamber_1][d] = contained\_within) \wedge \\
\neg(X.Lead_3.DistalTip[Heart.Chamber_1][d] = contained\_within) \wedge \text{(Fig. 4t)} \\
\hline
X.Heart.Chamber_1[Lead_1.DistalTip][front\_of] = covering \ \wedge \\
\neg(X.Heart.Chamber_2[Lead_1.DistalTip][front\_of] = covering) \ \wedge \\
\neg(X.Heart.Chamber_3[Lead_1.DistalTip][front\_of] = covering) \ \wedge \text{(Fig. 4r)} \\
\hline
X.PulseGenerator.Port_1[Lead_1.ProximalTip][d] = disjoint\_to \wedge \\
X.PulseGenerator.Port_2[Lead_1.ProximalTip][d] = disjoint\_to \wedge \\
X.PulseGenerator.Port_3[Lead_1.ProximalTip][d] = disjoint\_to \wedge \text{(Fig. 4s)} \\
\hline
X.Lead_1.ProximalTip[PulseGenerator.Port_1][d] = disjoint\_to \wedge \\
X.Lead_2.ProximalTip[PulseGenerator.Port_1][d] = disjoint\_to \wedge \\
X.Lead_3.ProximalTip[PulseGenerator.Port_1][d] = disjoint\_to \wedge \text{(Fig. 4u)} \\
\hline
(\forall m : \{pitch\_back, pitch\_forth, yaw\_left, yaw\_right, translate\_left, \\
translate\_right, position\_up, position\_down, position\_back, position\_forth\} \mid \\
Z.PulseGenerator[m] = 1) \ \wedge \text{(Fig. 4v)} \\
\hline
(\forall m : \{pitch\_back,\ pitch\_forth,\ yaw\_left,\ yaw\_right\} \mid \\
Z.Lead_1.ProximalTip[m] = 1) \text{ (Fig. 4aa)}
\end{array} \tag{4}
$$

The depicted input conditions support a desired instance of chamber-port connectibility because they specify a connection between port-1 and chamber-1. In (4), such a connection is established via lead-1. Input conditions for the eight remaining desired instances have similar syntax, but they reflect other correct connections via leads 1, 2, and 3, such as port-2 and chamber-2 via lead-3.

TABLE III
MODEL CHECKING RESULTS

Model	Result	Verification time (s)
First	*no witnesses*	10.29
Second	*witness*	20.16

C. Trace Evaluation

To support the trace evaluation, we instantiated two enumerated-list variables: *desired*, which incorporates three desired instances of the affordance (one for each lead), and *undesired*, which incorporates the 18 undesired instances. The selected desired instances include $ChamberPortConnectibility_1$, which supports connecting chamber-1 to port-1 via lead-1; $ChamberPortConnectibility_{14}$, which supports connecting chamber-2 to port-2 via lead-2; and $ChamberPortConnectibility_{27}$, which supports connecting chamber-3 to port-3 via lead-3. Using these variables, we verified the negative form of (3) in order to search for witnesses in both translated formal models.

Verification was conducted using SAL-INF-BMC [19] on a 3.5-GHz workstation with 64-GB RAM running the Ubuntu 16.04 desktop. Results are shown in Table III.

No witnesses were returned in the first model, indicating that if the pulse-generator ports' set screws have unchanging initial states, then the system cannot be considered both accurate and robust with respect to chamber-port connectibility.

The witness returned for the second model has five steps in which none of the 18 undesired instances emerged, but all three identified desired instances were actualized. In every step, all three lead distal tips are implanted within the three respective heart chambers as specified in (4). In the first step, the surgeon can move the pulse generator and the proximal tip of lead-2 as specified in (4). The set screw of port-2 is loosened, whereas the set screws of ports 1 and 3 are not. Thus, $ChamberPortConnectibility_{14}$ is actualized, resulting in a connection between chamber-2 and port-2 via lead-2. In the second step, the surgeon can move the pulse generator and the proximal tip of lead-3 as specified in (4). The set screw of port-3 is loosened, whereas the set screws of ports-1 and 2 are not. Thus, $ChamberPortConnectibility_{27}$ is actualized, resulting in a connection between chamber-3 and port-3 via lead-3. In the third step, the surgeon can move the pulse generator and the proximal tip of lead-1 as specified in (4). The set screw of port-1 is loosened, whereas the set screws of ports 2 and 3 are not. Thus, $ChamberPortConnectibility_1$ is actualized, resulting in a connection between chamber-1 and port-1 via lead-1. No transitions execute in the fourth or fifth step; thus, the final state is one in which each lead establishes a correct chamber-port connection. This result indicates that if the states of all three set screws can change, then it is possible for the pacemaker system to be both accurate and robust with respect to chamber-port connectibility.

VI. Discussion and Conclusions

This paper has introduced CAVEMEN, a formal approach to object-oriented connectibility affordances. CAVEMEN-XML enables the analyst to specify a formal representation of the HES, including parallel entity movements that are possible for a user and connectibility affordances with respect to duplicate entities in combination. The translator reduces the need for manually encoding model checking syntax by automatically generating all instances of the same connectibility affordance in the native syntax of SAL [19], including positive and negative instantiations of the accuracy and robustness specification. The inspection technique supports enumerating desired/undesired affordance instances with respect to translated model checking syntax. The trace evaluation technique supports analyses of accuracy and robustness in different environmental contexts.

We demonstrated CAVEMEN with a pacemaker system case study. The inspection showed that the HES supports 27 instances of chamber-port connectibility: nine desired and 18 undesired. The trace evaluation showed that different environmental contexts can affect accuracy and robustness: if an affordance input of interest never changes (whether a pulse generator port's set screw is loosened), then the system cannot be both accurate and robust; if the same affordance input can change nondeterministically, then the system can be both accurate and robust. These results indicate that CAVEMEN shows promise for analyzing object-oriented connectibility affordances of a safety-critical system.

A. Methodological Considerations

The case study illustrated methodological benefits of CAVEMEN with respect to the identified research contributions, including the enumeration of desired/undesired connectibility affordance instances, verification of accuracy and robustness, and modeling of emergent behaviors in different environmental contexts.

Regarding the enumeration of desired/undesired connectibility affordance instances, we showed that CAVEMEN can address the number of unique connection configurations defined in (1). In the case study, the translator successfully generated an expected result of 27 instances, whereas the inspection aided in identifying which instances are desired and undesired respectively. Thus, the inspection technique could be useful on its own—independently of the trace evaluation—as a way of determining what kinds of connections are possible due to duplicate entities and connection sources/targets.

Regarding the verification of accuracy and robustness, we showed that it is possible for trace evaluation results to correctly reflect an actual accuracy and robustness problem; i.e., the result of *no witnesses* for the first model indicates that it is possible for a surgeon to actualize an undesired instance of chamber-port connectibility—an expected result in light of the source material from a U.S. national database. This result indicates that the trace evaluation technique could be useful in a real-world application.

Regarding the modeling of emergent behaviors in different environmental contexts, we showed that nondeterministic state-transition behavior of one entity property can produce accuracy and robustness differences. Design insights gleaned from this technique come by imagining what could enable the behaviors observed in a returned witness. For example, one interpretation of the second model is that a hypothetical, automated control system adjusts set screws to ensure that only desired chamber-port connectibility instances emerge, such as by loosening the port-2 set screw and tightening the others if the surgeon is touching the lead-2 proximal tip (as in step-1 of the returned witness). Such an interpretation could inform the operational concept of an improved pacemaker system, although a different approach would be needed to model the envisioned automation.

An overarching benefit is reducing the amount of manually encoded syntax that is needed to support the analyses. For example, instantiating and translating the case study CAVEMEN-XML representation reflected a 2611-line reduction of manually encoded syntax. There is also a reduced need for knowledge of LTL semantics, as the translator generates both positive and negative instantiations of (3).

B. Limitations and Future Work

As shown with the case study analysis, our translation tool could generate thousands of lines of SAL code from a CAVEMEN-XML instantiation. This observation, which is mainly an artifact of duplicate-entity specifications, raises at least two scalability concerns: first, the translation algorithms could overwhelm a web browser, and second, infinite-bounded model checking could take an impracticably long time to execute. Future work should explore ways of reducing these concerns, such as by conducting benchmarking studies to inform a more efficient XML-to-SAL translation mechanism (see for example [29]).

Our interpretation of affordance is based on an extant formalism [4]; however, it could be limited with respect to other human-integrated system considerations. For example, Baber [30] indicates that we do not define the "rationale for performing the action [associated with an affordance] in the first place." He thus proposes an extended framework (called *Forms of Engagement*) that considers the user's perception, culture, and goals. Future work should explore ways of utilizing CAVEMEN in such a holistic context. One step toward achieving this could involve combining a CAVEMEN formal model with formal models of task behavior and perceptual artifacts. For example, a formal task model, such as EOFM [31], could aid in identifying whether connectibility affordances support a normative sequence of actions (or whether a normative procedure prevents undesired affordances from emerging), whereas a formal signifier model, such as BIGSIS [32], could aid in identifying whether perceivable properties of environment entities support correct connections (e.g., color-coded connection sources and targets).

Our interpretations of accuracy and robustness are inspired by extant HFE standards [1], [2]; however, there could be other interpretations that are not constrained to one type of connectibility affordance. For example, we did not demonstrate a way of verifying accuracy and robustness with respect to other affordances (e.g., "set screw tightenable/loosenable") or multiple connectibility affordances in combination. Future work should thus explore an extended set of accuracy and robustness specifications.

REFERENCES

[1] *Human Factors Engineering–Design of Medical Devices*," ANSI/AAMIHE75:2009/(R)2013, AAMI, Arlington, VA, USA, 2013.

[2] *Ergonomics of Human-System Interaction – Part 210: Human-Centered Design for Interactive Systems*, ISO 9241-210:2010, Geneva, Switzerland, 2017.

[3] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, J. B. Robert E Shaw, Ed. Hillsdale. Oxford, U.K.: N. J. Lawrence Erlbaum Assoc., 1977, pp. 67–82.

[4] T. A. Stoffregen, "Affordances as properties of the animal-environment system," *Ecol. Psychol.*, vol. 15, no. 2, pp. 115–134, 2003.

[5] A. J. Abbate and E. J. Bass, "Modeling affordance using formal methods," in *Proc. Annu. Meeting Human Factors Ergonom. Soc.*, 2017, pp. 723–727.

[6] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking.* Cambridge, MA, USA: MIT Press, 1999.

[7] A. J. Wells, "Gibson's affordances and Turing's theory of computation," *Ecol. Psychol.*, vol. 14, no. 3, pp. 140–180, 2002.

[8] A. M. Turing, "On computable numbers, with an application to the entscheidungs problem," *J. Math*, vol. 58, no. 345–363, pp. 230–265, 1936.

[9] M. T. Turvey, "Affordances and prospective control: An outline of the ontology," *Ecol. Psychol.*, vol. 4, no. 3, pp. 173–187, 1992.

[10] L. Rothrock, R. Wysk, N. Kim, D. Shin, Y.-J. Son, and J. Joo, "A modelling formalism for human-machine cooperative systems," *Int. J. Prod. Res.*, vol. 49, no. 14, pp. 4263–4273, 2011.

[11] J. Joo *et al.*, "Agent-based simulation of affordance-based human behaviors in emergency evacuation," *Simul. Model. Pract. Theory*, vol. 32, pp. 99–115, 2013.

[12] A. Lenarčič and M. Winter, "Affordances in situation theory," *Ecol. Psychol.*, vol. 25, no. 2, pp. 155–181, 2013.

[13] J. Barwise and J. Perry, "Situations and attitudes," *J. Philos.*, vol. 78, pp. 668–691, 1981.

[14] D. L. Parnas, "On the use of transition diagrams in the design of a user interface for an interactive computer system," in *Proc. 24th Nat. ACM Conf.*, 1969, pp. 379–385.

[15] E. M. Clarke, E. A. Emerson, and A. P. Sistla, "Automatic verification of finite-state concurrent systems using temporal logic specifications," *ACM Trans. Program. Lang. Syst.*, vol. 8, no. 2, pp. 244–263, 1986.

[16] M. J. Egenhofer and J. Herring, "A mathematical framework for the definition of topological relationships," in *Proc. 4th Int. Symp. Spatial Data Handling.*, 1990, pp. 803–813.

[17] A. U. Frank, "Qualitative spatial reasoning about distances and directions in geographic space," *J. Vis. Lang. Comput.*, vol. 3, no. 4, pp. 343–371, 1992.

[18] J. Bird and C. Ross, *Mechanical Engineering Principles*. New York, NY, USA: Routledge, 2012.

[19] L. De Moura, S. Owre, and N. Shankar, "The SAL language manual," Comput. Sci. Lab., SRI Int., Menlo Park, CA, USA, Tech. Rep. CSL-01-01, 2003.

[20] M. L. Bolton, R. I. Siminiceanu, and E. J. Bass, "A systematic approach to model checking human-automation interaction using task analytic models," *IEEE Trans. Syst., Man Cybern., Part A, Syst. Humans*, vol. 41, no. 5, pp. 961–976, Sep. 2011.

[21] M. L. Bolton and E. J. Bass, "Enhanced operator function model (EOFM): A task analytic modeling formalism for including human behavior in the verification of complex systems," in *The Handbook of Formal Methods in Human-Computer Interaction*, B. Weyers, J. Bowen, A. Dix, and P. Palanque, Eds. New York, NY: Springer, 2017, pp. 343–377.

[22] M. L. Bolton, N. Jimenez, M. M. van Paassen, and M. Trujillo, "Automatically generating specification properties from task models for the formal verification of human–automation interaction," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 5, pp. 561–575, Oct. 2014.

[23] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, and Y. Zhu, "Bounded model checking," *Adv. Comput.*, vol. 58, pp. 117–148, 2003.

[24] E. Şahin, M. Cakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adapt. Behav.*, vol. 15, no. 4, pp. 447–472, 2007.

[25] *Guide for Addressing Accessibility in Standards*, ISO, Geneva, Switzerland, ISO/IEC Guide 71:2014(E), 2014.

[26] A. J. Abbate, E. J. Bass, and A. L. Throckmorton, "A formal task analytic approach to medical device alarm troubleshooting instructions," *IEEE Trans. Human-Mach. Syst.*, vol. 41, no. 1, pp. 53–65, Feb. 2016.

[27] L. Quin, "Extensible markup language 1.0," 1997. [Online]. Available: http://www.w3.org/XML/Core/

[28] C. Sperberg-McQueen and H. Thompson, "The W3C XML schema 1.0," 2001. [Online]. Available: http://www.w3.org/XML/Schema

[29] M. L. Bolton, X. Zheng, K. Molinaro, A. Houser, and M. Li, "Improving the scalability of formal human–automation interaction verification analyses that use task-analytic models," *Innov. Syst. Softw. Eng.*, vol. 13, pp. 1–17, 2017.

[30] C. Baber, "Designing smart objects to support affording situations: Exploiting affordance through an understanding of forms of engagement," *Frontiers Psychol.*, vol. 9, 2018, Art. no. 292.

[31] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Generating phenotypical erroneous human behavior to evaluate human–automation interaction using model checking," *Int. J. Human-Comput. Stud.*, vol. 70, no. 11, pp. 888–906, 2012.

[32] A. J. Abbate and E. J. Bass, "A formal methods approach to semiotic engineering," *Int. J. Human-Comput. Stud.*, vol. 115, pp. 20–39, 2018.

Andrew J. Abbate (S'15–M'17) received the Ph.D. degree in biomedical science from the School of Biomedical Engineering, Science, and Health Systems, Drexel University, Philadelphia, PA, USA, in 2017.

He is currently a Senior Human Factors Engineer with the Pacific Science & Engineering Group, San Diego, CA, USA, where he specializes in visual analytics and human-autonomy integration. His research focuses on developing model-based tools and techniques that can be used to inform human-integrated systems design early in the engineering process.

Ellen J. Bass (M'98–SM'03) received the Ph.D. degree in industrial and systems engineering from the Georgia Institute of Technology, Atlanta, GA, USA.

She is currently a Professor and the Chair of the Department of Health Systems and Sciences Research, College of Nursing and Health Professions, Drexel University, Philadelphia, PA, USA, a Professor with the Department of Information Science, Drexel University's College of Computing and Informatics, and an Affiliate Professor with the Drexel University's School of Biomedical Engineering, Science and Health Systems. She has more than 30 years of human-centered systems engineering research and design experience in air transportation, healthcare, and other domains and the focus of her research is to develop theories of human performance, quantitative modeling methodologies, and associated analysis methods that can be used to evaluate human-system interaction in the context of total system performance.

Formalizing Human–Machine Interactions for Adaptive Automation in Smart Manufacturing

Taejong Joo and Dongmin Shin

***Abstract*—Human–machine interaction is one of the most crucial aspects of advanced manufacturing systems that have advanced to so-called *smart manufacturing systems*. In this regard, this paper presents a framework for formalizing human–machine manufacturing systems. The human–machine system considered in this paper consists of the following three main components: a human supervisor; several cells, each of which is composed of a human operator and a machine; and interfaces. A human operator interacts with a machine in a cell and performs manufacturing tasks based on commands given by the supervisor. Meanwhile, the supervisor is responsible for performing exception handling tasks in response to unanticipated events reported by the cells. With the proposed model, desirable specifications are constructed, which include a condition free of mode confusion, manufacturing task goal reachability, and exception handling task supportability in human–machine manufacturing systems. It is also suggested that adaptive automation with varying levels of information abstraction to humans can be accommodated by the proposed framework. As an illustrative example, we demonstrate the formal models and specifications and the applicability of adaptive automation with a case study of a simple chair assembly system.**

***Index Terms*—Adaptive automation, formal methods, human–machine interaction, human supervisory control, manufacturing system.**

I. Introduction

As SMART manufacturing systems have been recognized as promising factories of the future, the design of adaptive interfaces and the control of human–machine interaction (HMI) systems have become increasingly important issues in human factors/ergonomics (HF/E) and manufacturing systems engineering. In HMI manufacturing systems, machines relieve humans from performing repetitive tasks, and humans can play supervisory roles by focusing on creative and cognitive tasks. Such a collaboration results in improved quality, high reliability, and a flexible workflow [1].

The supervisory control theory (SCT) has gained popularity in the development of the control schemes in discrete-event systems so that the system's operations meet the desired specifications represented by formal language [2]. In other words, a formal model of the system's behavior, called a plant model, and the required specifications are essential to implement a supervisory controller that controls the system's behavior within the range of the desired specifications. Therefore, the possible behaviors of the system and a set of specifications should be unambiguously defined.

Compared to SCT-based computerized systems, the human–machine system is quite different in that it should cater to human operators and supervisors who can address undefined events that occur in unanticipated situations. As such, considering human-involved manufacturing systems, the complex nature of the system makes constructing well-defined formal models of system behaviors a challenging task because human behavior is flexible, nondeterministic, and even unpredictable. For this reason, formalizing a model of HMI systems remains an active research area.

In human–machine systems, *mode confusion* can be caused by an incorrect mental model of the human or a lack of feedback from the user interfaces (UIs) about the underlying states of an automation [3]–[6]. Catastrophes in human–machine systems, such as the accident at Three Mile Island, often result from mode confusion [7], [8]. Therefore, it is important to provide humans with correct and adequate information to prevent mode confusion while filtering out unnecessary information that would impose an excessive information load on the human. In an effort to identify and eliminate possible mode confusion problems in HMI, formal verification methods have been shown to be successful [5], [8], [9].

A human–machine cooperative manufacturing system should be able to complete the given goals (e.g., produce a final product) by exchanging the necessary information. Recently, advanced manufacturing processes have become complicated, which makes it difficult for humans to monitor and memorize all the process specifics. Thus, the appropriate level of task directives should be provided to the operator.

It is also important that the UI provides appropriate information to the human supervisor to respond to unanticipated situations. When an unexpected event is reported from the manufacturing cells, the supervisor should be able to perform responsive tasks by assessing its impact to other cells, and the UI needs to provide the necessary information about the situation.

Although automation has alleviated the need for humans to perform repetitive and tedious tasks, it has also caused humans to be uninformed about problems, which causes reduced human skills, increased mental workloads, and reduced situation awareness [10]–[14]. To mitigate these problems, many studies have

Manuscript received March 16, 2018; revised August 20, 2018 and December 5, 2018; accepted January 20, 2019. Date of publication March 27, 2019; date of current version November 21, 2019. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2017R1A2B4012582). This paper was recommended by Associate Editor Dr. Matthew Bolton. *(Corresponding author: Dongmin Shin.)*

The authors are with Department of Industrial and Management Engineering, Hanyang University, Ansan 15588, South Korea (e-mail: mrjoo512@hanyang.ac.kr; dmshin@hanyang.ac.kr).

Digital Object Identifier 10.1109/THMS.2019.2903402

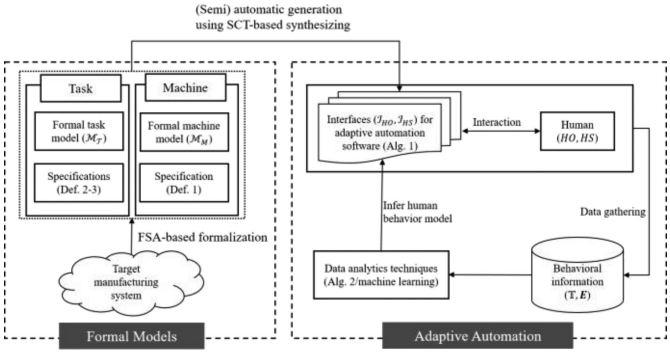


Fig. 1. Overall framework of the research.

been undertaken for adjusting the automation levels, which has resulted in *adaptive automation* [15], [16]. The underlying principle of adaptive automation is that the functional task allocation should be made dynamically in consideration of changes in an operating environment that includes humans [17]. In this regard, the adaptive mechanism for aiding the interactions between the human and the system should be considered an important aspect in designing HMI systems.

This paper aims to develop a framework for constructing formal models of HMI systems and representing their desirable specifications. Specifically, the key components within an HMI system are identified and formalized, and a specification for mode confusion is defined based on the model. In addition, two types of task reachability specifications are constructed that are necessary for the operator to complete a given task and for the supervisor to respond to unexpected events. Furthermore, it is suggested that adaptive automation can be accommodated within the proposed framework by considering different degrees of information abstraction to the human operator.

Another purpose of models is to provide not only a framework for human-involved manufacturing controllers but also a system that can gather the behavioral data of human supervisors and operators. The authors believe that the latter aspect is somewhat more important from the perspective of human-centered manufacturing systems. Algorithms for adaptive automation are suggested in consideration of machine learning techniques, which require sufficient data to train the prediction models. We note that the reasonable amount of behavioral data depends on statistical significance of the prediction levels of the employed machine learning techniques. The overall framework of the research is depicted in Fig. 1.

The rest of this paper is organized as follows. In Section II, we review the previous works. The human–machine system and desired specifications are discussed, and their formal representations are presented in Sections III and IV, respectively. In Section V, an example of the furniture assembly line is provided to illustrate the proposed approach, which is followed by concluding remarks in Section VI.

II. Previous Work

Formal methods have been known to be useful in describing the desired system behaviors in a consistent and unambiguous manner [18], [19]. Among them, a finite-state automata (FSA) has widely been employed in a variety of application domains as a formal model of human-automation interfaces [5], [8], [9], [19]. FSA enables the important properties of HMIs to be specified in a formal manner so that verification can be conducted in a systematic way.

An important characteristic of an FSA-based formal method is that there are known methods for automatically constructing recognition programs, called "lexical analyzers." A lexical analyzer recognizes a set of strings and allows external software modules to be executed upon recognition of the individual symbols in a string. By modifying the structure of an FSA to accommodate discrete-event-system-based models, automatic generation techniques can be adopted for mechanizing the model. In addition, model checking techniques can also be used to verify the desirable specifications.

Mathematically, an FSA $\mathcal{M}$ can be defined as 6-tuple as follows [20]:

$$\mathcal{M} = < \Sigma, Q, \delta, \Gamma, q_{(0)}, F > \tag{1}$$

where Σ is a finite set of events; Q is a finite set of states; $\delta : Q \times \Sigma \to Q$ is a transition function, which is a mapping from a pair of a state and an event to a changed state caused by the event; $\Gamma : Q \to 2^{\Sigma}$ is an active event function, which is a mapping from a state to possible events in the state; $q_{(0)}$ is an initial state; and F is a set of marking states. $\delta(q, \sigma)!$ denotes that a transition from a state $q \in Q$ by an event $\sigma \in \Sigma$ is defined. An extended transition function can be defined based on a string of events such that $\hat{\delta} : Q \times \Sigma^* \to Q$ and $\delta\ (\delta(q, \sigma_1), \sigma_2) = \hat{\delta}\ (q, \sigma_1 \sigma_2)$ for $q \in Q$.

A language is an equivalent representation of an FSA; that is, the language of $\mathcal{M}$ is defined as $\mathcal{L}\ (\mathcal{M}) = \{s \in \Sigma^* |\ \delta(q_{(0)}, s)!\}$ and the accepting language of $\mathcal{M}$ is defined as $\mathcal{L}_m\ (\mathcal{M}) = \{s \in \mathcal{L}(\mathcal{M}) |\ \delta(q_{(0)}, s) \in F\}$ [20]. The prefix closure of language $\mathcal{L}(\mathcal{M})$ is defined as $\overline{\mathcal{L}(\mathcal{M})} = \{s' | s' \text{ is a prefix of } s \wedge s \in \mathcal{L}(\mathcal{M})\}$ where a prefix of a string $s \in \Sigma^*$ is defined as s' such that $s = s'\ w$ for a string $w \in \Sigma^*$. It is known that automaton $\mathcal{M}$ is said to be nonblocking if $\overline{\mathcal{L}_m(\mathcal{M})} = \mathcal{L}(\mathcal{M})$. For more details, the readers are referred to [21].

Formal methods are used to describe the behavior of the human–machine system and to verify whether an HMI system adheres to a specified behavior, enabling the design of a control scheme of HMI systems. Shin *et al.* [19] employ communicating FSAs to describe the system where a human and an automated controller interact by exchanging massages with each other. Additionally, there have been efforts to describe the specifications of the desired behavior of the human–machine systems [8]. For instance, a functional specification of a human material handler is developed in [22] and the mathematically precise definition of mode confusion between the user and the machine in a shared-control system is developed in [23].

Formal verification methodologies have been utilized for verifying the appropriate interactions. Degani and Heymann [5] develop a formal framework for verifying the correctness of a UI. Interface properties for preventing automation surprise based

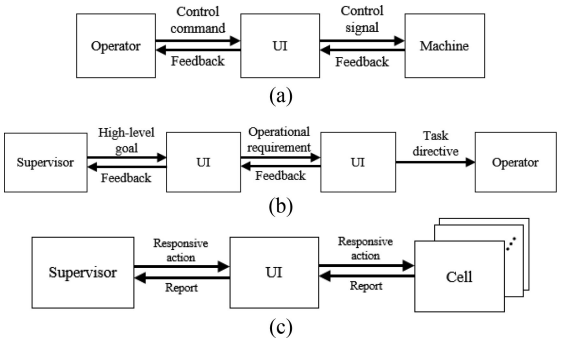


Fig. 2. Description of the human–machine system based on the type of interaction. (a) Operator-machine interaction. (b) Operator-task interaction. (c) Supervisor-system interaction.

on a simulation and bisimulation relationship are proposed in [9]. An extensive survey on the formal verification of human–automation interactions is provided in [8].

In recent years, there have been studies on automatically generating formal specification properties from task models [24], on minimal full control human mental models preventing automation surprise from system models [25], and on interfaces from task analytic behavior models [26], which were previously performed by human experts.

In [27], a method for cooperative supervision is suggested whereby a computer is responsible for the overall safety and the human is responsible for achieving unmodeled goals. In [27], *k*-bounded controllability is developed to accommodate human intervention in the middle of operating an HMI system. A control scheme with two supervisors is proposed in [28], which includes a product supervisor for routing based on the specifications of customized products and the resource capabilities and a resource supervisor for managing the operations to execute the plan of the product supervisor.

III. Informal Description of Human–Machine Systems

A. Interactions Among the Human, Machine, and Tasks

The structure of the human–machine manufacturing system considered in this paper is based on the human supervisory control model suggested in [29]. The system consists of a human supervisor and several manufacturing cells, each of which is composed of a human operator, a machine, and an interface. HMIs are further elaborated to include three types of interactions, as shown in Fig. 2.

Fig. 2(a) shows a direct HMI wherein the operator manipulates a machine. A human operator also performs the manufacturing and exception handling tasks issued by a supervisor, as depicted in Fig. 2(b). In Fig. 2(c), a human supervisor interacts with the cells to manage any unexpected events reported from the cells with the exception handling tasks. Herein, the supervisor acts as a decision maker with respect to the reported exceptions by assessing the impact of the event and planning the sequence of corresponding actions.

B. Interfaces for Operators and a Supervisor

The human–machine manufacturing system considered in this paper can be viewed from two layers, i.e., a cell layer and a supervisor layer. The interfaces at the cell layer are associated with the UIs for human operators who manipulate the machines to achieve the manufacturing goals and respond to unexpected situations either by reporting unanticipated events to a supervisor or performing the exception handling tasks issued by the supervisor.

Information about the machine state is delivered to a human operator through a UI, and then she/he manipulates the machine directly or issues commands to the machine, which causes a change in the machine state. While performing the actions for the next manufacturing tasks, the operator notifies a supervisor of an unexpected event, such as a nonconforming product, machine failure, or missing parts, and then takes certain actions in response to corrective commands from the supervisor.

This paper considers UIs from two perspectives, i.e., machine-related and task-related UIs. A machine-related UI (UI^{M}) is used to display the machine's state, and the two types of task-related UIs display the manufacturing-related task progress (UI^{T_m}) and exception handling task status (UI^{T_e}).

The interfaces at the supervisor layer are for a human supervising the human–machine manufacturing system. Some examples of the supervisory tasks include deciding the product mixture, assessing the key performance indices, and taking appropriate measures for trouble shooting. In normal situations, the supervisor can monitor the entire system through the UI^{T_m} to maintain the task progress. In the case of unexpected events, the exceptional situations are reported to the supervisor through a UI^{T_e} for responsive tasks and corrective commands to an operator.

As such, a UI^{M} is only related to the cell layer, supporting interactions between a human operator and a machine, whereas the UI^{T_m} and UI^{T_e} are involved in both the supervisor layer and the cell layer for task-related interactions.

C. Specifications for Human–Machine Manufacturing Systems

The specifications refer to the desired conditions or requirements of a system. For designing and operating a human–machine manufacturing system, special care needs to be taken for human operators and a supervisor to satisfy the desired specifications. This paper considers the specifications for a confusion-free mode with the ability to reach the manufacturing task goal and the exception handling tasks to ensure that the system can achieve the goal of manufacturing the products while responding to unanticipated events.

Mode confusion can occur when a gap between the human's mental model of the system state and the actual state of the system exists in the process of HMIs. Leveson *et al.* [30] attribute mode confusion to the authority limit and inconsistent behavior of a machine system. An authority limit is associated with the situation whereby the operator gives a command, but the machine refuses to execute the command. The inconsistent behavior of a machine refers to the situation where the operator obtains different outputs from what he/she expects.

In the human–machine manufacturing system considered in this paper, mode confusion is closely related to the design of UI^M, the machine-related UI. The UI^M should provide a human operator with information about the state of the machine in a way that the commands by an operator are guaranteed to be executed. The commands also need to result in the outcomes that the operator expects from the machine through an interface.

Manufacturing task goal reachability means that an operator should be able to complete the manufacturing tasks in a human–machine manufacturing system. Since a principal goal of a manufacturing system is to make final products, an operator is required to perform the manufacturing tasks that change the state of the product from a raw material into a final product. Furthermore, when an operator performs the manufacturing tasks, it is critical for her/him to be aware of the current progress of tasks. In this regard, the UI^{T_m} needs to be designed so that it can provide an operator with the information about the task progress. This design is an important aspect of recent manufacturing systems and emphasizes the convergence of human and advanced technologies to manufacture a wide range of customized products.

Exception handling task supportability is associated with responsive management concerning the uncertainty inherent in almost every manufacturing system. Therefore, it should be taken into account in designing a human–machine manufacturing system to assist the human operators or a supervisor in coping with an unanticipated situation.

If a machine malfunctions in a cell or a human operator fails to complete a task within a given time period, subsequent cells may need to wait until the machine is repaired or parts currently being processed are rerouted dynamically. In this case, information for the exception handling tasks needs to be provided, and the corresponding commands can be issued from a supervisor through the UI^{T_e}. Due to the extremely high degree of complexity, it is still a challenging problem to resolve exceptions only by means of automated systems, and this is one of the main reasons for human involvement.

D. Toward Adaptive Automation

To ensure the performance of an entire HMI system, it is necessary to cater to the humans' capabilities and limits, which change over time. In contrast to preprogrammed machine systems, humans tend to behave in a nondeterministic and flexible manner by adapting themselves to a changing environment. This means that the design and control of an HMI system need to be adaptive in the sense that the appropriate level of information should be provided to the humans depending on the different system states [15], [17]. This design is closely related to the types and levels of automation [12], [13], [31]. This paper considers this aspect with regard to abstraction of information when a human performs repetitive tasks.

In the proposed framework, the adaptive automation is considered such that a UI gives feedback information to the human operators/supervisor with different levels of abstraction. For instance, the UI can just inform humans whether a machine is working or idle, or it can provide more specific information, such as the previous processes performed and detailed task procedures of the machine, similar to a task manual.

IV. Formal Representations of Human–Machine Manufacturing Systems

In this section, the formal representations of an HMI system and the desired specifications are presented. Based on the structural model of an HMI system with FSA, the specifications of the human-manufacturing system are described, and their implications for adaptive automation are further discussed.

In the model, $\Sigma_{i,M}^{T_m}$ and $\Sigma_{i,M}^{T_e}$ indicate sets of observable manufacturing operations and exception reporting actions, respectively, by a machine in cell *i*. For a human operator in cell *i*, $\Sigma_{i,\mathrm{HO}}^{T_m}$, $\Sigma_{i,\mathrm{HO}}^{M}$, and $\Sigma_{i,\mathrm{HO}}^{T_e}$ denote the actions for manufacturing tasks, commands to the machine, and reporting actions for exceptional events to a supervisor, respectively. Similarly, for a human supervisor, $\Sigma_{\mathrm{HS}}^{T_m}$ and $\Sigma_{\mathrm{HS}}^{T_e}$ are the commands associated with manufacturing and exception handling tasks, respectively. Collectively, we refer to a set of events associated with a human operator in cell i as $\Sigma_{i,\mathrm{HO}} = \Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,\mathrm{HO}}^{T_e}$ and with a human supervisor as $\Sigma_{\mathrm{HS}} = \Sigma_{\mathrm{HS}}^{T_m} \cup \Sigma_{\mathrm{HS}}^{T_e}$.

A. Formal Representations of Components in HMI

1) Cell Level Components: The machine model in cell *i*, which describes the machine's responses to the operator's commands, is represented by the following:

$$\mathcal{M}_{i,M} = \left\langle \Sigma_{i,M} \cup \Sigma_{i,\mathrm{HO}}^{M}, Q_{i,M}, \delta_{i,M}, \Gamma_{i,M}, q_{(0)_{i,M}}, F_{i,M} \right\rangle \tag{2}$$

where $\Sigma_{i,M} \cup \Sigma_{i,\mathrm{HO}}^{M}$ is a set of events such that $\Sigma_{i,M} = \Sigma_{i,M}^{T_m} \cup \Sigma_{i,M}^{T_e}$; $Q_{i,M}$ is a finite set of states of the machine; $\delta_{i,M} : Q_{i,M} \times (\Sigma_{i,M} \cup \Sigma_{i,\mathrm{HO}}^{M}) \to Q_{i,M}$ is a state transition function; and $\Gamma_{i,M} : Q_{i,M} \to 2^{(\Sigma_{i,M} \cup \Sigma_{i,\mathrm{HO}}^{M})}$ is an active event function that indicates the possible machine operations and commands that the human operator can execute at a certain machine state.

The task model represents the task progress, which is changed either by the operator's actions or the machine's operations. For the two types of tasks, i.e., manufacturing tasks (T_m) and exception handling tasks (T_e), the task models take the same form, as follows:

$$\mathcal{M}_{\mathcal{T}} = \left\langle \Sigma_{\mathcal{T}}, Q_{\mathcal{T}}, \delta_{\mathcal{T}}, \Gamma_{\mathcal{T}}, q_{(0)_{\mathcal{T}}}, F_{\mathcal{T}} \right\rangle \tag{3}$$

for the manufacturing task model if $\mathcal{T} = T_m$ and for the exception handling task model if $\mathcal{T} = T_e$. Each of the tuples is defined as the same as in an FSA model; F_{T_m} and F_{T_e} are the final state sets associated with the completion of the manufacturing tasks and recovery from an exceptional situation to a normal situation, respectively. For local behaviors associated with T_m and T_e in cell i, we denote $\mathcal{M}_{i,T_m} = P_{\Sigma_{\mathcal{T}} \to (\Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M})}(\mathcal{M}_{T_m})$ and $\mathcal{M}_{i,T_e} = P_{\Sigma_{\mathcal{T}} \to (\Sigma_{\mathrm{HS}}^{T_e} \cup \Sigma_{i,\mathrm{HO}} \cup \Sigma_{i,M})}(\mathcal{M}_{T_e})$, meaning that these models are parts of $\mathcal{M}_{T_m}$ and $\mathcal{M}_{T_e}$.

A process of completing a task, either manufacturing or exception handling, consists of all the possible sequences of operations or actions, collectively called events, toward the final state of the task. A specific sequence of events of task $\mathcal{T}$ up to a

specific point of time, e.g., until the occurrence of the mth event, is represented by $\mathcal{B}_{\mathcal{T}}^{(m)} = \sigma_1\, \sigma_2 \cdots \sigma_m,\ \sigma_k \in \Sigma_{\mathcal{T}}\ \forall k \in [1, m]$, and it is said to be a proper running of an HMI system to reach the final state if $\mathcal{B}_{\mathcal{T}}^{(m)} \in \overline{\mathcal{L}_m(\mathcal{M}_{\mathcal{T}})}$.

The user model of a human operator for task $\mathcal{T}$ is associated with the behavior of a human operator toward the achievement of the task in a cell. For task $\mathcal{T}$ in cell i, it is defined as follows:

$$\mathcal{M}_{i,\mathrm{HO}}^{\mathcal{T}} = \left\langle \Sigma_{i,\mathrm{HO}}^{\mathcal{T}}, Q_{i,\mathrm{HO}}^{UI^{\mathcal{T}}}, \delta_{i,\mathrm{HO}}^{\mathcal{T}}, \Gamma_{i,\mathrm{HO}}^{\mathcal{T}}, {q_{(0)}}_{i,\mathrm{HO}}^{\mathcal{T}}, F_{i,\mathrm{HO}}^{\mathcal{T}} \right\rangle \tag{4}$$

where $\Sigma_{i,\mathrm{HO}}^{\mathcal{T}}$ is a set of events related to task $\mathcal{T}$, and it is defined by $\Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M}$ if $\mathcal{T} = T_m$, $\Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M}$ if $\mathcal{T} = T_m$, and $\Sigma_{i,M} \cup \Sigma_{i,\mathrm{HO}} \cup \Sigma_{\mathrm{HS}}^{T_e}$ if $\mathcal{T} = T_e$. $Q_{i,\mathrm{HO}}^{\mathrm{UI}^{\mathcal{T}}}$ is the set of states of task $\mathcal{T}$ offered by the UI. The set $Q_{i,\mathrm{HO}}^{\mathrm{UI}^{\mathcal{T}}}$ is a subset of $Q_{i,\mathcal{T}}$, meaning that the interface provides a part of the state-related information to the operator. $\delta_{i,\mathrm{HO}}^{\mathcal{T}} : Q_{i,\mathrm{HO}}^{\mathrm{UI}^{\mathcal{T}}} \times \Sigma_{i,\mathrm{HO}}^{\mathcal{T}} \to Q_{i,\mathrm{HO}}^{\mathrm{UI}^{\mathcal{T}}}$ is a state transition function and $\Gamma_{i,\mathrm{HO}}^{\mathcal{T}} : Q_{i,\mathrm{HO}}^{\mathrm{UI}^{\mathcal{T}}} \to 2^{\Sigma_{i,\mathrm{HO}}^{\mathcal{T}}}$ is an active event function. $F_{i,\mathrm{HO}}^{\mathcal{T}}$ is the set of final states of task $\mathcal{T}$.

Based on the user model for task $\mathcal{T}$ defined earlier, a human operator in cell i is considered to be a combination of the three models, such that $\mathcal{M}_{i,\mathrm{HO}} = \mathcal{M}_{i,\mathrm{HO}}^{M} \times \mathcal{M}_{i,\mathrm{HO}}^{T_m} \times \mathcal{M}_{i,\mathrm{HO}}^{T_e}$. Under the current states of the machine, manufacturing tasks, and exception handling tasks, $\mathfrak{q} = (q_{i,\mathrm{HO}}^{M} \in Q_{i,\mathrm{HO}}^{\mathrm{UI}^{M}}, q_{i,\mathrm{HO}}^{T_m} \in Q_{i,\mathrm{HO}}^{\mathrm{UI}^{T_m}}, q_{i,\mathrm{HO}}^{T_e} \in Q_{i,\mathrm{HO}}^{\mathrm{UI}^{T_e}})$, and through the interfaces, the possible machine operations and human actions are identified by active event functions. Depending on the executed machine operations or human actions, the resulting state changes of the system are then defined by the state transition functions.

The states of the machines and task progresses are described in terms of information provided by the interfaces, e.g., UI^{M}, UI^{T_m}, and UI^{T_e}. In this regard, the projection operators in an FSA can be adopted to describe the various information abstraction levels of the interfaces in relation to different levels of automation.

We refer to a set of all the events of discourse in these models as Σ. Therefore, the information abstraction of the interfaces can be described as $P_{\Sigma\to\Sigma'}(\mathcal{L}_m(\mathcal{M}))$ for model $\mathcal{M}$, or equivalently, $P_{\Sigma\to\Sigma'}(Q)$ for a set of states Q in $\mathcal{M}$. Furthermore, the level of abstraction can be assessed by the difference between Σ and Σ', such as $|\Sigma - \Sigma'|$. This abstraction can be useful in investigating how much information about the actual states of an HMI system is to be provided to an operator. For instance, a small difference between Σ and Σ' of the $\mathcal{M}_{i,\mathrm{HO}}^{M}$ model may indicate that a fine-grained description of the machine states is given to an operator.

The interface models are then described as follows:

$$\mathcal{I}_{i,\mathrm{HO}}^{M} \subseteq Q_{i,M} \times Q_{i,\mathrm{HO}}^{\mathrm{UI}^{M}} \tag{5}$$

$$\mathcal{I}_{i,\mathrm{HO}}^{T_m} \subseteq Q_{i,T_m} \times Q_{i,\mathrm{HO}}^{\mathrm{UI}^{T_m}} \tag{6}$$

$$\mathcal{I}_{i,\mathrm{HO}}^{T_e} \subseteq Q_{i,\,T_e} \times Q_{i,\mathrm{HO}}^{\mathrm{UI}^{T_e}}. \tag{7}$$

As such, the whole UI for a human operator in cell i can be defined as $\mathcal{I}_{i,\mathrm{HO}} = \mathcal{I}_{i,\mathrm{HO}}^{M} \times \mathcal{I}_{i,\mathrm{HO}}^{T_m} \times \mathcal{I}_{i,\mathrm{HO}}^{T_e}$.

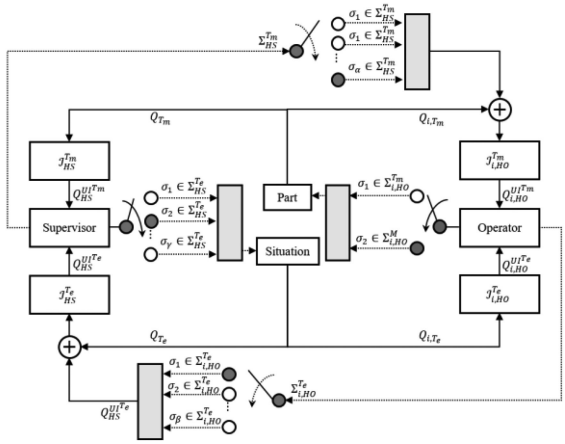


Fig. 3. Interaction between a human supervisor and a human operator. When a supervisor issues the command of a manufacturing task $\Sigma_{\mathrm{HS}}^{T_m}$, an operator performs the manufacturing task by executing either $\Sigma_{i,\mathrm{HO}}^{T_m}$ or $\Sigma_{i,\mathrm{HO}}^{M}$. Conversely, when an operator reports an unanticipated event $\Sigma_{i,\mathrm{HO}}^{T_e}$, the supervisor performs the exception handling task by executing one of the events in $\Sigma_{\mathrm{HS}}^{T_e}$. In the figure, the solid and dotted arrows represent information flows about events and states, respectively. The shaded rectangles represent the execution of events. The plus sign symbol inside a circle represents the composition of information.

2) Supervisor Level Components: The user model of a human supervisor for task $\mathcal{T}$ is associated with the behavior of a human supervisor toward the achievement of the task in the system, and it is defined as follows:

$$\mathcal{M}_{\mathrm{HS}}^{\mathcal{T}} = \left\langle \Sigma_{\mathrm{HS}}^{\mathcal{T}}, Q_{\mathrm{HS}}^{\mathrm{UI}^{\mathcal{T}}}, \delta_{\mathrm{HS}}^{\mathcal{T}}, \Gamma_{\mathrm{HS}}^{\mathcal{T}}, {q_{(0)}}_{\mathrm{HS}}^{\mathcal{T}}, F_{\mathrm{HS}}^{\mathcal{T}} \right\rangle \tag{8}$$

where $\Sigma_{\mathrm{HS}}^{\mathcal{T}}$ is a set of events associated with task $\mathcal{T}$, defined by $\cup_{i\in I}(\Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M}^{T_m}) \cup \Sigma_{\mathrm{HS}}^{T_m}$ if $\mathcal{T} = T_m$ and $\cup_{i\in I}(\Sigma_{i,\mathrm{HO}}^{T_e} \cup \Sigma_{i,M}^{T_e}) \cup \Sigma_{\mathrm{HS}}^{T_e}$ if $\mathcal{T} = T_e$. $Q_{\mathrm{HS}}^{\mathrm{UI}^{\mathcal{T}}}$ is a set of task states given by the UI. Again, the remaining tuples are defined as in an FSA. Note that the supervisor models correspond to both the manufacturing and exception handling tasks of each cells.

In contrast to the interfaces in the cell level, only two types of interfaces are involved at the supervisor level, and they provide information about the manufacturing task progress and the exception handling progress. Mathematically, these are defined as follows:

$$\mathcal{I}_{\mathrm{HS}}^{T_m} \subseteq Q_{T_m} \times Q_{\mathrm{HS}}^{\mathrm{UI}^{T_m}} \tag{9}$$

$$\mathcal{I}_{\mathrm{HS}}^{T_e} \subseteq Q_{T_e} \times Q_{\mathrm{HS}}^{\mathrm{UI}^{T_e}}. \tag{10}$$

Collectively, the model of a human supervisor consists of two models, i.e., $\mathcal{M}_{\mathrm{HS}} = \mathcal{M}_{\mathrm{HS}}^{T_m} \times \mathcal{M}_{\mathrm{HS}}^{T_e}$, and the UI for a human supervisor is defined as $\mathcal{I}_{\mathrm{HS}} = \mathcal{I}_{\mathrm{HS}}^{T_m} \times \mathcal{I}_{\mathrm{HS}}^{T_e}$. Fig. 3 illustrates the interaction between a human supervisor and a human operator based on the formal representations of an HMI system.

3) Specifications for Human–Machine Manufacturing Systems: The specifications describe what a certain system should satisfy to guarantee the required properties of the system. It tends to become more difficult to ensure specifications as the system becomes larger. This difficulty imposes a challenge on a system

if it is composed of a variety of heterogeneous components interacting with each other. Even worse, failure to satisfy some of the specifications may cause difficulty in achieving the system's ultimate goal(s). Designing and operating human-manufacturing systems also require special specifications to achieve the goals of manufacturing products while handling unexpected situations.

This paper considers three types of specifications for human–machine manufacturing systems and develops formal models to represent them. The formal models of these specifications can enable the validation of the HMI systems and, more importantly, synthesize the computerized system controllers by imbedding the presented models into the process of developing computerized controllers of the manufacturing systems.

Definition 1 (Mode confusion): For an interaction between a human operator and a machine $(\mathcal{M}_{i,M},\ \mathcal{M}^M_{i,\mathrm{HO}}, \mathcal{I}^M_{i,\mathrm{HO}})$, the system is said to be free of mode confusion if the following conditions hold:

1) $\forall q_M \in Q_{i,M} \wedge q^M_{\mathrm{HO}} \in Q^{\mathrm{UI}^M}_{i,\mathrm{HO}}$ s.t. $(q_M, q^M_{\mathrm{HO}}) \in \mathcal{I}^M_{i,\mathrm{HO}}$, $\sigma \in (\Gamma^M_{i,\mathrm{HO}}(q^M_{\mathrm{HO}}) \cap \Sigma^M_{i,\mathrm{HO}}) \Rightarrow \sigma \in \Gamma_{i,M}(q_M)$.
2) $\forall q^M_{\mathrm{HO}} \in Q^{\mathrm{UI}^M}_{i,\mathrm{HO}} \wedge \sigma \in \Sigma^M_{i,\mathrm{HO}}, |\delta^M_{i,\mathrm{HO}}(q^M_{\mathrm{HO}}, \sigma)| = 1$.

Condition 1, corresponding to the *authority limit* in [30], implies that an operator can execute only the events displayed by the interface that are executable by the machine at its current state and condition 2, called *consistent behavior*, dictates that an outcome of the human operator's command to a machine through the interface should be the same as he or she expects.

It should be noted that condition 2 cannot be explicitly evaluated only with the specification itself because it does not model what the human operator expects, which is an internal process of the human. Rather, it can be evaluated after the human behavioral log data and the corresponding operational data from machines and interfaces are analyzed.

Definition 2 (Manufacturing task goal reachability): For an interaction between a human operator and manufacturing tasks $(\mathcal{M}_{i,T_m}, \mathcal{M}^{T_m}_{i,\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}})$, the manufacturing task goal reachability from q to q', denoted by $q \mapsto q'$, holds if for $q, q' \in Q^{\mathrm{UI}^{T_m}}_{i,\mathrm{HO}}$ $\exists s \in (\Sigma^{T_m}_{i,\mathrm{HO}} \cup \Sigma^M_{i,\mathrm{HO}} \cup \Sigma_{i,M})^*$ such that $\hat{\delta}^{T_m}_{i,\mathrm{HO}}(q, s) = q'$.

Definition 2 indicates that a human operator can proceed with the manufacturing task through a sequence of events, each of which is executed by manual operations, issuing commands, or observing machine operations.

Theorem 1: The manufacturing task issued by a supervisor's command $\sigma \in \Sigma^{T_m}_{\mathrm{HS}}$ can be performed if $\exists\{q_t : q_t \in Q^{\mathrm{UI}^{T_m}}_{i,\mathrm{HO}} \wedge q_{t-1} \mapsto q_t\}^T_{t=1}$ where $(\delta_{i,T_m}(q_{(0)_{T_m}}, \sigma), q_0) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$ and $q_T \in F_{T_m}$.

Proof: Let $\sigma \in \Sigma^{T_m}_{\mathrm{HS}}$ be any command of a supervisor and take q_0 such that $(\delta_{i,T_m}(q_{(0)_{T_m}}, \sigma), q_0) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$. Then, by the definition of $\mathcal{M}^{T_m}_{i,\mathrm{HO}}$, if σ can be performed, $\exists s_{T_m} \in (\Sigma^{T_m}_{i,\mathrm{HO}} \cup \Sigma^M_{i,\mathrm{HO}} \cup \Sigma_{i,M})^*$ such that $\hat{\delta}^{T_m}_{i,\mathrm{HO}}(q_0, \sigma s_{T_m}) \in F^{T_m}_{i,\mathrm{HO}}$. Then, we can recursively find s_t and q_t for $t \in [1, T]$ such that $s_t = \arg\min_s |\{\sigma_j \in (\Sigma^{T_m}_{i,\mathrm{HO}} \cup \Sigma^M_{i,\mathrm{HO}} \cup \Sigma_{i,M}) : \hat{\delta}^{T_m}_{i,\mathrm{HO}}(q_{t-1}, \sigma_1 \cdots \sigma_J) \neq q_{t-1} \wedge s_1 \cdots s_{t-1}\sigma_1 \cdots \sigma_J \in \overline{s_{T_m}}\}^J_{j=1}|$ and $q_t = \hat{\delta}^{T_m}_{i,\mathrm{HO}}(q_{t-1}, s_t)$ until $q_T = \hat{\delta}^{T_m}_{i,\mathrm{HO}}(q_0, \sigma s_{T_m}) \in F^{T_m}_{i,\mathrm{HO}}$. Therefore, by definition 2, $q_{t-1} \mapsto q_t$ is satisfied for all t, which shows that σ can be realized if $\exists\{q_t : q_t \in Q^{\mathrm{UI}^{T_m}}_{i,\mathrm{HO}} \wedge q_{t-1} \mapsto q_t\}^T_{t=1}$.

While Definition 2 is about a single transition between two states, Theorem 1 is about a sequence of transitions between the final state and a state where a command from a human supervisor is given. Specifically, the supervisor *activates* a manufacturing task model by giving a command ($\sigma \in \Sigma^{T_m}_{\mathrm{HS}}$), which *activates* a transition from an initial state ($q_{(0)_{T_m}}$) to another (q_0), and Theorem 1 presents the condition that the HMI system can eventually reach the final state ($q_T \in F^{T_m}_{i,\mathrm{HO}}$) from the activated state. The condition recursively guarantees the supportability in that each successive transition eventually leads to reaching the final state.

Definition 3 (Exception handling task supportability): For an interaction between a supervisor and exception handling tasks for responsive management $(\mathcal{M}_{T_e}, \mathcal{M}^{T_e}_{\mathrm{HS}}, \mathcal{I}^{T_e}_{\mathrm{HS}})$, the supervisor exception handling task supportability from q to q', denoted by $q \rightarrowtail q'$, holds if for $q, q' \in Q^{\mathrm{UI}^{T_e}}_{\mathrm{HS}}$ such that $q' \neq q$, $\exists s \in (\cup_{i\in I}(\Sigma^{T_e}_{i,\mathrm{HO}} \cup \Sigma^{T_e}_{i,M}) \cup \Sigma^{T_e}_{\mathrm{HS}})^*$ such that $\hat{\delta}^{T_e}_{\mathrm{HS}}(q, s) = q'$.

Theorem 2: The exception handling task issued by reporting the events of the operators and machines on the shop floor $\sigma \in \cup_{i\in I}(\Sigma^{T_e}_{i,\mathrm{HO}} \cup \Sigma^{T_e}_{i,M})$ can be resolved if $\exists\{q_t : q_t \in Q^{UI^{T_e}}_{HS} \wedge q_{t-1} \rightarrowtail q_t\}^T_{t=1}$, where $(\delta_{T_e}(q_{(0)_{T_e}}, \sigma), q_0) \in \mathcal{I}^{T_e}_{\mathrm{HS}}$ and $q_T \in F_{T_e}$.

Proof: Let $\sigma \in \cup_{i\in I}(\Sigma^{T_e}_{i,\mathrm{HO}} \cup \Sigma^{T_e}_{i,M})$ be any reporting event incurred by an operator or a machines command of a supervisor and take q_0 such that $(\delta_{T_e}(q_{(0)_{T_e}}, \sigma), q_0) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$. Then, based on the proof of Theorem 1, it is can be shown that σ can be achieved if $\exists\{q_t : q_t \in Q^{UI^{T_e}}_{HS} \wedge q_{t-1} \rightarrowtail q_t\}^T_{t=1}$.

It should be noted that the definitions of the HMI system specifications are based on how the UIs provide humans with information about the actions and operations performed by the system components rather than inferring the cognitive states of the humans. This basis can be useful in analyzing the procedural knowledge of humans through the behavioral records of humans and machines while performing manufacturing tasks. More importantly, these definitions can serve in the automatic generation of interfaces, as in [9], in consideration of the desirable specifications of HMI systems.

4) Toward Adaptive Automation: To accommodate an adaptive automation that satisfies the specifications defined earlier, the state transition functions and active event functions are considered to be key factors. The interfaces for a human operator/supervisor are to be designed so that they can provide the system state information dynamically in accordance with the outcomes collected from the operation of the HMI system.

In the presented model, the state transition function and the active event function are exploited to record the actions that the human operators and the supervisor have taken among the possible actions. The recoded action data then compose the empirical state transition function, denoted by $\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}$, and the empirical active event function, denoted by $\Gamma^{T_m}_{i,\widehat{\mathrm{HO}}}$. From these empirical models, the functions defined for the $Q^{\mathrm{UI}^{T_m}}_{i,\mathrm{HO}}$ can be obtained by defining the transition $\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}(q, s) = q'$ if $\forall q_u \exists(q_v)$ such that $\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}(q_u, s) = q_v$, where $(q_u, q) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$ and $(q_v, q') \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$.

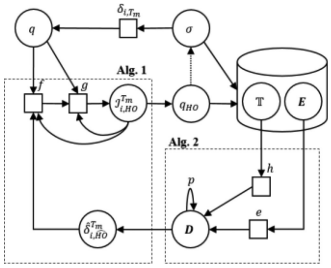


Fig. 4. Simplified graphical illustration of the implementation of adaptive automation in Algorithm 3. The circles and squares represent variables and functions, respectively. The solid arrows represent input–output relationships. A dashed arrow indicates an internal decision procedure of a human operator.

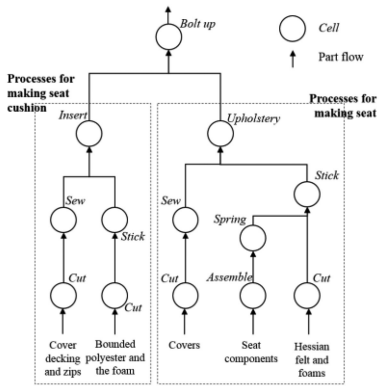


Fig. 5. Chair assembly process.

This relationship is encoded as $\hat{\delta}^{T_m}_{i,\mathrm{HO}} = \mathcal{P}(\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}})$. It means that the information provided by the interface is updated corresponding to the tasks that a human operator has performed.

The procedure of designing an adaptive interface for a human operator, denoted by $\mathcal{I}^{T_m}_{i,\mathrm{HO}}$, is described in Algorithm 1. Algorithm 1 is designed to adjust the projection operator so that the task-related information provided by the interfaces changes dynamically as manufacturing and exception handling tasks are processed.

We note that the interface is designed to be adaptive to the user model of an operator toward manufacturing tasks, which is *inferred* based on his/her behavioral data and is summarized by $\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}$. Algorithm 2 shows the procedure for the inference of the manufacturing task model of a human operator. For a human supervisor, the empirical models can also be obtained by replacing $\sigma \in \Sigma^{T_m}_{\mathrm{HS}}$, F_{T_m}, and $\mathcal{M}^{T_m}_{i,\mathrm{HO}}$ with $\sigma \in \cup_{i\in I}(\Sigma^{T_e}_{i,\mathrm{HO}} \cup \Sigma^{T_e}_{i,M})$, F_{T_e}, $\mathcal{M}^{T_e}_{\mathrm{HS}}$, and $\mathcal{I}^{T_m}_{i,\mathrm{HO}}$, respectively.

Based on Algorithms 1 and 2, the adaptive automation can support the human operator so that the specification of manufacturing task supportability holds under the dynamically changing capability of the operator. Algorithm 3 depicts this overall procedure, and Fig. 4 illustrates the relationship between the three algorithms with their key variables.

Algorithm 1: Adaptive interface design for manufacturing task goal reachability. $\mathcal{B}_{T_m}$ is a string of events that occurred in the system from $q_{(0)_{T_m}}$ to q. f is a function that counts the number of states in which the manufacturing task goal reachability is satisfied under the given conditions of UI such that $f\ (q_{\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}, \hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{B}_{T_m}) = |\{q|\ q_{\mathrm{HO}} \mapsto q \wedge \hat{\delta}^{T_m}_{i,\mathrm{HO}}(q_{\mathrm{HO}}, s) = q \wedge \hat{\delta}^{T_m}_{i,\mathrm{HO}} = \mathcal{P}(\hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}) \wedge \mathcal{B}_{T_m} s \in \overline{\mathcal{L}_m(\mathcal{M}_{i,T_m})}\}|$. g is a function with the output of the number of states of a task model whose projection is equal to q's projection given the interface $\mathcal{I}$ such that $g\ \ (q, \mathcal{I}^{T_m,(k)}_{i,\mathrm{HO}}) = |\{q' \in Q_{i,T_m} : p \in Q^{\mathrm{UI}^{T_m},(k)}_{i,\mathrm{HO}} \wedge (q,p) \in \mathcal{I}^{T_m,(k)}_{i,\mathrm{HO}} \wedge (q',p) \in \mathcal{I}^{T_m,(k)}_{i,\mathrm{HO}}\}|$.

Input: $\mathcal{M}_{i,T_m}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}, q, q_{\mathrm{HO}}, \mathcal{B}_{T_m}, \hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}$
Output: $q_{\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}$
01: **if** $f\ (q_{\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}, \hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{B}_{T_m}) = 0$ **then**
02: **while** $f(q_{\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}, \hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{B}_{T_m}) < 1$ **do**
03: select $\mathcal{I}^{T_m,prime}_{i,\mathrm{HO}}$ such that $g(q, \mathcal{I}^{T_m,'}_{i,\mathrm{HO}}) < g(q, \mathcal{I}^{T_m}_{i,\mathrm{HO}})$
04: $\mathcal{I}^{T_m}_{i,\mathrm{HO}} \leftarrow \mathcal{I}^{T_m,'}_{i,\mathrm{HO}}$
05: set q_{HO} such that $(q, q_{\mathrm{HO}}) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$
06: **else**
07: **while** $f(q_{\mathrm{HO}}, \mathcal{I}^{T_m}_{i,\mathrm{HO}}, \hat{\delta}^{T_m}_{i,\widehat{\mathrm{HO}}}, \mathcal{B}_{T_m}) \geq 1$ **do**
08: select $\mathcal{I}^{T_m,'}_{i,\mathrm{HO}}$ such that $g(q, \mathcal{I}^{T_m,'}_{i,\mathrm{HO}}) < g(q, \mathcal{I}^{T_m}_{i,\mathrm{HO}})$
09: $\mathcal{I}^{T_m}_{i,\mathrm{HO}} \leftarrow \mathcal{I}^{T_m,'}_{i,\mathrm{HO}}$
10: set q_{HO} such that $(q, q_{\mathrm{HO}}) \in \mathcal{I}^{T_m}_{i,\mathrm{HO}}$

V. Illustrative Example

In this section, the proposed approach is applied to a chair assembly system described in [32] (see Fig. 5). In this system, parts are processed through eight cells, each of which consists of a human operator and a machine, from the cutting to assembly processes to collaboratively achieve the goals of chair assembly and cushion processing. Specifically, in each cell, a human operator performs the tasks of loading/unloading parts, starting a machine process, and moving parts to the subsequent cell as per the supervisor's commands.

A human supervisor is responsible for assigning parts to the cells and for responding to machine breakdowns, which are unanticipated events reported from the cells. Specifically, in case of a machine breakdown in the middle of a manufacturing task, the operator also reports the situation to the supervisor; the supervisor then sends a command for a repairing task and notifies the other cells of the breakdown with a changed part routing. The event list associated with the system is shown in Table I.

It should be noted that the case-study model is intended to illustrate that a human-involved manufacturing system can be represented by a formal model. As a system becomes more complex in term of increasing numbers of tasks, desirable specifications and human operators and supervisors, the algorithms

TABLE I
SYMBOLIC DESCRIPTION OF THE EVENTS USED IN THE EXAMPLE

Symbol	Event
$\alpha^{i,T_m}(p) \in \Sigma_{i,M}^{T_m}$	Process part p by the machine in cell i
$\beta_1^{i,T_m}(p) \in \Sigma_{i,HO}^{T_m}$	Load part p to the machine in cell i
$\beta_2^{i,T_m}(p) \in \Sigma_{i,HO}^{T_m}$	Start the machine in cell i to process part p
$\beta_3^{i,T_m}(p,j) \in \Sigma_{i,HO}^{T_m}$	Move part p to cell j from cell i
$\beta_1^{i,T_e} \in \Sigma_{i,HO}^{T_e}$	Machine breakdown in cell i
$\beta_2^{i,T_e} \in \Sigma_{i,HO}^{T_e}$	Repair performed at cell i
$\gamma_1^{T_m}(g,i) \in \Sigma_{HS}^{T_m}$	Command to set a goal of cell i to be g
$\gamma_1^{T_e}(i) \in \Sigma_{HS}^{T_e}$	Command to repair the machine in cell i
$\gamma_2^{T_e}(i \to j) \in \Sigma_{HS}^{T_e}$	Notify cell j of a machine breakdown in cell i
$\gamma_3^{T_e}(i \to j) \in \Sigma_{HS}^{T_e}$	Notify cell j of a machine repair in cell i

Algorithm 2: Procedure for obtaining empirical models $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}$ and $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m}$. $\mathbb{T}^{(r)}$ is a set of tuples, each of which contains two states and a string, where the string enables a transition from one state to another state during the r-th repetition. $\boldsymbol{E} = \{e_{q\sigma}\}$ is a matrix of $|Q_{i,T_m}| \times |\Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M}|$ where $e_{q\sigma}$ is the number of errors that occurred in state q by event σ. Parameters m and p control the memory decay and production compilation, respectively. h is a function of inferring the operator's procedural knowledge based on observation and is defined as $h(\boldsymbol{t}) = \{(q \in Q_{i,T_m}, s \in (\Sigma_{i,T_m})^*) : \delta_{i,T_m}(t_1, s) = q \wedge s \in \overline{t_3}\}$. e is a mapping from a pair of a state and an event to a set of events, which is reachable after executing the event in the state and is defined as $e(q,s) = \{q' \in Q_{i,T_m} : \delta_{i,T_m}(q, \sigma s) = q' \wedge s \in (\Sigma_{i,T_m})^*\}$.

Input: $\mathcal{M}_{i,T_m}$, r, $\mathbb{T}^{(1)}, \ldots, \mathbb{T}^{(r)}$, $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(r)}$
Output: $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}$, $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m}$
Parameters: m, p
Initialization: $\mathbb{T} \leftarrow \cup_{k=r-m}^{r} \mathbb{T}^{(k)}$;
$\boldsymbol{E} \leftarrow \sum_{k=r-m}^{r} \boldsymbol{E}^{(k)}$; $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m} \leftarrow \emptyset$; $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m} \leftarrow \emptyset$;
$\boldsymbol{D} \leftarrow \{\emptyset\}_{|Q_{i,T_m}| \times |Q_{i,T_m}|}$
01: **for** $\boldsymbol{t} \in \mathbb{T}$ **do**
02: add s in $\boldsymbol{D}_{t_1,q}$ for all $(q,s) \in h(\boldsymbol{t})$
03: **for** $q, q', q'' \in Q_{i,T_m}$ such that $|d_{qq'}| > p$ and $|d_{q'q''}| > p$ **do**
04: add $s_1 s_2$ in $d_{q,q''}$ for all $(s_1, s_2) \in (d_{qq'}, d_{q'q''})$
05: **for** $(q,\sigma) \in (Q_{i,T_m}, \Sigma_{i,T_m})$ such that $e_{q\sigma}! = 0$ **do**
06: $d_{q,q'} \leftarrow \emptyset$ for all $q' \in e(q,\sigma)$
07: **for** $(q,q') \in (Q_{i,T_m}, Q_{i,T_m})$ such that $|d_{qq'}| > 0$ **do**
08: define $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}(q,t_3) = q'$ for all $\mathbf{t} \in d_{qq'}$
09: add σ in $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m}(q)$ for all $\sigma \in t_3$ and $\mathbf{t} \in d_{qq'}$

should be implemented by a computerized system. However, this case study only presents manual, but systematically verified, models with somewhat simplified tasks.

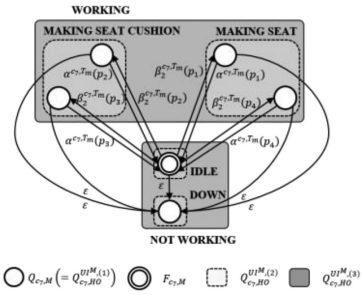


Fig. 6. Graphical representation of a machine model in cell 7 (c_7) and the different levels of information abstraction provided by the interfaces.

Algorithm 3: Adaptive automation for a human operator in performing manufacturing tasks. r and n_r are integers of repetition and the number of transitions in the r-th repetition, respectively. $\mathcal{B}_{T_m}^{(i)}$ and $\mathcal{B}_{T_m}$ denote a sequence of events in i-th transition and all the transitions, respectively.

Input: $\mathcal{M}_{i,T_m}$, $\mathcal{I}_{i,\mathrm{HO}}^{T_m}$, $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}$, $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m}$, r, $\mathbb{T}^{(1)}, \ldots, \mathbb{T}^{(r)}$, $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(r)}$
Output: $\mathcal{I}_{i,\mathrm{HO}}^{T_m}$, $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}$, $\Gamma_{i,\widehat{\mathrm{HO}}}^{T_m}$, r, $\mathbb{T}^{(1)}, \ldots, \mathbb{T}^{(r)}$, $\boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(r)}$
Initialization: $n_r = 0$; $\mathbb{T}^{(r)} \leftarrow \emptyset$;
$\boldsymbol{E}^{(r)} \leftarrow \{0\}_{|Q_{i,T_m}| \times |\Sigma_{i,\mathrm{HO}}^{T_m} \cup \Sigma_{i,\mathrm{HO}}^{M} \cup \Sigma_{i,M}|}$
01: $\sigma \in \Sigma_{\mathrm{HS}}^{T_m}$ occurs; $\mathcal{B}_{T_m}^{(0)} \leftarrow \sigma$; $\mathcal{B}_{T_m} \leftarrow \sigma$
02: $q \leftarrow \delta_{i,T_m}(q_{(0)_{T_m}}, \sigma)$; set q_{HO} such that $(q, q_{\mathrm{HO}}) \in \mathcal{I}_{i,\mathrm{HO}}^{T_m}$
03: **while** $q \in F_{i,T_m}$ **do**
04: $\sigma \in \Sigma_{i,T_m}$ occurs; $\mathcal{B}_{T_m}^{(n_r)} \leftarrow \mathcal{B}_{T_m}^{(n_r)}\sigma$; $\mathcal{B}_{T_m} \leftarrow \mathcal{B}_{T_m}\sigma$
05: **if** $\mathcal{B}_{T_m} \notin \overline{\mathcal{L}_m(\mathcal{M}_{i,T_m})}$ **then**
06: $e_{q\sigma} + = 1$; **break**
07: $q' \leftarrow \delta_{i,T_m}(q,\sigma)$; set q_{cand} such that $(q', q_{\mathrm{cand}}) \in \mathcal{I}_{i,\mathrm{HO}}^{T_m}$
08: **if** $q_{\mathrm{cand}} \neq q_{\mathrm{HO}}$ **then**
09: $q_{\mathrm{cand}}, \mathcal{I}_{i,\mathrm{HO}}^{T_m} \leftarrow$**Alg.1**$(\mathcal{I}_{i,\mathrm{HO}}^{T_m}, q', q_{\mathrm{cand}}, \mathcal{B}_{T_m}, \hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m})$
10: add $(q_{\mathrm{HO}}, q_{\mathrm{cand}}, \mathcal{B}_{T_m}^{(n_r)})$ in $\mathbb{T}^{(r)}$
11: $q_{\mathrm{HO}} \leftarrow q_{\mathrm{cand}}$; $n_r + = 1$; $\mathcal{B}_{T_m}^{(n_r)} \leftarrow \varepsilon$
12: $q \leftarrow q'$
13: $\hat{\delta}_{i,\widehat{\mathrm{HO}}}^{T_m}, \Gamma_{i,\widehat{\mathrm{HO}}}^{T_m} \leftarrow$**Alg.2**$(r, \mathbb{T}^{(1)}, \ldots, \mathbb{T}^{(r)}, \boldsymbol{E}^{(1)}, \ldots, \boldsymbol{E}^{(r)})$

A. Formal Description of the System

The machine model of each cell is defined as in (2) with $Q_{i,M} = \{\mathrm{IDLE}, \mathrm{WORKING(Op)}, \mathrm{DOWN}\}$ where Op denotes an operation allocated to a machine in cell i. As an example, Fig. 6 illustrates the machine model of cell 7, whose responsible task is a cutting task.

There are two manufacturing task models, i.e., $T_m =$ seat assembly, cushion processing, each of which corresponds to seat assembling and seat cushion processing in the form of (3), where $\Sigma_{T_m} = \cup_{i \in \{1,2,\ldots,8\}} \{\Sigma_{i,\mathrm{HO}}^{T_m} \cup$

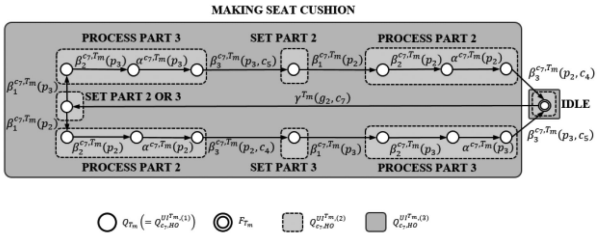


Fig. 7. Graphical representation of the manufacturing task of seat cushion processing.

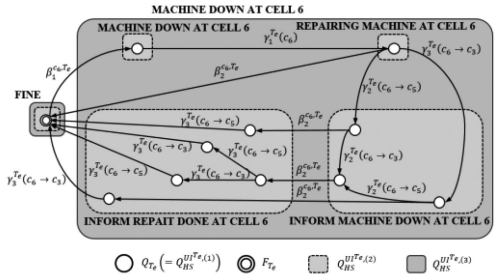


Fig. 8. Graphical representation of an exception handling task of machine breakdown.

$\Sigma^{M}_{i,\mathrm{HO}} \cup \Sigma^{T_m}_{i,M}\} \cup \Sigma^{T_m}_{\mathrm{HS}}$; Q_{T_m} is defined such that each of the states represents specific progress of the manufacturing task and $q_{(0)T_m} = F_{T_m} = \{\mathrm{IDLE}\}$ represents the completion of a seat assembly task. Fig. 7 depicts a manufacturing task model in cell 7.

The exception handling task model can also be defined as in (3), where $\Sigma_{T_e} = \cup_{i\in\{1,2,\ldots,8\}}(\Sigma^{T_e}_{i,\mathrm{HO}} \cup \Sigma^{T_e}_{i,M}) \cup \Sigma^{T_e}_{\mathrm{HS}}$; Q_{T_e} is defined as a set of progress levels in repairing a breakdown such that $Q_{T_e} = \{\mathrm{FINE}\} \cup \cup_{i\in\{1,2,\ldots,8\}}\{\mathrm{REPORTED\_DOWN}(c_i),\ \mathrm{REPAIR}(m_i)\} \cup \cup_{i\in\{1,2,\ldots,8\}}\{\mathrm{NOTIFY\_DOWN}(c_j), \mathrm{NOTIFY\_REPAIR\_DOWN}(c_j)\}$. Fig. 8 depicts a part of the machine breakdown handling task model.

The user models of a human operator toward a machine, a manufacturing task, and an exception handling task, each of which describes the operator's task performing process for the corresponding components, are defined by (4). Herein, $Q^{\mathrm{UI}^M}_{i,\mathrm{HO}}$, $Q^{\mathrm{UI}^{T_m}}_{i,\mathrm{HO}}$, and $Q^{\mathrm{UI}^{T_e}}_{i,\mathrm{HO}}$ are finite sets of the abstracted states of a machine and two types of tasks offered by the interfaces, respectively. These variables are represented as the shaded parts in Figs. 6 and 7.

The user model of a human supervisor toward a manufacturing task and an exception handling event, which dictates the supervisor's tasks, are defined by (5) and (6). Herein, $Q^{\mathrm{UI}^{T_m}}_{\mathrm{HS}}$ and $Q^{\mathrm{UI}^{T_e}}_{\mathrm{HS}}$ are finite sets of states of manufacturing and exception handling task models, respectively, offered by the interfaces of a human supervisor. Fig. 8 shows a model of a human supervisor in dealing with a machine breakdown in cell 6.

B. Human–Machine Manufacturing System Specifications

1) Mode Confusion: According to Definition 1, the mode confusion in HMIs can be evaluated by comparing the outputs of a state transition function and an active event function of a machine and a user model. Table II summarizes the results of evaluating the mode confusion in c_7 (see Fig. 6) depending on the three different interfaces. If a human operator is working with interface ($(q_M, q^M_H) \in \mathcal{I}^{M,(3)}_{c_7,\mathrm{HO}}$), mode confusion can occur in the form of an authority limit. When the machine is in a breakdown state ($q_M = \mathrm{DOWN}$) rather than in an idle state, the interface indicates only that the machine is not working. The operator may expect the machine to be idle ($q^M_H = \mathrm{NOT\ WORKING}$) and issues a command to process a part ($\sigma = \beta^{c_7,T_m}_2(p_1)$), which is rejected by the *broken* machine ($\sigma \notin \Gamma_{c_7,M}(q_M)$).

TABLE II
RESULTS OF THE MODE CONFUSION CONDITIONS FOR THREE INTERFACES

Interface	$\Gamma^M_{i,HO}(q^M_{HO}) \cap \Sigma^M_{i,HO} \subseteq \Gamma_{i,M}(q_M)$	$\lvert\delta^M_{i,HO}(q^M_{HO},\sigma)\rvert = 1$
$\mathcal{I}^{M,(1)}_{c_7,HO}$	O	O
$\mathcal{I}^{M,(2)}_{c_7,HO}$	O	O
$\mathcal{I}^{M,(3)}_{c_7,HO}$	O	X

TABLE III
TRANSITIONS REQUIRED FOR MANUFACTURING TASK REACHABILITY FOR THE THREE INTERFACES

Interface	Necessary transitions for holding Theorem 1
$\mathcal{I}^{T_m,(1)}_{c_7,HO}$	$\sigma \in \Sigma_{T_m}$
$\mathcal{I}^{T_m,(2)}_{c_7,HO}$	$\breve{\alpha}^{c_7}_2(p)$ for $p \in \{p_2, p_3\}$, $\breve{\alpha}^{c_7}_1(p)\breve{\gamma}^{c_7}_1(p)\breve{\gamma}^{c_7}_2(p)\breve{\alpha}^{c_7}_3(p)\breve{\alpha}^{c_7}_4(p,c)$ for $p \in \{p_2, p_3\}$ and $c \in \{c_4, c_5\}$
$\mathcal{I}^{T_m,(3)}_{c_7,HO}$	$\breve{\alpha}^{c_7}_2(p_3)\breve{\alpha}^{c_7}_1(p_3)\breve{\gamma}^{c_7}_1(p_3)\breve{\gamma}^{c_7}_2(p_3)\breve{\alpha}^{c_7}_3(p_3)\breve{\alpha}^{c_7}_4(p_3,c_5)$ $\breve{\alpha}^{c_7}_2(p_2)\breve{\alpha}^{c_7}_1(p_2)\breve{\gamma}^{c_7}_1(p_2)\breve{\gamma}^{c_7}_2(p_2)\breve{\alpha}^{c_7}_3(p_2)\breve{\alpha}^{c_7}_4(p_2,c_4)$

As an interface is designed to provide more abstracted information, such as $\mathcal{I}^{M,(2)}_{c_7,\mathrm{HO}}$ compared with $\mathcal{I}^{M,(1)}_{c_7,\mathrm{HO}}$, the details of operating a machine decrease. This aspect can be investigated by exploiting the cardinality of the state set of an interface as a quantitative measure.

2) Manufacturing Task Goal Reachability: Three different alternative designs of $\mathcal{I}^{T_m}_{i,\mathrm{HO}}$ for an operator to perform manufacturing tasks in c_7 are illustrated as blank circles, dotted and solid shaded rectangles in Fig. 7. For each interface, the required transition functions of $\mathcal{M}^{T_m}_{i,\mathrm{HO}}$ for assuring the manufacturing task reachability are identified, and they are listed in Table III.

In Table III, it is shown that the manufacturing task goal reachability always holds with the interface of $\mathcal{I}^{T_m,(1)}_{c_7,\mathrm{HO}}$ if an operator performs the atomic operations contained in Σ_{T_m}. That is, the operator performs only physical manufacturing tasks and the interface provides all the required steps and information to proceed with the task progress.

Conversely, the operator with either $\mathcal{I}^{T_m,(2)}_{c_7,\mathrm{HO}}$ or $\mathcal{I}^{T_m,(3)}_{c_7,\mathrm{HO}}$ is provided with abstracted information about the system. This type of interface can give the human operator more flexibility in performing the tasks at the expense of the decreased possibility of guaranteeing the manufacturing task goal reachability.

3) Exception Handling Task Supportability: In response to the exceptions reported from the cells, the human supervisor is

TABLE IV
TRANSITIONS REQUIRED FOR EXCEPTION HANDLING TASK SUPPORTABILITY FOR THE THREE INTERFACES

Interface	Necessary transitions for holding Theorem
$\mathcal{I}_{HS}^{T_e,(1)}$	$\sigma \in \Sigma_{T_e}$
$\mathcal{I}_{HS}^{T_e,(2)}$	$\sigma \in \Sigma_{T_e}$
$\mathcal{I}_{HS}^{T_e,(3)}$	$\overleftarrow{\beta}_1(\neg g_1, c_2)$ or $\overleftarrow{\beta}_1(g_1, c_x)\overleftarrow{\beta}_1(g_1, c_y)\overleftarrow{\beta}_1(g_1, c_z)$ where (x, y, z) is any permutation of (4,5,7)

responsible for managing them with an interface of $\mathcal{I}_{\mathrm{HS}}^{T_e}$. Depending on the different interfaces, the transition functions of $\mathcal{M}_{\mathrm{HS}}^{T_e}$ that are required to guarantee that the exception handling tasks are supported are identified, as shown in Table IV.

The interface of $\mathcal{I}_{\mathrm{HS}}^{T_e,(1)}$ is shown to provide the detailed step-by-step procedure specifics for handling exceptions. As in the case of the human operator, it provides all the available actions to deal with managing exceptions.

4) Adaptive Automation: Adaptive automation involves dynamically changing the abstraction levels of information provided through the interfaces to a human operator or a supervisor. Thus, based on the accumulated records of tasks performed by the human depending on the different interfaces, the abstraction levels of the information provided to her/him need to be adjusted.

The operator in c_7 performs a sequence of tasks to process p_3 and p_2 following the steps provided by $\mathcal{I}_{c_7,\mathrm{HO}}^{T_m(1)}$. For instance, the operator starts a machine ($\beta_2^{c_7,T_m}(p)$), processes the part ($\alpha^{c_7,T_m}(p)$), and moves the part to next cell ($\beta_3^{c_7,T_m}(p,c)$).

When the relationship between the abstraction level of information provided to a human operator and the number of violations of manufacturing task reachability is predicted with a certain threshold of a predefined acceptable performance criterion, such as precision/recall, the degree of information through the interface needs to be changed. In the case of this example, if the task sequences proceed for a statistically significant amount of time without violation of the manufacturing task reachability, the interface $\mathcal{I}_{c_7,\mathrm{HO}}^{T_m(1)}$ is changed to $\mathcal{I}_{c_7,\mathrm{HO}}^{T_m,(2)}$ to reduce the amount of detailed information for processing the seat cushion. Depending on the performance of the manufacturing task reachability, the interface can also change to $\mathcal{I}_{c_7,\mathrm{HO}}^{T_m(3)}$ or $\mathcal{I}_{c_7,\mathrm{HO}}^{T_m(1)}$.

Suppose the supervisor with the interface of $\mathcal{I}_{\mathrm{HS}}^{T_e,(2)}$ has satisfied the exception handling supportability for cell c_6 with regard to ($\mathcal{M}_{T_e}, \mathcal{M}_{\mathrm{HS}}^{T_e,(2)}, \mathcal{I}_{\mathrm{HS}}^{T_e,(2)}$). This result means that for the report $\beta_1^{c_6,T_e}$, the supervisor has issued a command to repair the machine ($\gamma_1^{T_e}(c_6)$) and has notified the adjacent cells of the machine breakdown ($\gamma_2^{T_e}(c_6 \to j)$) and the associated subsequent events ($\beta_2^{c_6,T_e}$) and ($\gamma_3^{T_e}(c_6 \to j)$).

If the exception of $\beta_1^{c_6,T_e}$ fails to be reported with $\mathcal{I}_{\mathrm{HS}}^{T_e,(2)}$, the abstraction level of the information should be adjusted and the interface $\mathcal{I}_{\mathrm{HS}}^{T_e,(1)}$, which provides detailed information about the machine breakdown at c_6, can be provided.

VI. CONCLUSION

In the HMI discipline, two separate research areas have long been pursued, i.e., descriptive methodologies based on experiments and normative approaches with formal models. However, data analytics and machine learning techniques have been reported to be successful in a variety of domains.

Therefore, this paper presents a framework that can contribute to bridge the gap between descriptive methodologies and normative approaches in designing and analyzing HMI. In this paper, formal models of HMI are constructed in an effort to describe HMI systems and represent specifications in a human-involved manufacturing system, which is the first contribution that this paper aims to achieve from a normative approach perspective.

Second, our intention in formalizing an HMI manufacturing system is that each of the components can be used to generate software that controls a manufacturing system where humans and machines collaborate. This software can serve as an experimental apparatus for descriptive methodologies.

Finally, the authors envision that the behavioral data of supervisors and operators along with machine and task data can be collected by the implemented system based on the proposed formal models and that the behavioral data can then be analyzed systematically. In this way, the adaptive automation can be investigated by use of machine learning techniques, as suggested in the algorithms.

We note that further experimental studies can be useful for improving the practicality of the proposed mode. For instance, the decision making regarding the manufacturing activities in the presented model is assumed to be determined mainly by the information about the system states. However, the behavior of the human is also significantly affected by the operator's functional state, including situation awareness, mental workload, stress, and many others. Therefore, experimental studies for identifying such factors are highly desirable, as this paper focuses only on the observed events.

Despite these weaknesses, we believe that the presented framework can contribute to bridge the gap between descriptive methodologies and normative approaches in designing and analyzing HMI systems. Furthermore, findings from descriptive approaches by means of extensive experiments can be added to the proposed model. Accordingly, as a direction for future work, the formal constructs in the proposed model can be employed to describe the behaviors of HMI systems and to investigate several specifications that have been studied through extensive experiments involving various HFs issues.

REFERENCES

[1] T. B. Sheridan, *Humans and Automation: System Design and Research Issues*. Hoboken, NJ, USA: Wiley, 2002.
[2] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete event processes," *SIAM J. Control Optim.*, vol. 25, no. 1, pp. 206–230, Jan. 1987.
[3] D. A. Norman, "The problem with automation - inappropriate feedback and interaction, not over-automation," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 327, no. 1241, pp. 585–593, Apr. 12, 1990.
[4] N. B. Sarter and D. D. Woods, "How in the world did we ever get into that mode? Mode error and awareness in supervisory control," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 37, no. 1, pp. 5–19, 1995.

[5] A. Degani and M. Heymann, "Formal verification of human-automation interaction," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 44, no. 1, pp. 28–43, Spring 2002.
[6] G. A. Jamieson and K. J. Vicente, "Designing effective human-automation-plant interfaces: A control-theoretic perspective," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 47, no. 1, pp. 12–34, Spring 2005.
[7] C. Perrow, *Normal Accidents: Living With High Risk Technologies*. Princeton, NJ, USA: Princeton Univ. Press, 2011.
[8] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Using formal verification to evaluate human-automation interaction: A review," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 3, pp. 488–503, May 2013.
[9] M. Adachi, T. Ushio, and Y. Ukawa, "Design of user-interface without automation surprises for discrete event systems," *Control Eng. Pract.*, vol. 14, no. 10, pp. 1249–1258, Oct. 2006.
[10] D. B. Kaber and M. R. Endsley, "The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task," *Theor. Issues Ergonom. Sci.*, vol. 5, no. 2, pp. 113–153, 2004.
[11] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 39, no. 2, pp. 230–253, Jun. 1997.
[12] M. R. Endsley and D. B. Kaber, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics*, vol. 42, no. 3, pp. 462–492, Mar. 1999.
[13] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Hum.*, vol. 30, no. 3, pp. 286–297, May 2000.
[14] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs," *J. Cogn. Eng. Decis. Making*, vol. 2, no. 2, pp. 140–160, 2008.
[15] E. de Visser and R. Parasuraman, "Adaptive aiding of human-robot teaming effects of imperfect automation on performance, trust, and workload," *J. Cogn. Eng. Decis. Making*, vol. 5, no. 2, pp. 209–231, 2011.
[16] P. A. Hancock, R. J. Jagacinski, R. Parasuraman, C. D. Wickens, G. F. Wilson, and D. B. Kaber, "Human-automation interaction research past, present, and future," *Ergonom. Des., Quart. Hum. Factors Appl.*, vol. 21, no. 2, pp. 9–14, 2013.
[17] T. Inagaki, "Adaptive automation: Sharing and trading of control," in *Handbook of Cognitive Task Design*. Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, 2003, vol. 8, pp. 147–169.
[18] E. M. Clarke and J. M. Wing, "Formal methods: State of the art and future directions," *ACM Comput. Surv.*, vol. 28, no. 4, pp. 626–643, Dec. 1996.
[19] D. Shin, R. Wysk, and L. Rothrock, "A formal control-theoretic model of a human–automation interactive manufacturing system control," *Int. J. Prod. Res.*, vol. 44, no. 20, pp. 4273–4295, 2006.
[20] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*. Berlin, Germany: Springer Science & Business Media, 2009.
[21] W. M. Wonham, "Supervisory control of discrete-event systems," in *Encyclopedia of Systems and Control*. Berlin, Germany: Springer, 2015, pp. 1396–1404.
[22] D. Shin, R. A. Wysk, and L. Rothrock, "Formal model of human material-handling tasks for control of manufacturing systems," *IEEE Trans. Syst., Man, Cybern. A, Syst. Hum.*, vol. 36, no. 4, pp. 685–696, Jul. 2006.
[23] J. Bredereke and A. Lankenau, "Safety-relevant mode confusions - modelling and reducing them," *Rel. Eng. Syst. Saf.*, vol. 88, no. 3, pp. 229–245, Jun. 2005.
[24] M. L. Bolton, N. Jiménez, M. M. van Paassen, and M. Trujillo, "Automatically generating specification properties from task models for the formal verification of human–automation interaction," *IEEE Trans. Hum. Mach. Syst.*, vol. 44, no. 5, pp. 561–575, Oct. 2014.
[25] S. Combéfis, D. Giannakopoulou, and C. Pecheur, "Automatic detection of potential automation surprises for ADEPT models," *IEEE Trans. Hum. Mach. Syst.*, vol. 46, no. 2, pp. 267–278, Apr. 2016.
[26] M. Li, J. Wei, X. Zheng, and M. L. Bolton, "A formal machine–learning approach to generating human–machine interfaces from task models," *IEEE Trans. Hum. Mach. Syst.*, vol. 47, no. 6, pp. 822–833, Dec. 2017.
[27] K. Akesson, S. Jain, and P. M. Ferreira, "Hybrid computer-human supervision of discrete event systems," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2002, vol. 3, pp. 2321–2326.
[28] J. F. Petin, D. Gouyon, and G. Morel, "Supervisory synthesis for product-driven automation and its application to a flexible assembly cell," *Control Eng. Pract.*, vol. 15, no. 5, pp. 595–614, May 2007.
[29] T. B. Sheridan, "Human supervisory control," in *Handbook of Human Factors and Ergonomics*, 4th ed. Hoboken, NJ, USA: Wiley, 2012, pp. 990–1015.
[30] N. Leveson, L. D. Pinnel, S. D. Sandys, S. Koga, and J. D. Reese, "Analyzing software specifications for mode confusion potential," in *Proc. Workshop Hum. Error Syst. Develop.*, 1997, pp. 132–146.
[31] T. B. Sheridan and W. L. Verplank, "Human and computer control of undersea teleoperators," Defense Tech. Inf. Center, Fort Belvoir, VA, USA, DTIC Document ADA057655, 1978.
[32] J. P. Wilsten and E. Shayan, "Layout design of a furniture production line using formal methods," *J. Ind. Syst. Eng.*, vol. 1, no. 1, pp. 81–96, 2007.

Taejong Joo received the B.S. and M.S. degrees in industrial engineering with a minor in applied mathematics from Hanyang University, Seoul, South Korea, in 2016 and 2018, respectively.

He is currently a Researcher with the ESTsoft, Inc., Seoul, South Korea. His research interests include deep learning, human–artificial intelligence interactions, and human–machine systems.

Dongmin Shin received the B.S. degree in industrial engineering from Hanyang University, Seoul, South Korea, in 1994, and the M.S. degree in industrial engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 1996.

He is currently a Professor with the Department of Industrial and Management Engineering and the Director of the Smart Manufacturing Learning Center, Hanyang University, Ansan, South Korea. His research interests include human–automation interactions, smart manufacturing systems, and machine learning.

Operator Strategy Model Development in UAV Hacking Detection

Haibei Zhu, Mary L. Cummings, *Senior Member, IEEE*, Mahmoud Elfar, Ziyao Wang, and Miroslav Pajic, *Member, IEEE*

***Abstract*—An increasingly relevant security issue for unmanned aerial vehicles (UAVs, also known as drones) is the possibility of a global positioning system (GPS) spoofing attack. Given the existing problems in current GPS spoofing detection techniques and human visual advantages in searching and localizing targets, we propose a human-autonomy collaborative approach of human geo-location to assist UAV control systems in detecting GPS spoofing attacks. An interactive testbed and experiment were designed and used to evaluate this approach, which demonstrated that human-autonomy collaborative hacking detection is a viable concept. Using the hidden Markov model (HMM) approach, operator behavior patterns and strategies from the experiment were modeled via hidden states and transitions among them. These models revealed two dominant hacking detection strategies. Statistical results and expert performer evaluations show no significant difference between different hacking detection strategies in terms of correct detection. The detection strategy model suggests areas of future research in decision support tool design for UAV hacking detection. Also, the development of HMMs presents the feasibility of quantitatively investigating operator behavior patterns and strategies in human supervisory control scenarios.**

***Index Terms*—Cyber-attack detection, hidden Markov model (HMM), human geo-location, human supervisory control, strategy classification, unmanned aerial vehicle (UAV).**

I. Introduction

UNMANNED aerial vehicles (UAVs) have significantly increasing use in commercial and military applications. The continued growth in numbers and functionalities of UAVs has been accompanied by many security, privacy, and regulatory concerns. One common security concern is UAV global positioning system (GPS) spoofing, in which attackers deceive GPS receivers by providing counterfeit GPS signals in order to override UAV navigation systems and redirect UAVs to unexpected destinations [1], [2]. One such well known incident garnered public attention in 2011 when an RQ-170 Sentinel UAV was captured using GPS spoofing attacks [3]. Therefore, successfully detecting GPS spoofing attacks is important for UAV control systems.

Understanding that human vision has advantages in complex searching and localizing tasks [4]–[6], we demonstrated a human-autonomy collaborative approach through geo-location in which humans can assist autonomous systems in the detection of possible GPS spoofing attacks on UAVs. In this study, this approach was evaluated via an experiment, which was designed and conducted using the security-aware research environment for supervisory control of heterogeneous unmanned vehicles (RESCHU-SA) platform [7], extension of the platform from [8]. Experimental sessions simulated human supervisory multiUAV control scenarios with potential UAV GPS spoofing attacks. Operators were able to successfully detect hacking events such that 65% of total experimental sessions exhibited at least 80% correct hacking identification. We also discovered that operators with significant video game experience were the best performers in hacking detection [9].

While this initial study demonstrated that human operators could successfully identify UAV GPS spoofing attacks through geo-location, given that such research has never before been conducted, our goal is to better understand what strategies emerged as novices attempted to determine if they had been hacked. To this end, it was advantageous to develop human behavior models to investigate operator behavior patterns, both in the execution of their primary task of supervising UAVs, and in attempting to thwart hacking attempts. Such models could be particularly useful as they could highlight training problems or interface design anomalies. Finally, such models could be used to develop predictive decision support tools that could assist human operators, particularly under areas of high workload and stress. The rest of this paper presents our efforts to develop strategy models of humans supervising multiple UAVs and determining whether a UAV had been hacked through human geo-location.

Manuscript received April 1, 2018; revised August 26, 2018; accepted November 17, 2018. Date of publication February 26, 2019; date of current version November 21, 2019. This work was supported in part by the NSF under Grant CNS-1652544 and in part by the ONR under Agreement N00014-17-1-2012 and Agreement N00014-17-1-2504. This paper was recommended by Associate Editor M. L. Bolton. *(Corresponding author: Haibei Zhu.)*

H. Zhu, M. Elfar, and M. Pajic are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: haibei.zhu@duke.edu; mahmoud.elfar@duke.edu; miroslav.pajic@duke.edu).

M. L. Cummings, and Z. Wang are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708 USA (e-mail: mary.cummings@duke.edu; ziyao.wang@duke.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2018.2888578

II. Background

A. UAV GPS Spoofing Detection

Remotely controlled UAVs typically rely on an embedded navigation system known as the GPS, which provides accurate localization information including position, velocity, and time for UAV GPS receivers. GPS receivers can calculate the precise latitude, longitude, height, and speed based on received satellite signals. However, GPS receivers are vulnerable to GPS spoofing attacks, in which GPS receivers are attacked by counterfeit

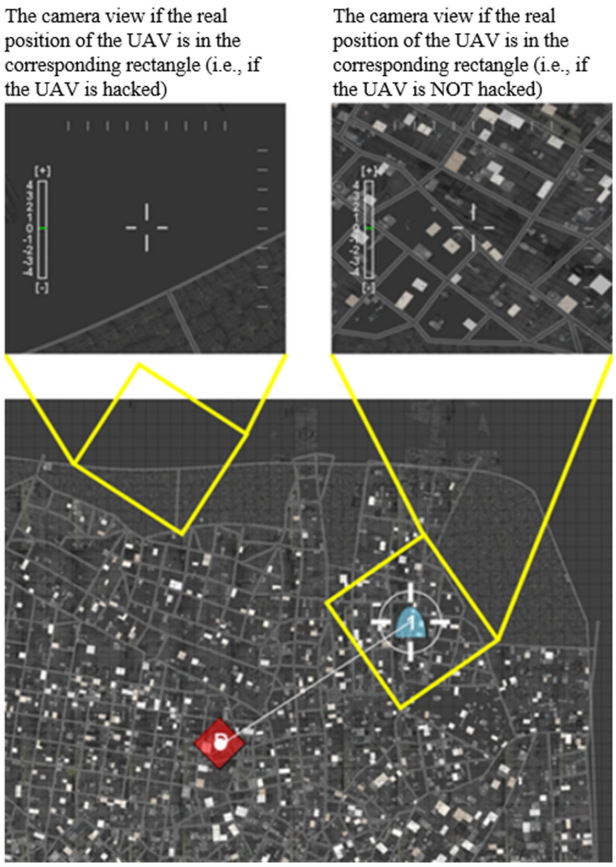


Fig. 1. Example of GPS reported locations on the map.

signals generated from GPS spoofers [10]–[14]. Many autonomous GPS spoofing detection methods have been proposed in recent studies [11]–[17]. However, false alarms and detection failures still exist while applying autonomous GPS spoofing detection [10], [11], [15]. Therefore, more research is needed to improve autonomous detection systems.

UAVs are commonly equipped with both a GPS navigation system and payload camera, whose signal is independent of the UAV GPS signal. Thus, if these two signals are independent, the payload camera view can be used as a reference to assist autonomous detection systems in detecting UAV GPS spoofing attacks. Based on the precondition that UAV payload camera views can provide the unbiased surrounding scene of UAVs, we propose that human operators can act as supplementary sensors and assist autonomous systems to detect UAV hacking attacks through the comparative human geo-location method.

In human geo-location, an operator can compare the nontampered video feed from the UAV payload camera to the potentially falsified GPS reported location on the map. This approach allows operators to detect inconsistencies, which indicate potential hacking attacks, between the location interpreted from the camera view and the GPS location reported on the map. In theory, such cross referencing could be accomplished automatically through autonomous localization and sensor-fusion techniques (e.g., [18], [19]), but these have not been very successful [20], particularly in military applications [21].

Based on feature integration theory, the first stage of human vision obtaining information from targets is the preattentive

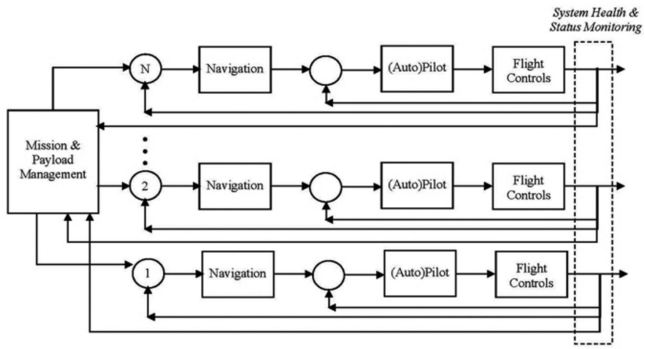


Fig. 2. Human supervisory multiple UAV control architecture [22].

stage, in which a human observer can gather basic information about a target even before the observer becomes conscious of it [4]. Thus, human vision can process target information efficiently in complex environments. Human observers also tend to choose areas that maximize information of the target in a salience-driven visual search strategy [5]. In addition, the direction discrimination threshold of human vision has a low average of 1.8 degrees [6], which suggests that human observers can precisely detect small changes in target movement orientation. Considering these human visual advantages, human operators can potentially assist UAV localization and detect potential UAV GPS spoofing attacks.

An example of human geo-location in UAV GPS spoofing detection is shown in Fig. 1. The GPS-reported location of the UAV is shown as the blue dome on the map in the upper right. If the UAV is under attack, the operator will observe the scene below the UAV through the camera, which would be different from the surrounding environment of the GPS-reported location on the map; e.g., as in the upper left camera-view in Fig. 1. If the UAV is not under attack, the operator will observe the scene below the UAV, as in the upper-right camera view in Fig. 1, matching the reported location. When a GPS spoofing attack is confirmed, the operator can prevent losing the hacked UAV by overriding the physical controls.

B. Modeling Operator Behavior

The human geo-location approach to hacking detection is an example of a common UAV control scheme which incorporates human supervisory control, in which a human operator monitors a multiUAV system, intermittently navigating UAVs, and conducting other higher level tasks [23]. The hierarchical architecture of a human supervisory UAV control loop of single operator with multiple UAVs is shown in Fig. 2 [22]. In this architecture, multiple parallel outermost loops represent the highest-level control of managing missions and payloads by human operators. The inner loops represent lower level navigation and flight controls by autonomous systems or operators. This architecture can be introduced with various levels of automation. The successful control of higher level operator loops depends on the success of lower level autonomous system loops. In this study, we assume that human operators keep higher level decision-making processes, and autonomous systems are in charge of lower level UAV control and navigation operations [22].

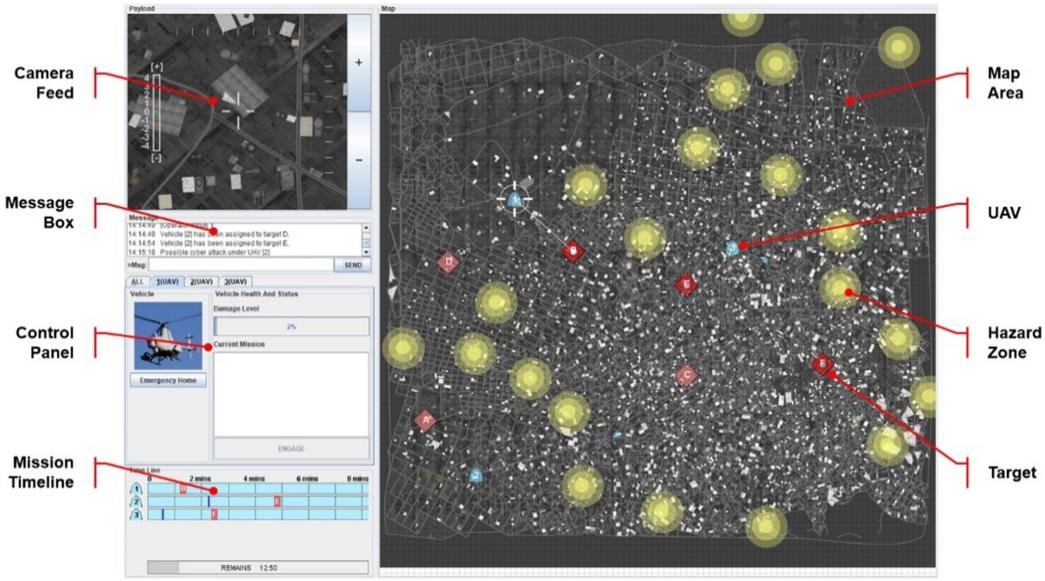


Fig. 3. RESCHU-SA experiment platform interface [7].

In supervisory control settings where humans are supervising one or more autonomous systems, human operator behavior models are needed for multiple reasons:

1) To investigate general operator behavior patterns, in order to determine if observed behaviors match the expected behaviors;
2) To investigate operators' strategies, in order to identify points of inefficiency or error;
3) Study both endogenous and exogenous factors that impact operator behavior patterns such as video game experience and task load;
4) Study how automation can improve operators' performance and success rate in task performance, including the use of predictive operator behavior models.

In terms of the hacking detection supervisory control setting we consider, we need a way to determine strategies that operators develop in their attempts to detect and mitigate hacking attempts, and how to improve upon those strategies that could include the use of automated decision support.

One problem with the generation of such models is that while interactions between a human operator and a supervisory control system can be directly observed through human physical interaction with an interface, such observations cannot be directly associated with a human thought, goal, plan, or strategy. In order to develop operator models that link actions and behaviors to plans, goals, and strategies, we need a method that abstracts low-level physical interface interactions into higher operator behavioral states and strategies. We believe that a hidden Markov modeling approach provides the foundation to do this, as described in the next section.

C. Markov Modeling Approaches

Markov models are widely used to capture stochastic evolution of state transitions in the state-space [24]. Many studies have used Markov models to investigate low-level human actions [25], [26]. However, Markov models only capture observable interactions between human operators and control systems, which may not accurately reflect operators' high-level behavioral states. Therefore, hidden Markov models (HMM), which are an extension of Markov models, could be a useful alternative in this regard.

An HMM is a two-layer stochastic model that describes a Markov process with a higher layer of indirectly observable system states and a lower layer of observable emissions from each state. The HMM formalism is widely used in machine learning, especially in speech recognition [27] and development of human operator behavior models in driving [28]. HMMs using an unsupervised approach to model training have been shown to provide more accurate operator behavior models over supervised learning approaches [29], [30]. Because an HMM can present higher level operator behavioral states using hidden system states based on lower level operator interactions with a supervisory control system like a UAV ground control station, the HMM was selected as the modeling framework for this effort.

III. Data Generation

In order to develop models of operator behavior in the UAV supervisory control environment with potential hacking events, user interactions with such a system were needed to provide the underlying training data. To this end, we developed the RESCHU-SA (now freely available to interested parties) [7], [8], [31]. RESCHU-SA is a Java-based simulation platform for a single operator with multiUAV supervisory control scenarios. It provides the flexibility to design multitasking scenarios including both navigational and imagery analysis tasks. Moreover, this platform provides the capability of simulating UAV GPS spoofing attacks, in which hacked UAVs deviate from the originally assigned paths and target unexpected destinations, along with real or false notifications that simulate autonomous GPS spoofing detection systems.

The interface of the RESCHU-SA platform is shown in Fig. 3. Five main components are featured in this interface, including

the payload camera view, message box, control panel, timeline, and map area. Specifically, the camera view displays the video stream of the surrounding environment beneath the selected UAV. The primary purpose of this view is to conduct imagery analysis tasks and can be used to determine the actual location of UAVs for detecting potential hacking events. The map displays the surveillance area with real-time locations of all UAVs, hazard areas, and targets.

A. Experiment Design

To collect enough data to develop operator models, a set of experiments was conducted using RESCHU-SA. The primary objectives of operators using RESCHU-SA are to control multiple UAVs to: 1) determine whether UAVs are under GPS spoofing attacks; 2) perform reconnaissance imagery tasks of counting road intersections when UAVs reach assigned targets; and 3) ensure that UAVs do not encounter hazard areas.

Given that a previous study demonstrated that the task load can significantly impact an operator's performance, and thus strategies [8], task load was the only controlled experimental variable in this experiment. Two objective task load levels, high versus low, were introduced, and each participant had both task load scenarios in the experiment. In the low-task load scenario, operators navigate three UAVs with six targets and six hacking notifications, including three real hacking notifications and three false alarms. In the high-task load scenario, operators navigate six UAVs with nine targets and nine hacking notifications, including five real hacking notifications and four false alarms. To simplify the hacking detection, no notification miss was introduced in the experiment that all real hacking events come with notifications.

In RESCHU-SA, operators are responsible for safely navigating UAVs to targets. Hazard areas can appear and disappear randomly, which require replanning the vehicle around these threat areas. In the experiment, GPS spoofing attack events with notifications followed a predefined schedule but appeared to randomly occur while an operator navigated the UAVs. Once an operator received a notification that a certain UAV was under possible cyber-attack, the operator could then investigate the potential UAV hacking by checking the UAV camera view and matching it against the position of the UAV on the map. Although UAV position drifts may be caused by GPS degradation, we assumed that all position drifts were caused by GPS spoofing attacks to simplify the hacking detection scenarios in this experiment.

When UAVs that were not hacked reached a target, the operator engaged in an imagery task of counting the road intersections from the UAV's camera view at a prespecified zoom level. This task represents the primary purpose of the mission, which is information gathering. The imagery counting task was the participants' primary work load task, and it allowed us to assess their performance based on the number of attempted tasks and the task correctness percentage.

B. Experiment Subjects and Procedure

Thirty-six participants took part in this experiment, including 22 males and 14 females. Age ranged from 19 to 34 years with an average of 25.2 and a standard deviation of 3.8 years. Among all participants, 18 participants had little video game experience, six participants had monthly gaming experience, five participants played video game several times a week, another five participants had weekly gaming experience, and only two participants had daily gaming experience. The experimental procedure consisted of four main sections including a self-paced tutorial section, a practice section, a test section, and a debriefing section. Specifically, in the test section, each participant finished two test sessions, including a counterbalanced high- and a low-task load scenario. Thus, we had 72 test sessions and collected data from all these sessions.

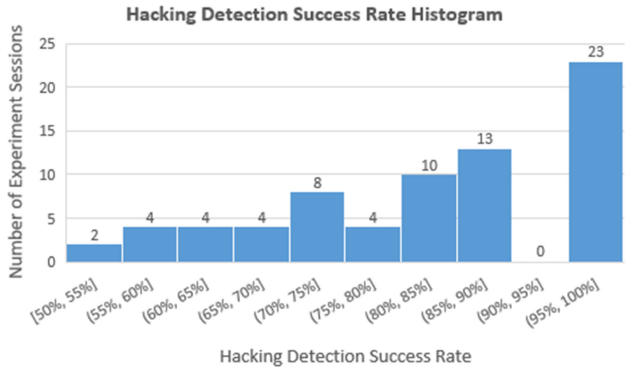


Fig. 4. Histogram of the hacking detection success rate.

TABLE I
CONFUSION MATRIX OF HACKING DETECTION DECISIONS IN DIFFERENT NOTIFICATIONS

	Real hacking notification	False alarm notification
Decision of considering UAV was hacked	224	40
Decision of considering UAV was not hacked	63	207

C. Experiment Results

In this experiment, 23 out of the total 72 test sessions (32%) resulted in 100% successful hack identifications, while another 24 (33%) reached above 80% successful attack identification. Thus, as shown in Fig. 4, 65% of total test sessions reached 80% correct hacking detection or better without having any prior formal hacking detection training.

Specifically focusing on the difference between real hacking notification and false alarms, as shown in Table I, out of all the 287 (224 + 63) real hacking notifications across all participants, the overall success rate was 78% (224 ÷ 287), and for all the 247 (40 + 207) false alarms, the success rate was 84% (207 ÷ 247). In other words, the type one error (false positive, operators considered UAV not hacked with real hacking notification) was 22% (63 ÷ 287), which was slightly higher than the type two error (false negative, operators considered UAV hacked with false alarm notification) of 16% (40 ÷ 247). Thus, operators were slightly better at detecting false alarms than identifying real hacking notifications.

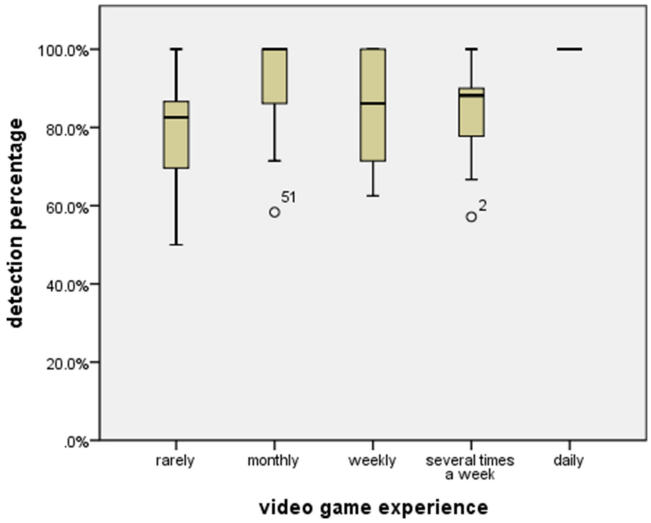


Fig. 5. Boxplot of hacking detection success rate based on different video game experience.

Task load, as a major experimental factor, only affected UAV damage level (MANOVA $F(1,31) = 32.93$, $p < 0.001$, alpha $= 0.05$), but it did not affect any other performance metric. However, the video game experience covariate had a significant effect on participants' correct hacking detections ($F(1,31) = 4.652, p = 0.039$), as shown in the boxplot in Fig. 5. This means that the more the video game experience, the higher the chance of a correct hacking detection. Not surprisingly, seven participants who lost UAVs had no video game experience, and the other five who lost UAVs ranged from little to moderate gaming experience. Participants with daily gaming experience did not lose any UAVs and were 100% correct in hacking identification.

These statistical results of our experiment provide a high-level understanding of the factors that impacted operator performance. However, we need to further investigate the underlying nature of why such factors had certain effects on performance. In addition, operators' hacking detection strategies cannot be inferred via statistical results. Therefore, human operator models are needed for further investigating operator behavior patterns and detection strategies in such UAV supervisory control scenarios.

IV. HMM Structure, Training, and Selection

As discussed previously, human operator behavior models can illustrate operator behavior patterns and strategies in high-level tasks. Considering that HMMs can infer hidden higher level operator behavioral states from observable lower level interactions between the operators and autonomous systems, HMMs were chosen for modeling the observable behaviors from the RESCHU-SA experiment.

A. HMM Structure

Based on the classic notation of HMM, the HMM can be formally defined as a tuple [32]

TABLE II
Observations (Emissions) of HMMs From RESCHU-SA Experiment Interface

Index	1	2	3	4
Observation	Add waypoint	Move waypoint	Delete waypoint	Move endpoint
Index	5	6	7	8
Observation	Switch target	Engage task	Select UAV	Confirm notification
Index	9	10	11	12
Observation	Ignore notification	Consider UAV hacked	Consider UAV not hacked	Adjust zoom level

$$H = \{S, V, A, B\}.$$

Here, $S = \{S_1, S_2, ..., S_N\}$ represents N different hidden states, $V = \{V_1, V_2, ..., V_M\}$ represents M different observations. Also, $A = \{a_{ij}\}$ is an $N \times N$ transition probability matrix, where $a_{ij} = P\{S_j^{t+1} | S_i^t\}$, $i, j = 1, 2, ..., N$, and $B = \{b_{ik}\}$ is an $N \times M$ emission probability matrix, where $b_{ik} = P\{V_k | S_i\}$, $i = 1, 2, ..., N$, $k = 1, 2, ..., M$. In addition, both $a_{ij}, b_{ik} \geq 0$. In HMMs, each hidden state can be considered as a cluster of observations with different weights, which are emission probabilities. The system states (or operator behavioral states, in this paper) transfer among hidden states based on the time sequence, and the probabilities of switching from the current state to the next state are the transition probabilities.

B. HMM Training and Selection

The first step in the HMM training process is state space reduction. In RESCHU-SA, every key stroke and mouse action were recorded in log files, along with the system status. In an HMM, the hidden higher level behavioral states are clusters of operator actions, so the interaction data should be aggregations of observations based on a predefined state reduction grammar. In this manner, there were 12 possible places for operators to click in RESCHU-SA, which yielded 12 observations, as presented in Table II.

The multisequence Baum–Welch algorithm, an unsupervised model training method, was used in model training [33]. HMM training results were then selected (number of hidden states) using the Bayesian information criterion (BIC) [27], [34] and the number of rare states (NRS) method [35] to achieve both high model likelihood values and reasonable model structures. Models with the lowest BIC values are preferred. The BIC balances the increase of model complexity, which is caused by the increase of the model features, by penalizing the number of free parameters in the model training process. The NRS method maintains the simplicity and interpretability of a descriptive model by monitoring all rare states whose occurrence frequencies are lower than a certain threshold value, which is usually 5%. Generally, HMMs without any rare state are preferred. When BIC curves are monotonically decreasing, the NRS method can suggest the model with the highest number of hidden states without any rare state.

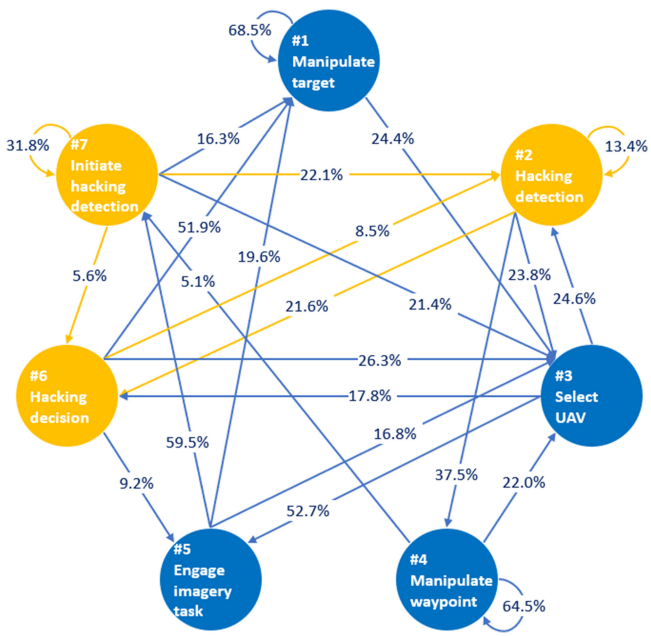


Fig. 6. General human operator behavior HMM.

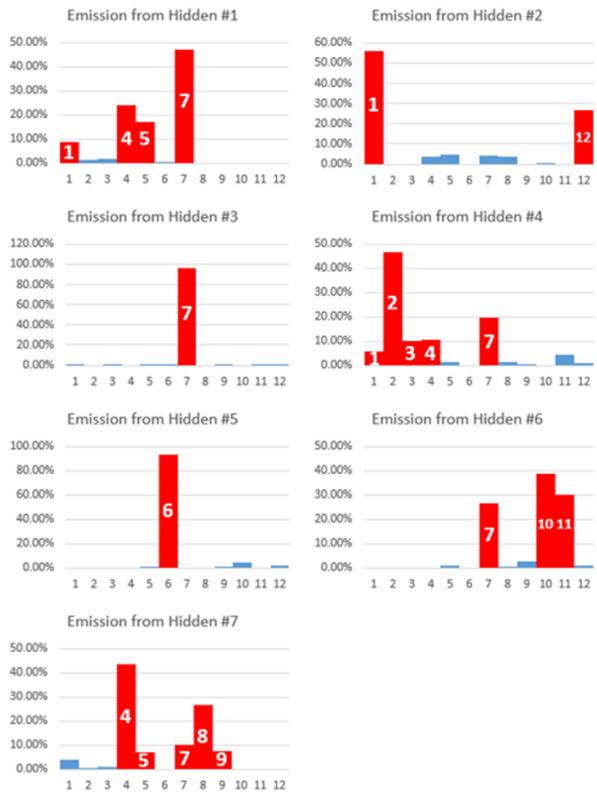


Fig. 7. Emission probabilities for the HMM capturing general operator behavior.

V. General Operator Behavior Model

Understanding that task load did not affect operators' overall performance and success rate in hacking detections and imagery tasks, the general operator behavior model was trained using data from both high- and low-task load scenarios. As shown in Table II, the general operator behavior HMM was trained using observation sequences with 12 different observations. Based on the model selection process described previously, the HMM with seven states had the lowest BIC value. Also considering that the 7-state model did not have any rare states and the HMMs with eight or more states had at least one rare state, the general operator behavior model was determined to be a 7-state HMM, as shown in Fig. 6. The interpretation for each hidden state was determined by the emission probabilities, shown in Fig. 7.

The first state was interpreted as "Manipulate target" because it was mainly a cluster of observation 4 (Move endpoint), 5 (Switch target), and 7 (Select UAV), which were directly related to UAV target manipulations. The second state was interpreted as "Hacking detection" because this was the only state that had significant emission to observation 12 (Adjust zoom level), which indicated the typical operation of using a UAV's camera to compare against the map. The third state was interpreted as "Select UAV" because its only major emission was observation 7 (Select UAV). The fourth state was interpreted as "Manipulate waypoint" because it was a cluster of observation 1 (Add waypoint), 2 (Move waypoint), 3 (Delete waypoint), and 7 (Select UAV), which were directly related to waypoint management. The fifth state was interpreted as "Engage imagery task" because its only major emission was observation 6 (Engage task), indicating that people were executing the intersection counting task. The sixth state was interpreted as "Hacking decision" because it was the only state that had major emissions to observation 10 (Consider UAV hacked) and 11 (Consider UAV not hacked) which were decisions to hacking events. The seventh state was interpreted as "Initiate hacking detection" because it was the only state that had emissions to observation 8 (Confirm notification) and 9 (Ignore notification) which indicated the initiation of hacking detection.

The general operator behavior model represents the operator behavioral states in navigating UAVs, conducting imagery search, and dealing with potential hacking events. The first interesting fact shown in the model is that the UAV navigation (highlighted in blue) and hacking detection (highlighted in orange) functional groups can be distinguished clearly. The transitions between these two functional groups represent the probabilities of switching functional groups in operator behavioral states. This distinction shows that operators typically conducted tasks either in UAV navigation or hacking detection, reflecting that operators were switching between two primary objectives of navigating the UAVs and detecting hacking.

Interestingly, a previous study on the original RESCHU platform, which only dealt with the navigation of UAVs and did not have any hacking considerations [30], exhibited just four similar states to those blue states in Fig. 6. This is an important finding since it means that the addition of a new set of tasks did not dramatically change the underlying states, rather the added functionality of hacking detection simply added more states. This suggests that at least in some supervisory control environments, functions may be modeled in a modular fashion, which would reduce the workload in adapting older models as new functions are added.

In addition, the general RESCHU-SA model in Fig. 6 shows some potential inefficiencies in operator behavior patterns. In

TABLE III
OBSERVATIONS (EMISSIONS) OF THE HACKING DETECTION STRATEGY HMM FROM RESCHU-SA EXPERIMENT INTERFACE

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

the navigation functional set of states, the first state of "Manipulate target" and the fourth state of "Manipulate waypoint" have high self-transition probabilities. These high self-transition probabilities indicate that once operators entered these two behavioral states, operators tended to conduct repeated operations. For instance, the participant who repeated manipulating targets the most (91 times compared to the average of 35 times), was enmeshed in the "Manipulate target" state and actually had a low overall performance score of 236 (compared to the average of 303). These repeated operations indicate potential inefficiencies that could be improved with future designs for the UAV supervisory control interface.

Two hidden states, "Hacking detection" and "Initiate hacking detection," in the hacking detection functional group also revealed potential problems with self-transitions. Based on statistical analyses, the time consumption in hacking detection was negatively correlated with the hacking detection success rate (Pearson $= -0.375, p = 0.001$). Thus, this fact implies that the longer the person spent investigating a potential hacking event, the less likely a successful detection would occur. This result was curious because as people gather more information, they should increase their probability of successful detection. These results then led us to develop more detailed HMMs about just operator hacking detection strategies in order to shed more light about this unexpected result. These more specific HMMs are detailed in the following section.

VI. HACKING DETECTION STRATEGY MODEL

The HMM in Fig. 6 provides an overall view into how operators approached the overall tasks of navigating the UAVs in support of their primary reconnaissance missions, while also dealing with hacking events. However, since this model does not provide enough detail about just how exactly people formed strategies for dealing with the hacking events, we elected to focus on those operator interactions from the beginning to the end of each hacking event. Overall, there were 15 such hacking events per participant. The resulting hacking detection model was trained based on ten observations instead of the original 12 observations, as shown in Table III. In the revised model training, original observations of "Confirm notification" and "Ignore notification" were combined to "Perceive hacking," and "Consider UAV hacked" and "Consider UAV not hacked" were combined to "Detection decision."

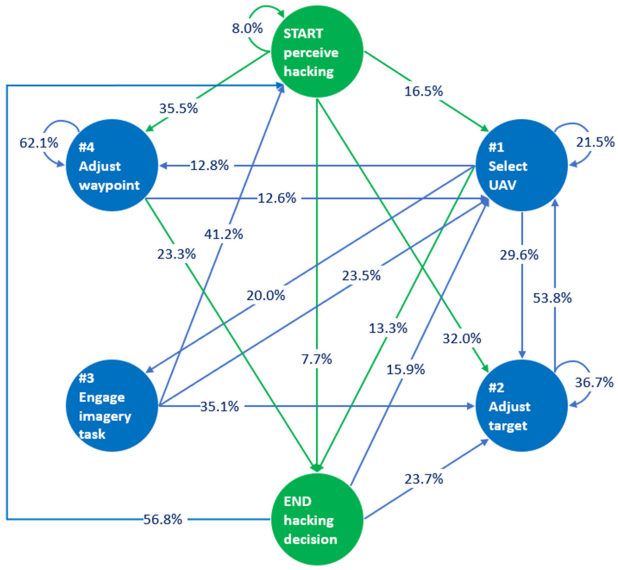


Fig. 8. Operator hacking detection strategy model.

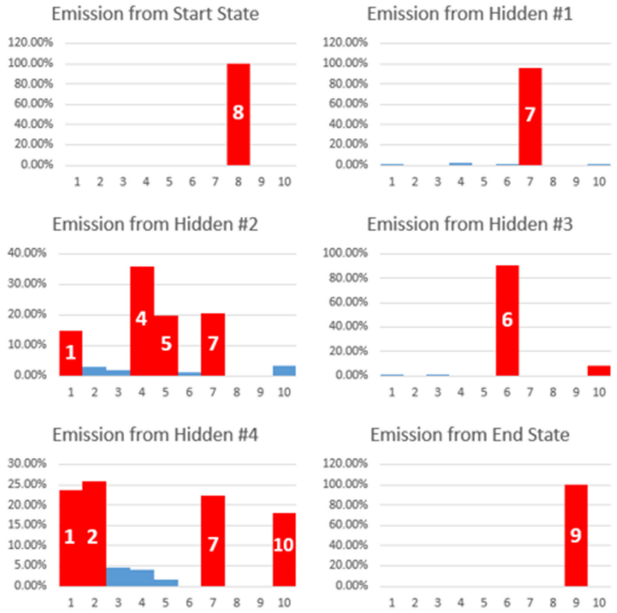


Fig. 9. Emission probabilities of the hacking detection strategy model.

As shown in Fig. 8, the obtained hacking detection strategy model is a 6-state HMM based on the similar model selection process as used for the general operator behavior model. The interpretation for each hidden state was determined by the emission probabilities shown in Fig. 9. Although the observations were slightly different, the interpretation criteria were similar to the general behavior model. The six hidden states were interpreted as: 1) the start state of "Perceive Hacking"; 2) "Select UAV"; 3) "Adjust target"; 4) "Engage imagery task"; 5) "Adjust waypoint"; and 6) the end state of "Hacking decision." The 56.8% transition from the END state to the START state represents overlapping hacking detections. This means once operators finished a hacking detection, roughly half the operators

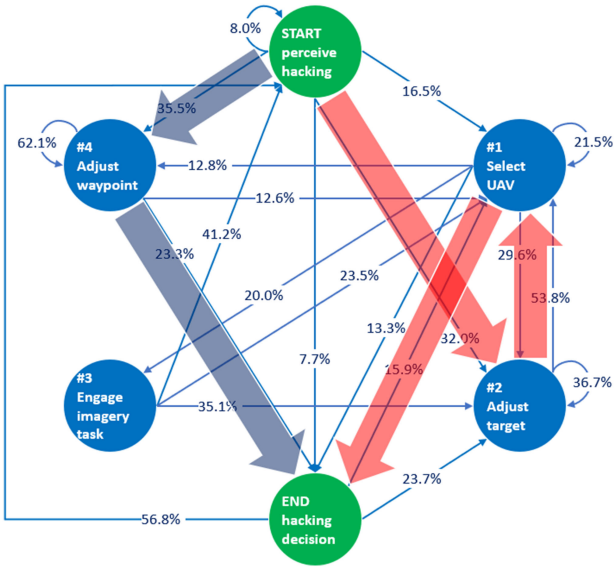


Fig. 10. Master participant strategies in hacking detection strategy model.

then went on to solve another hacking event that occurred almost coincidentally with the current event.

A. Hacking Detection Strategies

Two major behavioral state transitions (also known as operation flows) in the hacking detection HMM can be observed based on transition probabilities, as shown in Fig. 10. Such transitions are considered as detection strategies because they start from the START state, in which operators perceived hacking events, to the END state, in which operators determined detection results. The first major flow, indicated by blue arrows, has one single intermediate state of "Adjust waypoint" between the start and the end state. The second major flow, indicated by red arrows, has two intermediate states of "Adjust target" and "Select UAV" between the start and the end. These two major operation flows suggest two dominant hacking detection strategies, termed "waypoint-oriented strategy" and "target-oriented strategy."

In the waypoint-oriented strategy, operators tended to manipulate UAV waypoints, including adding and moving waypoints, to detect hacking events. In this hacking detection strategy, to investigate the potential differences in the scene between the camera view and the surrounding map area, operators typically either manipulated or introduced waypoints. Operators who used this strategy typically fixated on comparing the effects of turning the UAV and the appearance of the ground in the camera feed to that expected while turning on the map. This can be considered a dynamic strategy as motion was a key element in the determination of location.

In the target-oriented strategy, operators tended to directly switch UAV targets to detect hacking events. In this strategy, operators were typically focused more on the specific landmarks that the UAVs would fly over, such as unusual intersections or buildings. This can be considered a static strategy as operators would wait until the UAV reached a place of interest to make a hacked or not hacked decision. Both strategies revealed inefficiencies, primarily through the self-transition probabilities. For example, in the waypoint-oriented strategy, 62% of people stayed in this state, repeatedly adding, moving, and deleting waypoints. Similarly, 37% of people repeatedly redirected vehicles to other targets, suggesting an inefficient target selection process. These actions suggest inefficiencies that potentially could be made better with advanced decision support, which is an area of future work.

TABLE IV
PARTICIPANT CLASSIFICATION BASED ON DIFFERENT HACKING DETECTION STRATEGIES

Index	Strategy	Number	Percentage
1	Waypoint strong dominant	10	27.8%
2	Waypoint weak dominant	7	19.4%
3	Target weak dominant	11	30.6%
4	Target strong dominant	8	22.2%

The occurrence frequency and percentages of the waypoint- and target-oriented strategies for each participant was obtained by applying the hacking detection HMM to each participant's data using the Viterbi algorithm [27]. Based on the occurrence percentage of the adjust waypoint and adjust target states, participants were classified into different hacking detection categories. As shown in Table IV, participants were classified into four categories: 1) waypoint strong dominant strategy; 2) waypoint weak dominant strategy; 3) target weak dominant strategy; and 4) target strong dominant strategy. The population of each strategy category was approximately one-fourth the total participant population.

Another repeated-measure multivariate ANOVA model with a significance level of 0.05 was used to analyze the impact of different hacking detection strategies on participant performance and hacking detection success rate. In this rm-MANOVA model, strategy categories were considered as a between-subject factor. The rm-MANOVA model showed that different hacking detection strategies did not affect the overall participant performance ($F(3, 27) = 0.754, p = 0.530$), their hacking detection success rate ($F(3, 27) = 0.086, p = 0.967$), and their imagery counting task success rate ($F(3, 27) = 1.528, p = 0.230$). Thus, when examining the aggregate group, no strategy dominated in terms of performance. However, given that the only operator who had perfect performance were the two operators with daily game experience, we examined their strategies in detail in the next section to shed more light on which strategies could potentially produce the best outcomes.

B. Master Participant Hacking Strategies

Developing separate HMMs for the two master participants was not possible due to the limited data; however, operator state paths can provide a map of individual strategies. As shown in Fig. 10, the two dominant hacking detection strategies are

highlighted separately to represent the two master participant strategies. The red path represents the first master participant's operation flow and the blue path represents the second master participant's operation flow.

As depicted in Fig. 10, the master participants represented the two dominant strategies shown in the hacking detection model of Fig. 8. The first master participant exhibited the target-oriented strategy, spending an average of 81.1 s in each hacking detection (overall average for target-dominant people was 100.3 s). The second master participant exhibited the waypoint-adjusted strategy, spending an average of 50.5 s in each hacking detection (overall average for waypoint-dominant people was 81.8 s). The two master participants demonstrated the two dominant strategies shown in the model for all participants and both master participants achieved 100% detection, so there is no clear dominant strategy in terms of quality of final decision. However, there was a clear difference in speed with the waypoint-oriented strategy taking, on average, 30 s less to accomplish, which can be seen in the two strategies in Fig. 10 where the target-oriented strategy has an additional state. This is a practically significant number as intervening as quickly as possible in the middle of a hacking event is paramount. So, while this analysis reveals no dominant strategy in terms of detecting a hacking event, it does suggest that the waypoint-oriented strategy is likely to lead to faster results, which could be very important in prosecuting actual events.

VII. Conclusion

The human operator behavior models in this study present the feasibility of investigating operator behavior patterns and strategies in conducting supervisory control tasks through the use of HMMs. From operator behavior models, we can investigate factors that potentially impact operator behavior patterns and their higher level strategies. Observed strategies from a single HMM can provide engineers and researchers a practical approach to investigating human operators' strategies in human supervisory control scenarios.

The general behavior model, derived using RESCHU-SA-based experiments, shows seven major human operator behavioral states for supervision of UAVs that could be subject to hacking events. In this model, two functional groups emerged, including a hacking detection group with three behavioral states and a UAV navigation group with four states. Operators generally switched between functional groups as demands dictated, i.e., when a hacking event emerged, operators moved from the navigation flow to the hacking flow, indicating that such functions could be seen as modular.

A 6-state hacking detection strategy model allowed us to investigate operator hacking detection strategies in detail. Two major strategies can be observed from the model, including waypoint-oriented and target-oriented strategies. Based on statistical results, different hacking detection strategies did not affect operators' overall performance and success rate in hacking detection. Although no single best hacking detection strategy emerged in terms of quality, one strategy was superior in terms of the time to correct decision.

Although this geo-location approach for UAV hacking detection is still in an experimental stage, these initial results suggest that such an approach could enhance the security of future supervisory UAV control systems if hacking notifications are provided. Considering that no hacking notification misses were introduced in this experiment, as a future study we will investigate the potential effects on operators' performance and detection strategies if the autonomous system fails to provide notifications. In addition, certain limitations still exist in our HMM method, including limited model training data and required experimenter subjective judgment in hidden state interpretation, which is a fundamental issue for all unsupervised machine learning approaches. Current research is underway to determine how to make such model interpretation more straightforward as well as improve sensitivity analysis methods to reveal weaknesses in employed assumptions.

These descriptive operator behavior models highlight the fact that even effective strategies can be inefficient. Further work is needed to determine why people adopt different strategies and whether additional assistance can be used to improve operator strategies, either through training or a decision support system. Finally, the development and utilization of predictive behavior models can contribute to the future development of real-time guidance systems, which monitor operators constantly and provide real-time operational guidance.

References

[1] T. E. Humphreys, B. M. Ledvina, M. L. Psiaki, B. W. O'Hanlon, and P. M. Kintner, Jr., "Assessing the spoofing threat: Development of a portable GPS civilian spoofer," in *Proc. 21st Int. Tech. Meeting Satell. Division Inst. Navigat.*, 2008, pp. 2314–2325.

[2] D. P. Shepard, T. E. Humphreys, and A. A. Fansler, "Evaluation of the vulnerability of phasor measurement units to GPS spoofing attacks," *Int. J. Crit. Infrastructure Protection*, vol. 5, no. 3/4, pp. 146–153, 2012.

[3] S. Shane and D. E. Sanger, "Drone crash in Iran reveals secret US surveillance effort," *The New York Times*, vol. 7, Dec. 7, 2011.

[4] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[6] B. de Bruyn and G. A. Orban, "Human velocity and direction discrimination measured with random dot patterns," *Vis. Res.*, vol. 28, no. 12, pp. 1323–1335, 1988.

[7] M. Elfar *et al.*, "Platform for security-aware design of human-on-the-loop cyber-physical systems," in *Proc. 8th Int. Conf. Cyber-Phys. Syst.*, 2017, pp. 93–94.

[8] B. Donmez, C. Nehme, and M. L. Cummings, "Modeling workload impact in multiple unmanned vehicle supervisory control," *IEEE Trans. Syst., Man, Cybern. Part A, Syst., Humans*, vol. 40, no. 6, pp. 1180–1190, Nov. 2010.

[9] H. Zhu, M. Elfar, M. Pajic, Z. Wang, and M. Cummings, "Human augmentation of UAV cyber-attack detection," in *Augmented Cognition: Users and Contexts* (Lecture Notes in Computer Science, vol 10916), D. Schmorrow, C. Fidopiastis Eds. Cham, Switzerland: Springer, 2018.

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

[11] A. Broumandan, A. Jafarnia-Jahromi, V. Dehghanian, J. Nielsen, and G. Lachapelle, "GNSS spoofing detection in handheld receivers based on signal spatial correlation," in *Proc. IEEE/ION Position Location Navigat. Symp.*, 2012, pp. 479–487.

[12] K. D. Wesson, D. P. Shepard, J. A. Bhatti, and T. E. Humphreys, "An evaluation of the vestigial signal defense for civil GPS anti-spoofing," in *Proc. 24th Int. Tech. Meeting Satell. Division Inst. Navigat.*, 2011, pp. 2646–2656.

[13] K. Wesson, M. Rothlisberger, and T. Humphreys, "Practical cryptographic civil GPS signal authentication," *Navigation*, vol. 59, no. 3, pp. 177–193, 2012.
[14] M. Pajic, J. Weimer, N. Bezzo, O. Sokolsky, G. J. Pappas, and I. Lee, "Design and implementation of attack-resilient cyberphysical systems: With a focus on attack-resilient state estimators," *IEEE Control Syst.*, vol. 37, no. 2, pp. 66–81, Apr. 2017.
[15] K. D. Wesson, B. L. Evans, and T. E. Humphreys, "A combined symmetric difference and power monitoring GNSS anti-spoofing technique," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2013, pp. 217–220.
[16] T. E. Humphreys, "Detection strategy for cryptographic GNSS anti-spoofing," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 2, pp. 1073–1090, Apr. 2013.
[17] M. L. Psiaki, B. W. O'Hanlon, J. A. Bhatti, D. P. Shepard, and T. E. Humphreys, "Civilian GPS spoofing detection based on dual-receiver correlation of military signals," in *Proc. 24th Int. Techn. Meet. Satellite Division Inst. Navig. (ION GNSS 2011)*, Portland, OR, USA, Sep. 20–23, 2011, pp. 2619–2645.
[18] M. Pajic, I. Lee, and G. J. Pappas, "Attack-resilient state estimation for noisy dynamical systems," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 82–92, Mar. 2017.
[19] R. Ivanov, M. Pajic, and I. Lee, "Attack-resilient sensor fusion for safety-critical cyber-physical systems," *ACM Trans. Embedded Comput. Syst.*, vol. 15, no. 1, Feb. 2016, Art. no. 21.
[20] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, Mar. 2005.
[21] D. Blacknell and H. Griffiths, *Radar Automatic Target Recognition (ATR) and Non-Cooperative Target Recognition (NCTR)*. London, U.K.: Inst. Eng. Technol., 2013.
[22] M. L. Cummings, S. Bruni, S. Mercier, and P. J. Mitchell, "Automation architecture for single operator, multiple UAV command and control," *Int. Command Control J.*, vol. 1, no. 2, pp. 1–24, 2007.
[23] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA, USA: MIT Press, 1992.
[24] S. Asmussen, *Applied Probability and Queues*, vol. 51. New York, NY, USA: Springer-Verlag, 2008.
[25] A. Pentland and A. Liu, "Modeling and prediction of human behavior," *Neural Comput.*, vol. 11, no. 1, pp. 229–242, 1999.
[26] A. Galata, N. Johnson, and D. Hogg, "Learning variable-length Markov models of behavior," *Comput. Vis. Image Understanding*, vol. 81, no. 3, pp. 398–413, 2001.
[27] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[28] X. Meng, K. K. Lee, and Y. Xu, "Human driving behavior recognition based on hidden Markov models," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2006, pp. 274–279.
[29] Y. Boussemart, J. Las Fargeas, M. L. Cummings, and N. Roy, "Comparing learning techniques for hidden Markov models of human supervisory control behavior," in *Proc. AIAA Infotech@ Aerosp. Conf. AIAA Unmanned... Unlimited Conf.*, 2009, Paper 1842.
[30] Y. Boussemart, M. L. Cummings, J. L. Fargeas, and N. Roy, "Supervised vs. unsupervised learning for operator state modeling in unmanned vehicle settings," *J. Aerosp. Comput., Inf., Commun.*, vol. 8, no. 3, pp. 71–85, 2011.
[31] C. E. Nehme, "Modeling human supervisory control in heterogeneous unmanned vehicle systems," Ph.D. dissertation, Dept. Aeronaut. Astronaut., Massachusetts Inst. Technol., Cambridge, MA, USA, 2009.
[32] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.
[33] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, 1966.
[34] G. Schwarz *et al.*, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[35] V. Rodríguez-Fernández, A. Gonzalez-Pardo, and D. Camacho, "Finding behavioral patterns of UAV operators using multichannel hidden Markov models," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2016, pp. 1–8.

Haibei Zhu received the B.S. degree in electrical engineering from Rensselaer Polytechnic Institute, NY, USA, in 2015. He is currently working toward the Ph.D. degree in computer engineering at Duke University, NC, USA.

He is currently a Research Assistant with the Duke Humans and Autonomy Lab. His research interests include human computer interaction, data mining, and operator strategy prediction.

Mary L. Cummings (SM'03) received the Ph.D. degree in systems engineering from the University of Virginia, Virginia, VA, USA, in 2004.

She is currently a Professor with the Duke University Department of Mechanical Engineering and Materials Science, the Duke Institute of Brain Sciences, and the Duke Electrical and Computer Engineering Department. She is the Director of the Duke Humans and Autonomy Laboratory.

Mahmoud Elfar received the B.Sc. degree in mechatronics from Ain Shams University, Cairo, Egypt. He is currently working toward the Ph.D. degree in computer engineering at Duke University, NC, USA.

His research interests are in formal methods, model checking techniques, and their applications in building human-aware cyber-physical systems.

Ziyao Wang received the B.S. degree in mechanical engineering from Jilin University, Jilin, China, in 2016, and the M.S. degree in mechanical engineering and material science from Duke University, NC, USA, in 2018.

He is currently working in the Humans and Autonomy Lab at Duke University. His current research interests include human robot interaction.

Miroslac Pajic (S'06–M'13) received the Dipl. Ing. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 2003 and 2007, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2010 and 2012, respectively.

He is currently the Nortel Networks Assistant Professor with the Department of Electrical and Computer Engineering at Duke University. He also holds a secondary appointment with the Computer Science Department. His research interests focus on the design and analysis of cyber-physical systems and in particular real-time and embedded systems, distributed/networked control systems, and high-confidence medical devices and systems.

Interactive Context-Aware Anomaly Detection Guided by User Feedback

Yang Shi, Maoran Xu, Rongwen Zhao, Hao Fu, Tongshuang Wu, and Nan Cao

***Abstract*—Automatic anomaly detection techniques have been extensively used to support decision making in abnormal situations. However, existing approaches are limited in their capacity of effectively identifying anomalies due to the complexity of the real-world environment, the uncertainty of the data input, and the unavailability of ground truth. In this paper, we propose an interactive context-aware anomaly detection algorithm framework that incorporates human judgment in searching for anomalous regions within a large geographic environment. In specific, our framework, 1) estimates a focal region and detect anomalous situations in real time, through which the user can observe and analyze suspicious entities, 2) leverages user feedback to refine results and guide further analysis, and 3) tolerates potential fault feedback provided by the users and resignal dubious anomalous points. Based on the framework, we propose two algorithm implementations, respectively, employ Bayes' theorem and metric learning. We demonstrate the effectiveness of the proposed framework and corresponding implementations through two controlled user studies and a case study with a domain expert.**

***Index Terms*—Anomaly detection, interaction techniques.**

I. Introduction

ANOMALY detection refers to the problem of identifying patterns in the data that do not conform to expected behavior [1]. A variety of anomaly detection systems have been developed for different purposes such as finding environmental changes [2], improving information security on social media [3], and monitoring traffic [4].

Due to its importance, anomaly detection techniques have been extensively researched. Existing techniques primarily approach the problem through automated analysis models including classification-based [5], clustering-based [6], statistical [7], [8], and spectral methods [9]. However, the computation results of analysis models are often imprecise and even misleading due to the complexity of the real-world environment, the uncertainty of the data input, and the unavailability of ground truth [10]. Consequently, interactive strategies have also been proposed to facilitate anomaly detection such as acquiring new or updated labels from users [11]. While such human-in-the-loop approaches significantly refine the analysis models, most of them are built on the assumption that all anomalies in the environment are completely observable. This assumption, however, overlooks the impact of *environment complexity* on detecting anomalies in real-world scenarios. Here, environmental complexity measures how difficult an anomaly in an environment can be observed by humans. For example, a polluted region with a small number of monitoring stations has a low probability that its anomalous situation can be observed. When an analyst investigates such a region with high environmental complexity, he or she may incorrectly perceive that the region has not been polluted.

In response to the aforementioned limitations, we propose a more reliable and practical interactive framework that utilizes human supervision to search for anomalous regions within a large geographic environment. Our framework has the following three main advantages.

1) The framework can estimate the environment and detect anomalous situations in real time, through which a user can observe and analyze suspicious entities.
2) The framework can leverage user feedback to refine results and guide further analysis. Due to the lack of ground truth, we use user domain knowledge as the essential source for refinements, meaning that a labeling on local environment will trigger global updates and thereby guide further analysis.
3) The framework can tolerate potential false feedback by introducing environmental complexity. When the user investigates a region with high environmental complexity, he or she may provide false feedback. The fault tolerant update mechanism can re-signal dubious anomalous entities and require further investigation from the user.

Based on the proposed framework, we propose two algorithms, respectively, employ Bayes' theorem and metric learning. Bayes' theorem is selected as it can calculate the conditional probability of finding an anomaly in a specific region immediately after receiving user feedback one at a time. On the other hand, metric learning can process multiple user feedback simultaneously and apply user feedback to the data for situation update.

Manuscript received May 1, 2018; revised October 15, 2018, December 21, 2018, and April 29, 2019; accepted June 11, 2019. Date of publication July 11, 2019; date of current version November 21, 2019. This work was supported in part by the Fundamental Research Funds for the Central Universities in China and in part by the National Natural Science Foundation of China under Grant 61802283 and Grant 61602306. This paper was recommended by Associate Editor L. Rothrock. *(Corresponding author: Nan Cao.)*

Y. Shi, H. Fu, and N. Cao are with the Intelligent Big Data Visualization Lab, Tongji University, Shanghai 201804, China (e-mail: yangshi.idvx@tongji.edu.cn; fuhao@icrd.com.cn; nan.cao@tongji.edu.cn).

M. Xu is with the University of Florida, Gainesville, FL 32611 USA (e-mail: maoranxu@ufl.edu).

R. Zhao is with the University of California, Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: rzhao17@ucsc.edu).

T. Wu is with the University of Washington, Seattle, WA 98195 USA (e-mail: wtshuang@cs.washington.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2925195

The results of our two user studies suggests that the framework can effectively support anomaly detection. Within the framework, the *Metric-Update* implementation outperforms the *Bayes-Update* in terms of the detection accuracy and time cost. In addition, an expert walk through in an air quality monitoring scenario suggests that our framework can provide solid supports for real use cases.

In summary, our work has the following main contributions.

1) *Algorithm framework:* We introduce an online interactive context-aware anomaly detection algorithm framework. The proposed framework provides an efficient anomaly detection mechanism that interactively adopts human judgment and includes environmental complexity into computations.
2) *Situation update algorithms:* Based on the proposed framework, we provide two algorithmic implementations that employ Bayes' theorem and metric learning, respectively, to help detect regional anomalies in the environment.
3) *Evaluation:* We conducted two controlled user studies, each having 18 participants, and a case study with a domain expert to verify the usefulness of the framework and compare the effectiveness of the two proposed algorithmic implementations.

In the rest of this paper, we first provide a brief survey of related work (see Section II). Then, we ground our motivation with a use scenario (see Section III), and then introduce the framework and two implementations (see Section IV). For evaluation, Section V describes the details and rationales of the experiment design, followed by the first study, analysis results, and discussions (see Section VI). Section VII describes the second user study and an expert walk-through based on the prototype system.

II. Related Work

Anomaly detection is a well-established field aiming at finding patterns in data that deviate from normal behavior [12]. In response to the indecisive "anomaly" definition problem, prior studies have also proposed interactive strategies to help practitioners personalize their detection models based on their own decision boundaries. Researchers have proposed multiple forms of feedback. For instance, Konijn and Kowalczyk [13] enabled users to iteratively label outliers until no more interesting outliers can be found. Krasuski and Wasilewski [14] improved the detection of outlying Fire Service's reports by discussing the features and decision boundaries with domain experts. Cao *et al.* presented TargetVue [11], a visualization system that displays the analysis results of anomalous online user behaviors and collects feedback from analysts. Liao *et al.* [15] developed GPSvas, which embeds an active learning process in its visual analysis, allowing users to input manual labels for further training. Online learning approaches has also been applied to anomaly detection. For example, Ahmad *et al.* [16] detected anomalies in streaming data using an online sequence memory algorithm. Similarly, Ozkan *et al.* [17] utilized the Markov model to detect anomaly in time series data. Bastani *et al.* [18] provided an efficient framework using sequential Monte Carlo for surveillance video. The aforementioned work helps efficiently detect anomalies in large scale, streaming data without spatial information. Laxhammar and Falkman [19] proposed the Sequential Hausdorff Nearest-Neighbor Conformal Anomaly Detector for online learning and sequential anomaly detection on geo-spatial trajectory data. Cao *et al.* [4] designed a visual interactive system, Voila, that uses Bayesian approach to detect anomalies in an urban context. Our interactive framework incorporates human supervision in searching for anomalous regions within a large geographic environment. The framework extends Voila's Bayesian approach and also introduces metric learning implementation, which is able to process multiple feedback simultaneously. Also, we use a fault tolerant mechanism by introducing environmental complexity in our algorithms to resignal dubious anomalous entities and require further investigation from the user.

III. Use Scenario

To better motivate the design of our framework, we first describe a use scenario. Suppose Alice, an urban security officer in the department of urban traffic monitoring and management, attempts to use an anomaly detection system to monitor urban traffic and find abnormal traffic incidents (e.g., the traffic flow within a region is significantly higher compared to its historical statistics). Alice first observes and analyzes the most suspicious regions ranked by the automated analysis models. However, she finds that the results are not aligned with her domain knowledge and, thus, refines the results by confirming or rejecting the anomalies detected by the system. After the refinement, the results are more accurate. However, there are tons of regions, manually checking each of them would be an onerous task. She expects that the system will use her labels as an indicator to update other regions in similar situations automatically. Unfortunately, the system fails to do so: the underlying algorithm is not designed for a spatial context, thus, all the dimensions are treated equally. Moreover, Alice notices that the system fails to consider cases when she incorrectly identifies anomalous cases and provides false feedback. The reason is that her judgment is based on the analysis of incomplete observed data, for example, traffic cameras are sparsely distributed among specific areas, resulting in difficulties in monitoring the situation for the system.

According to the use scenario, we identified the following design requirements for our system.

- *DR.1* *Updating analytics based on human feedback:* The system should provide an interaction mechanism that accepts feedback from users in real-time to dynamically rectify the anomaly detection model.
- *DR.2* *Tolerating potential false feedback:* Users might produce false feedback due to the difficulty in perceiving the environmental context. The system should include environment complexity into computations and better tolerate false feedback.
- *DR.3* *Allowing the integration of different anomaly detection algorithms:* To meet different detection requirements, any algorithm that yields the probability of finding anomalies in a given environment can be embedded into the framework.

IV. Algorithm Framework and Implementation

In this section, we introduce our interactive context-aware anomaly detection algorithm framework and two algorithm implementations, respectively, based on Bayes' theorem and metric learning.

Algorithm 1: Interactive Context-Aware Anomaly Detection Framework.

Input: $\mathbf{E} = \{r_1, ..., r_n\}$; $\mathbf{Q} = \{q_1, ..., q_n | q_i \in [0, 1]\}$

1 $\mathbf{P^{(0)}} = \{p_1^{(0)}, p_2^{(0)}, ..., p_n^{(0)}\} \leftarrow Initialization(\mathbf{E}, \mathbf{Q})$
2 **while** *true* **do**
3 **if** *user cannot find more anomalies* **then**
4 break;
5 **end**
6 $\mathbf{J} \leftarrow GetJudgement(\{(r_{i_1}, f_{i_1}), ...(r_{j_k}, f_{j_k}) | f_{y_x} \in \{0, 1\}\})$
7 $\mathbf{P^{(s)}} = \{p_1^{(s)}, ..., p_n^{(s)}\} \leftarrow update(\mathbf{J}, \mathbf{E}, \mathbf{Q}, \mathbf{P^{(s-1)}})$
8 **end**

A. Interactive Context-Aware Anomaly Detection Algorithm Framework

The goal of the framework is to help decision makers efficiently explore a large geographic environment and locate the regions that contain anomalous entities. Based on the design requirements, we achieve the goal by designing an anomaly detection mechanism that has the following characteristics:

1) interactively adopts human judgments to reflect users' decision boundary onto the global environment;
2) includes environmental contexts into computation;
3) light-weighted and independent of the underlying anomaly detection techniques.

As described in Algorithm 1, the environment E is uniformly partitioned into a set of n equal-size cells, i.e., n regions $\{r_1, r_2, \ldots, r_n\}$. We assign the environmental complexity Q to each of these n regions as $\{q_1, q_2, \ldots, q_n\}$. Then, the global environment E and the environmental complexity Q are used as the inputs for the probability initialization function $initialization(\cdot)$, which generates the initial probability P of finding an anomaly in each region as $\{p_1^{(0)}, p_2^{(0)}, \ldots, p_n^{(0)}\}$ (line 1). Every time when a user inspects a region, a binary feedback f_i indicating whether or not he or she finds an anomaly in the region r_i is recorded. Suppose that the user labels k regions in his inspection, $(r_{i_1}, f_{i_1}), \ldots, (r_{j_k}, f_{j_k})$ is used as the input to compute user judgment J (line 6). The probability update function $update(\cdot)$ receives the human judgment J, the global environment E, the environmental complexity Q, and the probability $P^{(s-1)}$ at step $s-1$ to assign the new probability $P^{(s)}$ of each region at step s (line 7). The iterative process described above continues until the user indicates that no more anomalies can be found (lines 3–5).

B. Implementation of Update Functions

We introduce two algorithm implementations based on the above-mentioned framework. The first algorithm employs Bayes' theorem [20], which handles user feedback one at a time. The second algorithm is designed based on metric learning [21], which is able to process multiple feedback simultaneously. *A priori* knowledge of the number of anomalies is required to use the first implementation while no such knowledge is required for the second implementation. Both of these implementations employ one-class support vector machines (SVM) [22] to generate initial probability values due to its benefits of unsupervised feature learning, computational efficiency, and a good performance [23].

1) Implementation I: The first implementation employs the Bayes' theorem [20] $P(A|B) = P(B|A)P(A)/P(B)$, which calculates the probability of event A given event B is observed. The Bayes' theorem updates the priori probability with the observation and yields the posterior probability. Therefore, it can be used in real world scenarios to search for regions of interest in a large environment [24]. In our case, we assume that there is only one anomaly in one of the investigation regions $\{r_1, \ldots, r_n\}$. Let A_i denote the event "an anomaly exists in the ith region" and let B_j denote the event "the user thinks that the jth region does not contain an anomaly." The probability $P(B_j)$ is positively correlated with the environmental complexity of this region, q_j. Thus, $P(A_i|B_j)$ represents the probability of an anomaly exists in the region i when a user perceives the region j as normal. Note that the assumption is simplified for computation and the Bayes' theorem based method also holds true when multiple anomalies are to be observed. We first normalize the probability for each region by the total number of anomalies. Every time when an anomaly is found, the denominator of the normalization is reduced by 1 and the probability of all the remaining regions are reduced with the same rate.

Based on the Bayes' theorem, we define the probability update function $update(\cdot)$ in Algorithm 1 as

$$\begin{aligned} \text{update}(r_i, p_i, q_i, f_j = 0) &= P(A_i|B_j) \\ &= \begin{cases} \frac{P(B_i|A_i)P(A_i)}{P(B_i)} = \frac{p_i q_i}{1-p_i(1-q_i)}, i = j \\ \frac{P(B_j|A_i)P(A_i)}{P(B_j)} = \frac{p_i}{1-p_j(1-q_j)}, i \neq j. \end{cases} \end{aligned} \tag{1}$$

Here, $P(A_i) = p_i$ indicates the probability of an anomaly in the region r_i. q_i and f_i represent the environmental complexity and user feedback of this region, respectively. Note that the above-mentioned update rules only take negative user feedback ($f_i = 0$) for update, that is, it updates the global situation only when the anomaly has not been found. Here, we assume that there is only one anomaly to be detected in the investigation space. Thus, a failed attempt of finding the anomaly at one region will increase the conditional probabilities of detecting it in other regions. Once the anomaly in one of the regions has been found ($f_i = 1$), this region r_i will be removed from the investigation space by setting $p_i \equiv 1$, and the user can start to find the next anomaly in the environment.

2) Implementation II: The second algorithm implementation embeds metric learning into one-class SVM. One-class SVM [22] computes the anomaly score using the distance from data point to the decision boundary, which can be modified by users' updates. Its key component, the kernel function, computes the distance in a hyperspace and projects the distance into the original data space. We modify the kernel based on metric learning [21]. Under the new metric, points in the same class will have small distance while points in different classes will have larger distance. Next, we will explain the how one-class SVM is implemented and how metric-learning is used to refine the distance function in one-class SVM that determines the anomaly score of each data point.

3) One-Class SVM: We use a set of data points $\{x_1, \ldots, x_m\}$ to represent entities to be observed in the environment E. Each data point $\mathbf{x}_i$ is associated with a multidimensional vector in the feature space. The data points are randomly distributed in

E, among which one or more data points are anomalies. The region that contains one or more anomalies is defined as an anomalous region. One-class SVM detects anomalies by finding a tight boundary in the feature space that encloses a majority of highly related data points, while points outside the boundary are identified as the anomalies. The distance between the points and the boundary indicates the anomaly score. The algorithm projects the data points into a latent hyperspace of a higher dimension, where the points can be easily separated by a hyperplane. Here, the equation representing the hyperplane is defined as

$$\sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x_i}) - \rho = 0 \tag{2}$$

where w_i and ρ are parameters learned from the input data, $\mathbf{x}$ and $\mathbf{x}_i$ represent the data point of interest and one of the data points in the dataset, respectively. The kernel function $K(\mathbf{x}, \mathbf{x_i})$ can be interpreted as estimating the distance between the two data points in a hyperspace. Usually, Gaussian kernel is used as it is the most frequent one used in nonlinear cases. The algorithm finds the tight boundary by ensuring the distance from a point to the hyperplane in the hyperspace to be equivalent to the distance from the same point to the boundary in the feature space. Here, the distance function is as follows:

$$\text{dist}(\mathbf{x}_i) = \sum_{j=1}^{m} w_j K(\mathbf{x_i}, \mathbf{x_j}) - \rho \tag{3}$$

where w_i and ρ are parameters learned from the input data, $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the data points in the dataset. Intuitively, when a data point lies on the hyperplane, the distance is equal to zero; otherwise, the distance has a positive or negative value, indicating the point lies inside (normal) or outside (abnormal) the boundary, respectively.

4) Metric Learning: To refine the distance function in one-class SVM, we first introduce a new kernel function (K_m) based one metric learning

$$K_m(\mathbf{x}, \mathbf{x_i}) = \exp\left(-\frac{d_M^2(\mathbf{x}, \mathbf{x_i})}{2\sigma^2}\right) \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}_i$ represent the data point of interest and one of the data points in the dataset, respectively. $d_M(\cdot)$ is a distance metric defined as

$$d_M(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{(\mathbf{x_i} - \mathbf{x_j})^T M (\mathbf{x_i} - \mathbf{x_j})}. \tag{5}$$

The goal of the update is to find one optimal matrix M^* that best matches the user judgment in terms of separating the normal and abnormal situations. To this end, we then employ the least-square metric learning (LSML) [25], in which M^* can be learned based on a set of constraints $\mathcal{C}$ in the form of

$$\mathcal{C} = \{(\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}, \mathbf{x_d}) : d_M(\mathbf{x_a}, \mathbf{x_b}) < d_M(\mathbf{x_c}, \mathbf{x_d})\} \tag{6}$$

where $\mathbf{x}_a$ and $\mathbf{x}_b$ are data points from the same class while $\mathbf{x}_c$ and $\mathbf{x}_d$ are data points in different classes. $\mathcal{C}$ ensures that the pairs of points within the same class to be closer than the pairs from different classes. $\mathcal{C}$ can be automatically generated based on user feedback f_i, the probability p_i, and the environmental complexity q_i for the ith region. Here, $\mathbf{M}(f_i, p_i, q_i)$ is used to denote the function that generates the optimal matrix M^* and Algorithm 2

Algorithm 2: Metric Learning Algorithm for $\mathbf{M}(f_i,p_i,q_i)$.

Input: f_i, p_i, q_i, $\mathcal{C} = \varnothing$;
1 $\mathcal{A} = \{r_j | p_j \in \mathbf{P}, p_j > threshold\}$;
2 $\mathcal{N} = \{r_k | p_k \in \mathbf{P}, p_k \leq threshold\}$;
3 **for** *j=1 to NumOfConstraints* **do**
4 $a, b, c \leftarrow sample(\mathcal{N})$, $d \leftarrow sample(\mathcal{A})$;
5 $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{(a, b, c, d)\}$
6 **end**
7 **if** $f_i == 1$ **then**
8 $\mathcal{A} = \mathcal{A} \bigcup r_i$;
9 **for** $j = 1$ *to* n **do**
10 $a \leftarrow sample(\mathcal{A})$, $b \leftarrow sample(\mathcal{N})$;
11 $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{(r_i, a, r_i, b)\}$, with probability $1 - q_i$;
12 **end**
13 **else**
14 $\mathcal{N} = \mathcal{N} \bigcup r_i$;
15 **for** $j = 1$ *to* n **do**
16 $a \leftarrow sample(\mathcal{N})$, $b \leftarrow sample(\mathcal{A})$;
17 $\mathcal{C} \leftarrow \mathcal{C} \bigcup \{(r_i, a, r_i, b)\}$, with probability $1 - q_i$;
18 **end**
19 **end**
20 $\mathbf{M}^* \leftarrow LSML(\mathcal{C})$;

describes the metric learning algorithm for $\mathbf{M}(f_i, p_i, q_i)$. $\mathcal{A}$ denotes the abnormal training set while $\mathcal{N}$ denotes the normal training set, p_j is the probability of observing an anomaly in the region j. If p_j is greater than a *threshold*, there is a high probability that the region j contains an anomaly. These high-anomaly regions constitute set $\mathcal{A}$ while those low-anomaly regions constitute set $\mathcal{N}$ (lines 1–2). In our case, we set the threshold as 0.5. The reason is that the values of p are evenly distributed between [0, 1], so the midpoint, 0.5, is selected. Accordingly, $\mathcal{A}$ and $\mathcal{N}$ are of roughly the same size. *NumOfConstraints* is a parameter that fixes the sample size of constraints, which is empirically set to the number of regions in our case. a, b, c, and d are sampled from $\mathcal{A}$ and $\mathcal{N}$ (lines 3–6). Here, $sample(\cdot)$ is a random sampling. If the user detects an anomaly in the region r_i (line 7), r_i is appended to the set $\mathcal{A}$ (line 8). Then, the new constraints (r_i, a, r_i, b) reported by the user are added to the constraints collection $\mathcal{C}$ with probability $1 - q_i$, meaning that the user judgment on the regions with less environment complexity are more reliable and these regions are more likely to change the metric (lines 9–12). Otherwise, the region r_i is appended to the set $\mathcal{N}$ and the related constraints are added likewise (lines 13–19).

Based on the (3)–(5), the probability update function $update(\cdot)$ in Algorithm 1 is as follows:

$$\text{update}(r_i, p_i, q_i, f_i) = N\left(\sum_{j=1}^{m} w_j \exp\left(-\frac{(\mathbf{x_i} - \mathbf{x_j})^T \mathbf{M}(f_i, p_i, q_i)(\mathbf{x_i} - \mathbf{x_j})}{2\sigma^2}\right) - \rho\right) \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ represent data points in the dataset, $\mathbf{M}(f_i, p_i, q_i)$ denotes the function that generates the optimal matrix M^*, $N(\cdot)$ normalizes the results into $[0, 1]$. σ is a free parameter to tune the fitness and smoothness of the decision boundary.

In our case, σ is set it to 1. $N(\cdot)$ is conducted on all the regions whenever the probability is updated. We use min–max normalization on each region and normalize the results into [0, 1]. The formula of normalization is: $N(r_i) = (p_i - p_m)/(p_M - p_m)$, where p_i is the anomaly score for region r_i, p_m, and p_M denote the minimum and maximum scores among the n regions, respectively.

V. Experiment Design

To evaluate the effectiveness of the algorithm framework in supporting detect anomalies and compare the performance of the two algorithm implementations under different conditions, we designed a controlled user study. In this section, we describe the details and rationales of the experiment design.

A. User Task

We design tasks that simulate real-world scenarios of anomaly detection, in which a small portion of potential anomalies needs to be found among a large collection of data points distributed in a two-dimensional (2-D) area. Users are required to find the anomalies and label the regions that contain these anomalies. Hence, the user task is described as follows.

Locate the regions that contain anomalous entities (i.e., the data points that have different feature values compared with that of other points) in a 2-D spatial environment.

In this task, the primary variable to be tested is the choice of the algorithms (i.e., *No-Update* where anomaly detection is implemented without situation update (baseline), *Bayes-Update*, and *Metric-Update*). Additionally, the study tested two more variables including 1) the scale of the investigation space, which is determined by the number of grids (denoted as g^2) and 2) the number of anomalies to be found in the investigation space (denoted as n_a). We determine the proper setting of the two variables g^2 and n_a by conducting a pilot study with 20 participants. Based on the results of the pilot study, the number of grids g^2 is set to either 10^2 (small) or 15^2 (large), and the number of anomalies n_a is set to either 3 (small) or 7 (large). In the formal user study, we defined four task including *T1 (G10-A3):* finding three anomalies in 10×10 grids, *T2 (G10-A7):* finding seven anomalies in 10×10 grids, *T3 (G15-A3):* finding three anomalies in 15×15 grids, and *T4 (G15-A7):* finding seven anomalies in 15×15 grids. Each task is repeated for three times to reduce random noise.

B. Dataset

We synthesized a testing dataset simulating all of the aforementioned task conditions. We first generated a collection of data points from multivariate Gaussian distribution with mean $\mu = (1, 1, 1, 1, 1, 1)^T$ and covariance matrix $\Sigma = \text{diag}\{1, 1, 1, 1, 1, 1\}$. These data points are grouped into the normal points (distance from the mean μ were less than 3σ) and abnormal ones (distance greater than 3σ). Each data point is associated with a 6-D vector. We then placed these points randomly in a 2-D plane. The probability p is initialized by the anomaly score output from one-class SVM with an untrained Gaussian Kernel.

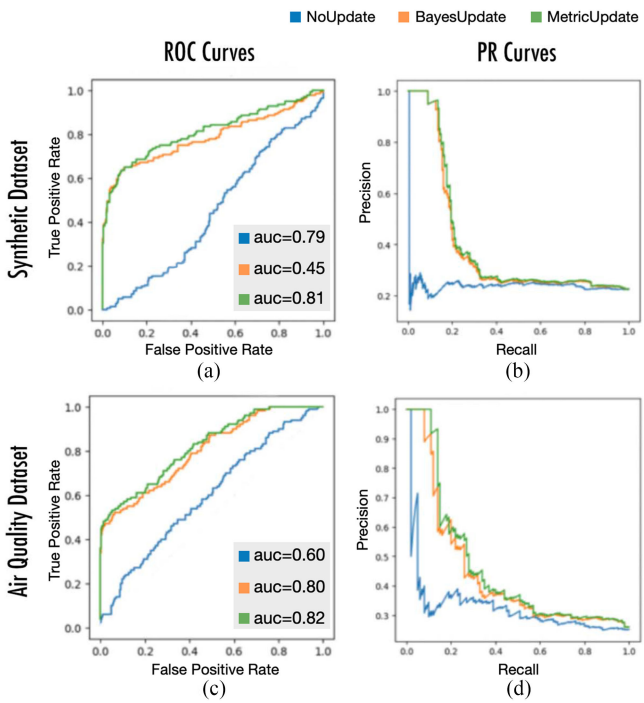


Fig. 1. Evaluation curves of the three algorithms. (a) and (b) Drawn from the testing dataset. (c) and (d) Drawn from the China air quality dataset on January 15, 2017.

C. Preliminary Test

We first ran a preliminary test to evaluate the performance of the three algorithms (*No-Update*, *Bayes-Update*, *Metric-Update*) on the testing dataset when no human interaction is involved. We let the three algorithms always pick the most anomalous region and make an update of the current situation. Receiver operating characteristic (ROC) curves were then generated by comparing the perceived labels with the original labels of these updated regions. Fig. 1(a) and (b) shows that both *Bayes-Update* and *Metric-Update* outperform *No-Update* while *Bayes-Update* is slightly better than *Metric-Update*.

D. Study Hypotheses

Our hypotheses were formed as follows.

- *H.1* The algorithm framework will enhance the user performance of detecting anomalous entities in the environment.
- *H.2* Users will achieve higher task accuracy in detecting anomalous regions using *Metric-Update* than *Bayes-Update* and *No-Update*.
- *H.3* User will spend less completion time in detecting anomalous regions using *Metric-Update* than *Bayes-Update* and *No-Update*.

The preliminary test suggests that *Metric-Update* outperforms *Bayes-Update* and *No-Update* when no human interaction is involved. As human might have different observations regarding our visualization and interaction design, we posed *H.2* to further compare task accuracy of the three algorithms when user interaction is involved. We hypothesized that users would spend less completion time using Metric-Update (*H.3*) as it accepts

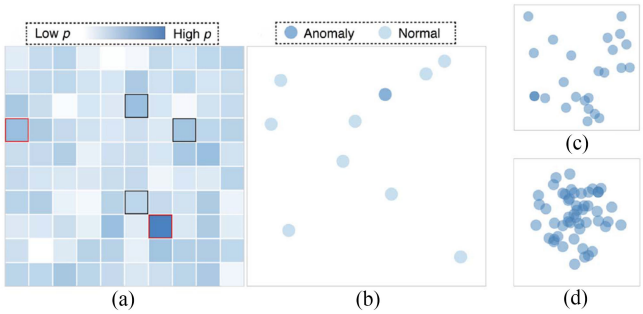


Fig. 2. User study interface. (a) Global view. (b) Detail view. The results in the detail view with different environmental complexity. (c) $q = 0.1$. and (d) $q = 0.8$.

multiple feedback simultaneously. Based on the accuracy (*H.2*) and time (*H.3*), we posed the overall hypothesis *H.1* that our algorithm will enhance user performance.

E. Task Performance Measures

To quantify user performance of detecting anomalies via different algorithms, we use task accuracy and completion time.

Task Accuracy: There are two measures of accuracy: the precision and recall rate. Our pilot study showed that the precision rate was not a proper measure to reflect the accuracy compared to the recall rate, as users rarely made mistakes in identifying anomalous grids. As a result, the recall rate was selected as the measure of accuracy in the user study.

Completion time: The completion time measures the duration starting at the time when the dataset is displayed to users, and stopping at the time when users click the "next" button to begin the next trial, which contains both the inspection time and response time. Users can click the "pause" button when having a break, during which the time recorder is held up.

F. User Interface

To evaluate how well each algorithm in our framework help users detect anomalies in the environment, we design an user interface with two coordinated views, the global view and detail view, as shown in Fig. 2.

Global view: The global view [see Fig. 2(a)] displays an overview of the anomalous information in the environment. We overlaid equal-sized grids to uniformly partition the environment to be investigated. Each grid represents a region, with the color illustrating the probability of containing an anomaly in this region. A grid with darker blue indicates that the region has a higher probability of containing an anomaly (i.e., high p value).

Detail view: The detail view [see Fig. 2(b)] shows individual entities in a specific region that the user has selected in the global view. Each data point represents an entity, with the opacity illustrating whether the entity is an anomaly. An opaque blue point indicates an anomaly while a translucent blue point encodes a normal entity. Here, we selected opacity to differentiate between anomalous and normal entities as it can be also used to show environmental complexity q through the overlapping rate of the points. That is, a set of highly overlapped translucent points will result in difficulties in identifying opaque ones. For example, by comparing Fig. 2(c) and (d), we found that the region with higher environment complexity [see Fig. 2(d)] is more difficult for users to identify the anomaly (i.e., opaque blue point).

We control the environmental complexity via the distribution of the data $N(0, 1 - q)$, where $N(\cdot)$ is the normal distribution. In the user study, the environmental complexity of each grid in each trial is randomly determined.

During the process of anomaly detection, a user can first investigate the most suspicious region in the global view by hovering on the grid that has the highest p value, that is, the region shown in the darkest blue color. Once the focal region is identified, he or she can then inspect its data points in the detail view. Based on his/her judgments of these data points, the user can label the grid in the global view; a single-click indicates the region indeed contains an anomaly, which can be canceled by a double-click. Additionally, the grids that have been hovered by the user are marked with thick black borders to avoid duplicated investigations. The user can also submit his/her judgments (i.e., the labels of grids) to the back end by a right-click whenever he or she wants, the algorithms will automatically update the p value for each grid and its corresponding color based on user feedback.

Before integrating the visual interface into the framework, the pilot study also investigates 1) whether the environmental complexity determined by $N(0, 1 - q)$ align with users' perception and 2) the "accuracy-q" correlation regarding different number of data points in a grid. The results suggest that our design of environmental complexity align with users perception and 30 points per grid is the best setting.

VI. User Study I

In this section, we first describe the study method, followed by the analysis results and discussions.

A. Method

We recruited 18 participants (9 females) with an average age of 22.06 (SD = 1.95) for the user study with the goal of evaluating and comparing three algorithms, *No-Update*, *Bayes-Update*, and *Metric-Update*, implemented in the interactive context-aware anomaly detection framework. Before the study, we conducted a 20-min tutorial session, during which the concept of anomaly detection and its application in real-world scenarios were briefly introduced. Next, we described in detail the framework with the proposed algorithms, the user interface, and the interactions. In the practice session, the participants were instructed to use the system with a sample dataset.

The study consisted of three sessions, each of which involved one of the three algorithms. In each session, participants completed four tasks, *T1 (G10-A3)*, *T2 (G10-A7)*, *T3 (G15-A3)*, *T4 (G15-A7)*). Both the task accuracy and completion time were recorded automatically for later analysis. We counterbalanced the order of the three sessions as well as the order of four tasks to avoid learning effects. Upon the completion of the three sessions, we asked the participant to complete a questionnaire. The user study took approximately 45–55 min.

This study was performed on a 13.3-in laptop computer with a display resolution of 2560 × 1600 and each trial was displayed in a 2000 × 1200 window with a white background. The size of

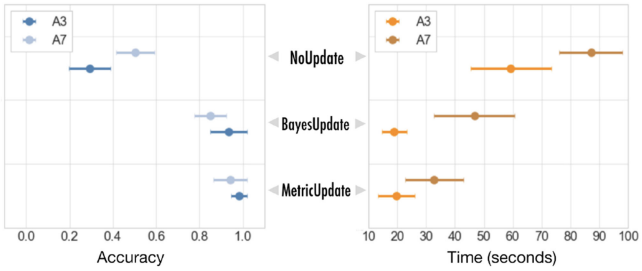


Fig. 3. (a) Accuracy and (b) completion time for 10 × 10 grids (small).

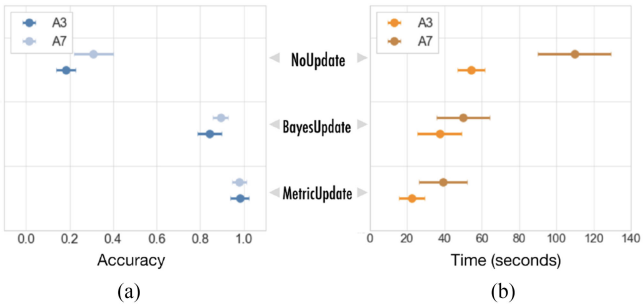


Fig. 4. (a) Accuracy and (b) completion time for 15 × 15 grids (large).

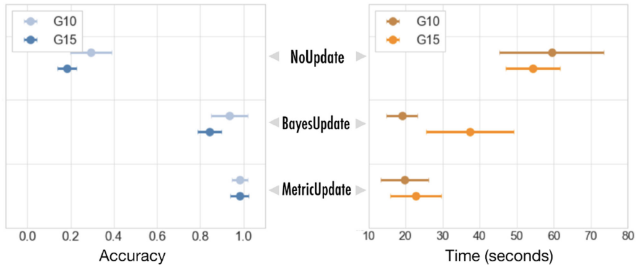


Fig. 5. (a) Accuracy and (b) completion time for three anomalies (small).

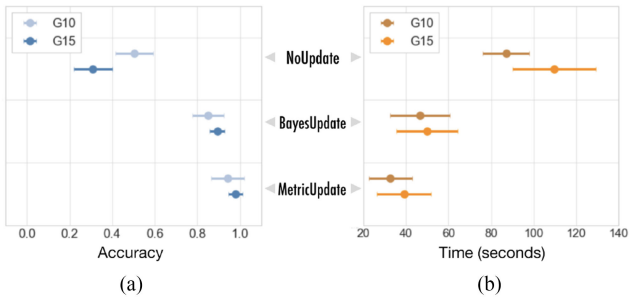


Fig. 6. (a) Accuracy and (b) completion time for seven anomalies (large).

each grid was adjusted automatically according to the number of grids g^2.

B. Result Analysis

We now report the quantitative results from the above-mentioned user study. We first analyze the effect of two study variables (number of the grids and anomalies) on the task performance. We then compare the accuracy and completion time of three algorithms (*No-Update*, *Bayes-Update*, and *Metric-Update*). Finally, we show the results form the poststudy questionnaire. Repeated Measures ANOVA (RM-ANOVA) was applied to examine if there is a significant difference. Bonferroni correction was used to conduct the pairwise comparisons.

1) Validation of Variables: To evaluate the effect of the number of grids and anomalies, we analyzed user performance under different conditions.

Small grid number (10 × 10): RM-ANOVA shows that the number of anomalies significantly affected user performance in terms of the accuracy ($F(2, 34) = 15.81, p < 0.05$) across all three algorithms [see Fig. 3(a)]. The analysis results showed no significant difference in completion time [see Fig. 3(b)]. Compared to *No/Bayes-Update*, *Metric-Update* was the least sensitive to the change of anomaly numbers in both accuracy and time.

Large grid number (15 × 15): Fig. 4(a) illustrates that the accuracy was significantly lower when the number of anomalies increased ($F(2, 34) = 12.65, p < 0.05$) across all algorithms. Fig. 4(b) suggested that the completion time was also significantly influenced by the anomaly number ($F(2, 34) = 68.23, p < 0.01$). RM-ANOVA showed that *Metric-Update* and *Bayes-Update* were less influenced in both task accuracy and completion time than *No-Update*.

Small anomaly number (3): Fig. 5(a) shows that the accuracy of both *No-Update* and *Bayes-Update* significantly decreased (*No-Update:* $F(2, 34) = 19.95, p < 0.05$; *Bayes-Update:* $F(2, 34) = 10.17, p < 0.05$) when the grid number increased. The accuracy of *Metric-Update* was less sensitive to the number of the grids. In terms of completion time, only *Bayes-Update* was significantly influenced ($F(2, 34) = 63.79, p < 0.01$) [see Fig. 5(b)].

Large anomaly number (7): Metric/Bayes-Update showed no significant difference between the two grid numbers while *No-Update* was significantly influenced in terms of both accuracy ($F(2, 34) = 23.28, p < 0.01$) [see Fig. 6(a)] and completion time ($F(2, 34) = 46.36, p < 0.01$) [see Fig. 6(b)].

2) Comparison of Algorithms: We compare the task accuracy and completion time of the three algorithms under four task conditions, T1,T2,T3, and T4.

Task accuracy: When grid number was either small or large, significant differences were observed among the three algorithms in the two levels of anomaly number, as shown in Fig. 7. Moreover, post-hoc analysis showed that in most of the cases (T1 (G10-A3): $F(2, 34) = 6.05, p < 0.05$, T2 (G10-A7): $F(2, 34) = 34.11, p < 0.01$, T4 (G15-A7): $F(2, 34) = 21.36, p < 0.01$), *Metric-Update* significantly outperformed *Bayes-Update* in accuracy (*H.2* accepted).

Completion time: Fig. 8 showed a significant difference among the three algorithms. Those showing a difference were associated with grid number as well as anomaly number. Moreover, a post-hoc analysis showed that in most of these cases (i.e., T1 (G10-A3): $F(2, 34) = 79.16, p < 0.01$, T2 (G10-A7): $F(2, 34) = 98.32, p < 0.01$, T4 (G15-A7)): $F(2, 34) = 147.21, p < 0.01$, *Metric-Update* significantly outperformed *Bayes-Update* in completion time (*H.3* accepted). As a result, we verified that our framework significantly enhances user performance in terms of accuracy and time (*H.1* accepted).

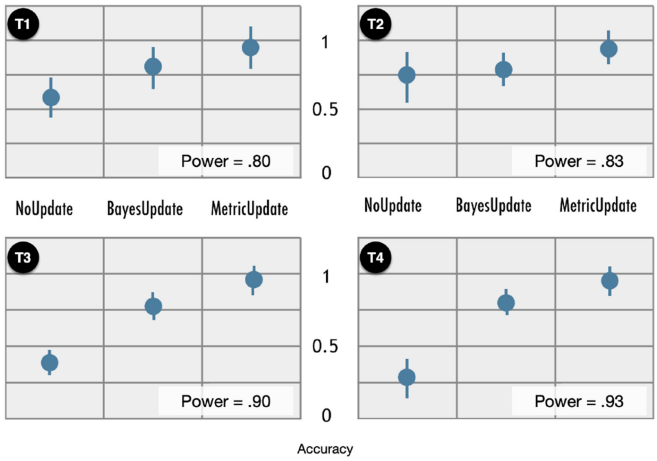


Fig. 7. Accuracy for four user tasks, including T1 (G10-A3): finding three anomalies in 10 × 10 grids, T2 (G10-A7): finding seven anomalies in 10 × 10 grids, T3 (G15-A3): finding three anomalies in 15 × 15 grids, and T4 (G15-A7): finding seven anomalies in 15 × 15 grids.

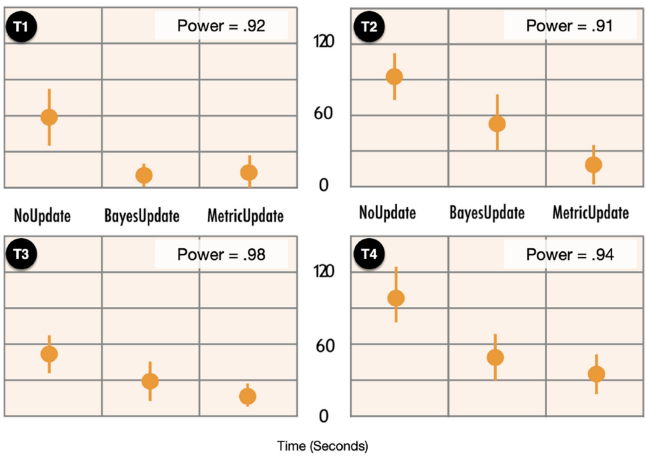


Fig. 8. Completion time for user four tasks, including T1 (G10-A3): finding three anomalies in 10 × 10 grids, T2 (G10-A7): finding seven anomalies in 10 × 10 grids, T3 (G15-A3): finding three anomalies in 15 × 15 grids, and T4 (G15-A7): finding seven anomalies in 15 × 15 grids.

3) Poststudy Questionnaire: The poststudy questionnaire was designed to qualitatively estimate the three algorithms. Questions 1–6 asked users to rate the ease of use and usefulness of each method for anomaly detection using a five-point Likert scale, as shown in Fig. 9(a). Questions 7–10 asked users to choose the method that they thought the most effective under various task conditions (i.e., larger or smaller grid number and larger or smaller anomaly number), as shown in Fig. 9(b). The results suggest that *Metric-Update* is favored by most of the participants, which supports the quantitative findings.

C. Discussion

We now discuss why and when *Metric/Bayes-Update* are useful and what are the challenges of using the framework.

Why did Metric-Update outperform Bayes-Update? According to the statistics, *Metric-Update* outperformed *Bayes-Update* on both the accuracy and completion time. *Bayes-Update* uses human judgment as an observation to update prior anomaly

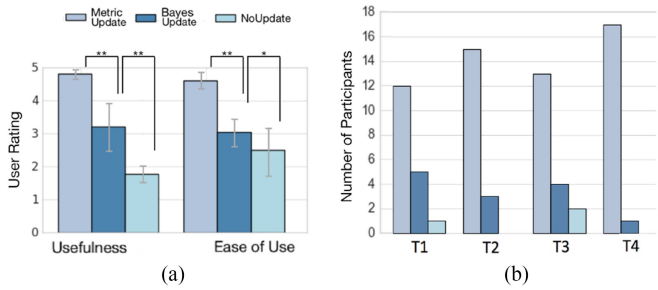


Fig. 9. Questionnaire statistics. (a) Users' rating of different algorithms with respect to their usefulness and ease of use (**: $p < 0.01$ and *: $p < 0.05$). (b) Users' preference of the three algorithms.

scores. Equation (1) also shows that *Bayes-Update* updates the anomaly scores of unchecked grids ($i \neq j$) with the same factor, resulting in the same level of color changes in grids. These changes may be difficult for users to perceive. *Metric-Update*, on the other hand, involves human judgment and data features into calculation. It uses human judgment as additional labels for training and update the hyperplane that separates abnormal and normal data points and, thus, achieves higher accuracy. In terms of completion time, users suggested that it was time consuming to update after each check when using *Bayes-Update* or to randomly guess when using *No-Update*. On the other hand, *Metric-Update* adapts a simultaneous update strategy and, thus, costs less time.

When should Metric-Update be used? Fig. 9 shows that *Metric-Update* was the most preferred method across all conditions. In terms of robustness, it was also less sensitive to the variation of the scale of the investigation space and the number of anomalies (see Figs. 3–6). Therefore, *Metric-Update* is recommended as the first choice in most real-world scenarios, especially in the complex and large-scale cases.

When should Bayes-Update be used? Fig. 9 shows that the preference for *Bayes-Update* was higher in T1 compared to T2, T3, and T4. Some participants suggested that it immediately responded to the clicks and, thus, provided good user experience. The reason is that Bayes-Update does not use feature values of data for calculation in each update, resulting in quick response. Figs. 7 and 8 also illustrate that *Bayes-Update* achieved good performances in T1 and T3 where the amount of anomalies is small. Therefore, *Bayes-Update* is applicable in simple cases due to its quick response.

VII. User Study II

To further evaluate the effectiveness of the framework and algorithms in real-world scenarios, we conducted another user study with 18 participants and an interview with an expert who is a professor in environmental engineering using a real-world dataset. Both the participants and the expert were required to monitor air quality in China and find anomalous regions where the air has been polluted.

A. Dataset

We used the air pollutant concentration dataset (http://pm25.in) collected from more than 1400 air quality monitoring stations

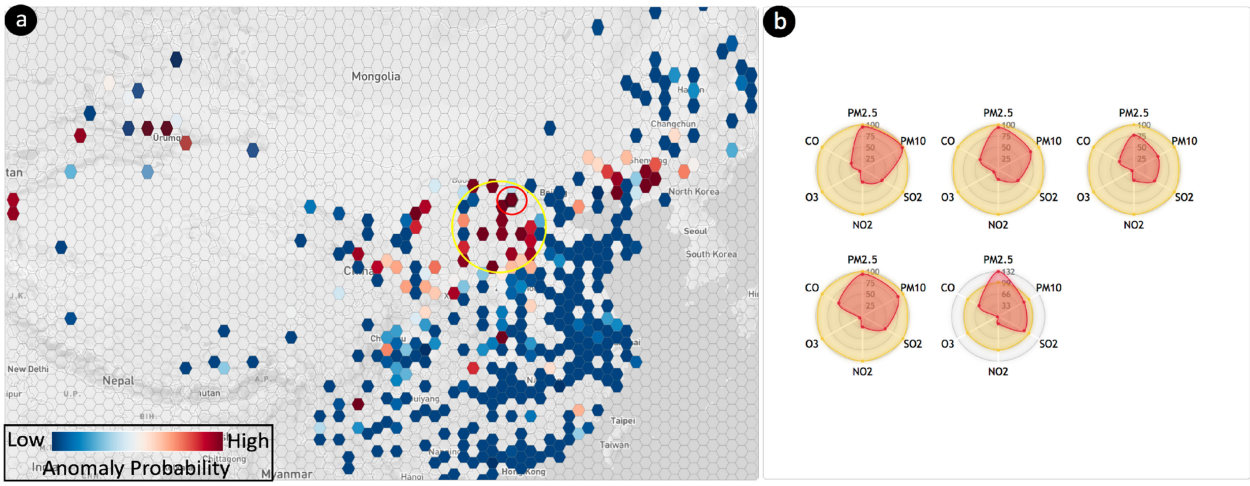


Fig. 10. User interface of the prototype system, iDetector, consists of two major views. (a) Global view. (b) Detailed view.

located in 367 urban cities as the real-world dataset. These monitoring stations record six pollutants, including nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), and particulate matter (PM2.5, PM10).

We first standardized the value of each pollutant based on the individual air quality index (IAQI) to facilitate a comparison. Then, we used a 6-D feature vector to capture air quality in a region recorded by the local monitoring station. The feature vector indicates the average IAQI value of the six pollutants. Specifically, the likelihood of the air in a region been polluted p is calculated using one-class SVM by comparing the features of the region to that of other regions as well as the historical data. The environmental complexity q is negatively proportional to the number of air quality monitoring stations inside the region (note that q is not visualized in iDetector). Based on the air quality index (AQI) and health implications, the monitoring stations whose AQIs are above 200 are defined as anomalies.

B. iDetector

We designed an interactive anomaly detection system, iDetector, for our second user study. iDetector consists of two coordinated views, the global view and detail view. The global view [see Fig. 10(a)] display a map of China overlaid with equal-sized hexagonal grids that uniformly segment the environment. It shows an overview of anomalous regions with the color of each grid illustrating p of a region. A region's color is blending between red and blue to, respectively, encode the high and low anomaly score. Gray grids indicate that no data have been collected from those regions.

Fig. 10(b) shows the detail view of air quality in a focal region. iDetector visualizes air quality recorded by each monitoring station as a radar chart with each axis indicating a pollutant. In the glyph, the current situation of air quality is drawn in red in the foreground while the standard baseline is drawn in yellow in the background. This design facilitates a fast comparison and allows users to quickly identify an anomaly when the current score is greater than the baseline (with the red region exceeding the yellow boundary).

C. User Study and Result Analysis

The second user study follows the same protocol used in the first user study. A total of 18 participants (11 females) with an average age of 24.94 (SD = 2.60) were recruited. In each session, participants used iDetector with one of the three algorithms implemented (*No-Update*, *Bayes-Update*, *Metric-Update*) to explore the air quality dataset. Note that the performance of the three algorithms on the real-world dataset was evaluated using ROC/PR curves [see Fig. 1(c) and (d)].

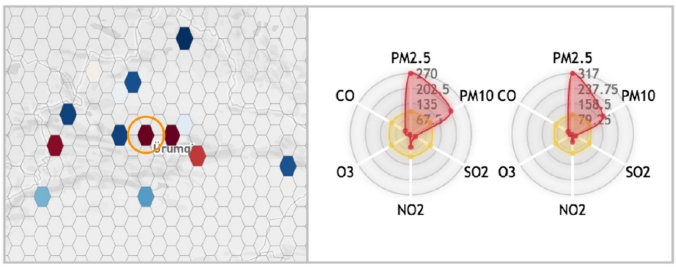
Task accuracy: The results show significant difference in accuracy ($F(2, 34) = 59.14, p < .01$). The post-hoc test suggests that participants achieved significantly higher accuracy when using *Metric-Update* than *Bayes-Update* and *No-Update* (*H.2* accepted).

Completion time: Significant difference is found in time ($F(2, 34) = 38.85, p < .01$). The post-hoc test suggests that participants spent significantly less time using *Metric-Update* than *Bayes-Update* and *No-Update* (*H.3* accepted). As a result, we verified that our framework significantly enhances user performance in terms of accuracy and time (*H.1* accepted).

D. Expert Interview

The expert interview used the air quality dataset on January 15, 2017 and iDetector with *Metric-Update* implemented. During the process, the expert first identified the most suspicious regions at first glance in the global view. "It is intuitive, I searched for the regions shown in the darkest red, and it turned out to be Henan Province based on the geographic information" [highlighted via a yellow circle in Fig. 10(a)]. He then hovered one of these regions with the highest anomaly score (highlighted via a red circle) and explored its detail view [see Fig. 10(b)].

The detail view shows radar charts that visualize the air quality recorded by the five monitoring stations in the focal region. By comparing the current situation (red) with the baseline (yellow) in each chart, the expert found that even though the values of several pollutants are high, most of them are within the normal range. He considered it as a normal situation, and thus, double-clicked to reject the result and label the region as



(a) (b)

Fig. 11. Anomalous region with high air pollution risk (a region in Xinjiang, China) revealed in (a) global view and (b) detail view.

normal and right-clicked to commit the change to iDetector for situation update. This update resulted in color changes in other regions in the global view. "I noticed that some regions turned from red to blue." The expert explored these grids and found that the regions now turned to normal were the ones similar to the region he had labeled. "I see, it learned from what I did and made similar judgment automatically. It is smart!" The expert also noticed that the color of some grids turned to darker red. He investigated these grids in the detail view and found that the radar charts display extremely high PM2.5 and PM10 values compared to the baseline. These regions were then marked as anomalies by the expert and the global environment was update accordingly.

Another suspicious area that caught the expert's attention is in Xinjiang Province shown in dark red in Fig. 11(a) (highlighted via an orange circle). He explored a specific region in this area in the detail view [see Fig. 11(b)]. By evaluating the pollutant values, he confirmed that this region had been polluted with high PM2.5 and PM10, and thus, single-click to mark it as the anomalous region. After situation update, he noted that "this time my judgment did not result in obvious color changes in other areas and I was wondering why." We explained to him that his decision is of low confidence due the high environmental complexity q in this region, that is, data collected from only two monitoring stations is available for analysis. Knowing this reason, the expert was impressed by our technique and suggested "this is brilliant! I can imagine this mechanism to be used in many applications in practice."

VIII. Conclusion

In this paper, we introduce an online interactive algorithm framework to support the analysis process of anomaly detection and two algorithm implementations based on Bayes and metric learning, respectively. The results of the two user studies indicate that the framework is useful to identify regions that contain anomalous entities and the *Metric-Update* algorithm significantly outperforms the *Bayes-Update* algorithm and the baseline in terms of accuracy and completion time. The case study with a domain expert further verified the usefulness of the proposed technique. Future work includes refining the algorithm to provide smoother update results, capturing temporal and spatial features for analysis, and applying our technique to more real-world applications.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.
[2] J. H. Faghmous and V. Kumar, "Spatio-temporal data mining for climate data: Advances, challenges, and opportunities," in *Proc. Data Mining Knowl. Discovery Big Data*, 2014, pp. 83–116.
[3] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins, "# fluxflow: Visual analysis of anomalous information spreading on social media," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1773–1782, Dec. 2014.
[4] N. Cao, C. Lin, Q. Zhu, Y.-R. Lin, X. Teng, and X. Wen, "Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 23–33, Jan. 2018.
[5] V. Roth, "Kernel fisher discriminants for outlier detection," *Neural Comput.*, vol. 18, no. 4, pp. 942–960, 2006.
[6] S. Budalakoti, A. N. Srivastava, R. Akella, and E. Turkov, "Anomaly detection in large sets of high-dimensional symbol sequences," NASA Ames Res. Center, Mountain View, CA, USA, Tech. Rep. NASA TM-2006-214553, 2006.
[7] Z. Ju and H. Liu, "Fuzzy gaussian mixture models," *Pattern Recognit.*, vol. 45, no. 3, pp. 1146–1158, 2012.
[8] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, vol. 589. Hoboken, NJ, USA: Wiley, 2005.
[9] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and B. Maglaris, "Hierarchical anomaly detection in distributed large-scale sensor networks," in *Proc. IEEE Symp. Comput. Commun.*, 2006, pp. 761–767.
[10] D. Overby, J. Wall, and J. Keyser, "Interactive analysis of situational awareness metrics," *Vis. Data Anal.*, vol. 8294, 2012, Art. no. 829406.
[11] N. Cao, C. Shi, S. Lin, J. Lu, Y.-R. Lin, and C.-Y. Lin, "Targetvue: Visual analysis of anomalous user behaviors in online communication systems," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 280–289, Jan. 2016.
[12] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.
[13] R. M. Konijn and W. Kowalczyk, "An interactive approach to outlier detection," in *Proc. Int. Conf. Rough Sets Know. Technol.*, 2010, pp. 379–385.
[14] A. Krasuski and P. Wasilewski, "Outlier detection by interaction with domain experts," *Fundam. Informaticae*, vol. 127, no. 1–4, pp. 529–544, 2013.
[15] Z. Liao, Y. Yu, and B. Chen, "Anomaly detection in GPS data based on visual analytics," in *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, 2010, pp. 51–58.
[16] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.
[17] H. Ozkan, F. Ozkan, and S. S. Kozat, "Online anomaly detection under Markov statistics with controllable type-i error," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1435–1445, Mar. 2016.
[18] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric Bayesian activity mining and analysis from surveillance video," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2089–2102, May 2016.
[19] R. Laxhammar and G. Falkman, "Online learning and sequential anomaly detection in trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1158–1173, Jun. 2014.
[20] A. O'Hagan and J. J. Forster, *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, vol. 2. London, U.K.: Arnold, 2004.
[21] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.
[22] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, no. 11–13, pp. 1191–1199, 1999.
[23] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[24] T. Furukawa, F. Bourgault, B. Lavis, and H. F. Durrant-Whyte, "Recursive bayesian search-and-tracking using coordinated UAVs for lost targets," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 2521–2526.
[25] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang, "Metric learning from relative comparisons by minimizing squared residual," in *Proc. IEEE Int. Conf. Data Mining*, 2012, pp. 978–983.

Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability

Yidu Lu and Nadine Sarter

***Abstract*—Trust miscalibration, a mismatch between a person's trust in automation and the system's actual reliability, can lead to either misuse or disuse of automation. Existing techniques to measure trust (e.g., subjective ratings) tend to be discrete and/or disruptive. To better understand and support the process of trust calibration, a nonintrusive continuous measure of trust is needed. The present study investigated whether eye tracking can serve this purpose. In the context of an unmanned aerial vehicle simulation, participants monitored six video feeds to detect predefined targets with the assistance of onboard automation. Automation reliability (95% versus 50% reliable) and priming (reliability information provided or not) were manipulated. Eye movement data, subjective trust ratings, and performance data were collected. The eye tracking data show that people visit more frequently and spend more time on low reliable automation. Priming information could also affect the participants' trust level and trigger different types of searching behavior, as reflected in eye tracking data such as mean saccade amplitude. In summary, these findings confirm that eye tracking is a very promising tool for inferring trust and supporting future research into trust calibration.**

***Index Terms*—Eye tracking, trust calibration, trust measurement, visual search.**

I. Introduction

THE introduction of automation technologies to various application domains, such as the military, aviation and health care, has helped improve the safety and efficiency of operations [1]. At the same time, automation has also created challenges. One important issue that has recently received considerable attention is people's trust in highly automated systems. Several studies have shown that trust, i.e., "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [2], is a critical intervening variable affecting automation usage [3]. If people's trust in automation is miscalibrated, i.e., if it does not match the actual capabilities and reliability of the system [2], [4], people are likely to misuse or disuse systems [5]. Misuse happens when people trust an automated system too much and over rely on the automation—a problem that has contributed to incidents and accidents in high-risk domains, such as aviation. Monitoring failures resulting from pilots' overtrust in automation have been reported frequently to the Aviation Safety Reporting System [6], [7]. Disuse happens when people lack trust in an automated system. This can lead to the slow adoption, and even complete rejection of those systems. Fostering safe and appropriate use of modern technologies requires a better understanding of, and support for the trust calibration process.

Trust in human–automation interaction can be affected by many different factors. A recent literature review of human trust in automation [8] divided these factors into three categories: human related, automation related, and environment related. Among automation-related trust factors, system reliability is considered to be most critical for shaping operators' trust in a system [2], [8], [9]. Past studies indicated that operator trust increased as a function of system reliability [10] and that automation failures led to a loss or decrease in trust [10]–[13]. It is important to note that, in many of these studies, system reliability was set as a stable, between-subject variable [10], [14]. Studies that did examine reliability changes over time, as a within-subject variable, most often assessed trust using repeated discrete ratings, across several very short trials [9], [15]. For example, Akash and his colleagues built a model of dynamic human trust in an obstacle detection sensor based on subjective trust ratings that were collected over 100 short trials. More work is needed to develop means of measuring or inferring trust on a continuous basis, over extended periods of time, to examine how it coevolves with changes in system reliability at the level of individuals.

In addition to system reliability, priming—a human-related factor—can influence trust formation. Priming has been defined as "the exposure to a stimulus that influences the response to another subsequent stimulus and thus influences behavior later on, without the individual necessarily being aware of this effect" [16]. For example, Muir [4] suggested that one way to facilitate trust calibration is to "improve the perception of trustworthiness" of an automated system, such as a decision aid, by providing operators with automation reliability/predictability information in advance of operations. This information can support the effective management of limited attentional resources across tasks and systems. Some studies have shown that priming indeed improves the trust formation process but does so only early on

Manuscript received May 25, 2018; revised March 14, 2019; accepted June 30, 2019. Date of publication August 20, 2019; date of current version November 21, 2019. This work was supported by the Rackham Graduate School, the University of Michigan, under Grant U057375. This article was recommended by Associate Editor K. Feigh. *(Corresponding author: Yidu Lu.)*

The authors are with the Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48019 USA (e-mail: luyd@umich.edu; sarter@umich.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2930980

[17]–[21]. More research is needed to establish long-term effects of priming on trust, and how priming (a top-down process [22]) may interact with, and possibly be overridden by, observations of actual system performance (a bottom-up process) during human interaction with automation.

To date, several types of trust measures have been employed to examine trust in human–automation interaction, such as subjective ratings, behavioral data, and psychophysiological measurements. Most studies have relied heavily on rating scales [3] which have the benefit of being easy to implement and employ. However, one shortcoming of this approach is that these ratings tend to be collected only once, often at the end of an experiment [20], [23], [24]. Used that way, trust ratings fail to capture how trust evolves over extended periods of human–machine interaction. Some studies have tried to measure trust changes over time by repeatedly asking for subjective ratings at short intervals throughout the experiment [25]–[28]. This technique can capture the temporal evolution of trust to some extent but is quite disruptive of task performance and thus undesirable and not feasible in real-world environments. In addition, most of these studies involved very short trials, lasting only several seconds. Another approach to inferring trust levels is based on observable operator/participant behavior. This approach assumes that, while trust is considered an attitude, this attitude forms the basis for intentions [2] and, ultimately, mediates behavior [11], [29], [30]. For example, a driver switching back to manual control can be an indicator of a loss or reduction in trust in the autonomous vehicle [31]. Like subjective ratings, behavioral measures tend to be discrete which makes it difficult to trace moment-to-moment changes in trust. Furthermore, behavioral indicators tend to be domain specific, thus making it difficult to compare findings across studies [32]. Finally, attempts have been made to infer trust in real time based on various psychophysiological measures. Initially, these measures were employed to study interpersonal trust [33], [34]. More recently, research has shown that psychophysiological measures such as electroencephalography and galvanic skin response can be used to build models of human trust in modern technologies [35], [36].

Another promising means of measuring trust in a continuous, real-time, and nonintrusive way is eye tracking [37]–[39]. Raw eye movement data can be used to calculate a variety of eye tracking metrics which have been used successfully to infer various challenges to perception and cognition in past studies, such as visual clutter [40] and attentional narrowing [41]. To date, there is very limited empirical evidence for the feasibility and validity of using eye tracking to infer human trust in automation. Hergeth and his colleagues [42] collected eye tracking data to quantify drivers' trust in a highly automated car. In their study, gaze behavior was measured in terms of the frequency of monitoring (defined as the number of fixations) of the driving scene during a non-driving-related task (NDRT) and in terms of the monitoring ratio (defined as the sum of fixation durations during a NDRT, scaled to the duration of that NDRT). Subjective trust ratings were collected also and related to the eye tracking data. The findings supported the assumption of a negative correlation between subjective trust and monitoring frequency. In other words, participants monitored automation-related displays more often when the system was trusted less, and vice versa. A small number of studies confirmed this correlation between people's trust and their visual scanning behavior [43]–[45]. However, these studies explored the diagnosticity of only a small number of eye tracking metrics.

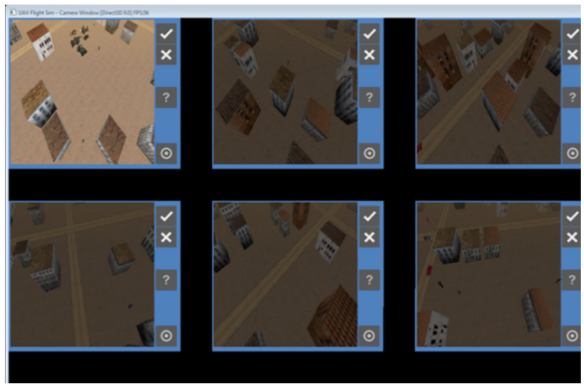


Fig. 1. UAV simulation.

The aim of this study is to identify the most useful eye tracking metrics for tracing, in real time, variations in trust as a function of system reliability and priming. This ability is a prerequisite for designing adaptive technologies that detect and overcome trust miscalibration and avoid unexpected performance breakdowns of joint human–machine systems.

II. Methods

A. Participants

Thirty-five University of Michigan undergraduate and graduate students participated in the experiment. Data from three participants were excluded due to malfunctions of the eye tracker or incomplete data. The 32 participants whose data were included in the data analysis were between the ages of 18–35 years ($M = 24.50$, SD $= 3.86$). Eighteen participants were males. None of the participants had any experience with unmanned aerial vehicle (UAV) control before this study. Because the eye tracker used in this experiment cannot be worn with additional eyewear (such as prescription glasses), all participants were required to have normal or corrected-to-normal vision (contact lenses were allowed). This study was approved by the University of Michigan Institutional Review Board (IRB Reference ID: HUM00126745).

B. Apparatus and Tasks

The application domain for this study was military reconnaissance and intelligence gathering with the assistance of UAVs. A UAV simulation replicating military target identification tasks was developed in our laboratory, as shown in Fig. 1.

During a 30-min scenario, automation onboard six UAVs scanned predefined regions to help with the detection of a military target (a truck carrying a gun; see Fig. 2). The simulated video feeds from the six UAVs were displayed in a 2*3 grid on a single 27 in monitor display. They included both actual targets and similar looking nontargets, as shown in Fig. 3.

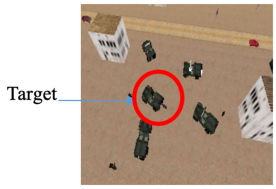


Fig. 2. Target example.

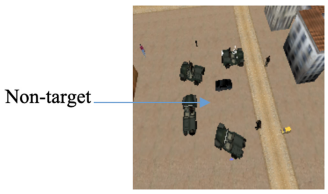


Fig. 3. Nontarget example.

TABLE I
OVERALL RELIABILITY AND CORRESPONDING NUMBERS OF HITS, CORRECT REJECTIONS, FALSE ALARMS AND MISSES

Reliability = 95%		Reliability = 50%	
Hits	35	Hits	30
Correct rejections	51	Correct rejections	15
False alarms	4	False alarms	9
Misses	0	Misses	36

Hit: UAV successfully detects a target and the screen is highlighted. Correct rejection: UAV correctly determines that an object is not a target and the screen is not highlighted. False alarm: UAV incorrectly identifies an object as a target and the screen is highlighted. Miss: UAV fails to detect a target and the screen is not highlighted.

A UAV's video feed was highlighted when it identified a possible target. The participant then reviewed the scene and pressed one of two buttons to either confirm ($\surd$) or reject (X) the presence of a target. In addition, participants were asked to scan the various UAV video feeds on a continuing basis to make sure no targets were missed. If a target was missed by a UAV but detected by a participant, she/he pressed a third "target" ($\odot$) button to record the event. Participants were instructed to press the "?" button if they were uncertain about the presence of a target, independent of whether the screen was highlighted or not.

C. Experiment Design

The experiment employed a 2 (reliability: high, low) * 2 (priming: reliability information, no reliability information) factorial design. Automation reliability was a within-subject factor. Half of the UAVs were highly reliable (95% correct) while the other three UAVs were only 50% reliable. For half of the participants, the upper three UAVs were the highly reliable ones whereas for the other half of the participants, the three highly reliable UAVs were shown at the bottom of the screen. Detailed information about the performance of the high- and low-reliability UAVs is shown in Table I. The total number of hits and false alarms was the same for the two levels of automation reliability to avoid biasing participants' attention allocation.

Priming was a between-subject factor. Half of the participants were informed about the overall reliability of the two groups of UAVs in advance of the experiment while the other half did not receive any reliability information. All participants assessed system reliability based on their observations of UAV performance throughout the experiment. They were not provided with feedback or scores to help them determine the reliability of the automation.

The dependent measures in this experiment included eye tracking data, performance on the search task (including response times and error rates), and subjective trust ratings. Throughout the experiment, participants were prompted every 2 minutes to rate their trust in the upper three UAVs and the lower three UAVs on a scale of 0 ("I do not trust these UAVs at all") to 9 ("I completely trust these UAVs"). Once participants completed the two ratings, the monitoring task automatically resumed. Eye movement data were collected using Tobii Pro Glasses 2 and the Tobii Pro Lab software. The sampling rate of the eye tracking glasses was 50 Hz.

The raw eye tracking data were used to calculate the metrics listed in Table II. These metrics fall into three commonly used categories [28], [33]: 1) temporal metrics; 2) spatial metrics; and 3) count metrics. The two temporal metrics are total and average fixation duration. A fixation is defined as "a relatively stable eye-in-head position within some threshold of dispersion (typically $\sim$2°) over some minimum duration (typically 100–200 ms), and with a velocity below some threshold (typically 15–100°/s)" [47]. The fixation filter in the Tobii Pro Lab Analyzer is based on the velocity of the directional shifts of the eye. The default threshold is 100°/s. The four spatial metrics—mean saccade amplitude, backtrack rate, rate of transitions, and scanpath length per second—relate to the efficiency and randomness of the scanpath. Finally, the two count metrics capture the number/frequency of fixations and transitions between areas of interest (AOI). An AOI is defined by the experimenter as the targeted area for which eye movement data are being analyzed. In this experiment, two sets of AOIs were used: 1) for the purpose of studying the effects of automation reliability and the relationship between eye tracking and subjective trust ratings, the three upper and lower UAV windows, respectively, were considered an AOI; and 2) to examine participants' scan patterns in more detail, each of the six UAV windows was defined as a separate AOI.

D. Experiment Procedure

Each experiment session started with participants being informed that the goal of the experiment was to study trust in human–automation interaction. Participants then read and signed the consent form. The eye tracker was calibrated, and participants were trained on the target identification task for 10 minutes. At the end of the training, participants were expected to correctly identify at least 90% of all targets. If they did not meet this criterion, they received additional training until their performance was acceptable. Participants in the priming group were informed about the overall reliability of the two groups of UAVs (95% versus 50%) in advance of the experiment. However,

TABLE II
EYE TRACKING METRICS CALCULATED IN THIS STUDY

Metric	Definition	Prediction
	TEMPORAL METRICS	
Total fixation duration(s)	The total time each participant fixated each AOI	Low system reliability results in low trust which, in turn, is expected to lead to longer fixation duration for affected UAVs (both average and total). Priming will, at least initially, result in a larger difference in fixation duration between low and high reliability automation.
Average fixation duration (s)	The average duration of the fixations within each AOI	
	SPATIAL METRICS	
Mean saccade amplitude (pixel)	The average amplitude of all saccades	Low trust levels resulting from low system reliability will lead to less efficient and organized search behavior. The no-priming group needs time to build up appropriate trust levels. Therefore, their search behavior will initially be more random and less efficient (longer mean saccade amplitude, larger backtrack rate, larger transition rate between AOIs and longer scanpath length per second).
Backtrack rate (/s)	The number of saccade angles larger than 90 degrees, divided by the total time	
Rate of transitions (/s)	The number of transitions between AOIs, divided by the total time	
Scanpath length per second (pixel/s)	The total length of the scanpath, divided by the total time	
	COUNT METRICS	
Total fixation count	The number of fixations within each AOI	Participants are likely to monitor low reliability UAVs more often; therefore, the total fixation count is expected to be larger for those vehicles. Participants are also likely to switch more often between the three low-reliability UAV windows to detect misses, resulting in a higher transition count. Priming information will further differentiate participants' trust in low and high reliability UAVs, leading to more efficient search behavior (smaller total fixation counts and transition counts).
Transition count	The number of transitions between AOIs	

they were not told about the distribution of particular types of errors (false alarms versus misses), nor were they informed that there would be no misses in the high-reliability condition. Participants in the no priming group did not receive any information about overall system reliability. Following the 30-min experiment session, a debriefing was conducted to ask participants for feedback about various aspects of the experiment, such as their overall monitoring strategy and the effectiveness of the automatic highlighting of UAV windows.

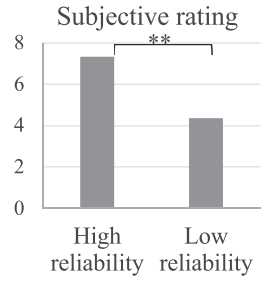


Fig. 4. Ratings as a function of reliability (**: $p < 0.01$).

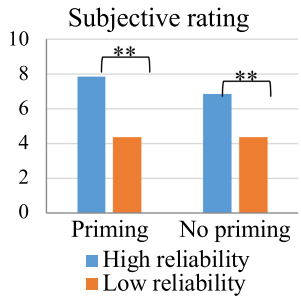


Fig. 5. Ratings as a function of priming (*: $p < 0.05$).

III. RESULTS

A. Subjective Trust Ratings

Subjective trust ratings were analyzed using a 2 (reliability: high versus low) *2 (priming: reliability information provided versus not) linear mixed model. The significance level was set at 0.05. Participants rated the highly reliable UAVs ($M = 7.34$, SD $= 1.06$) as significantly more trustworthy than the low-reliability UAVs ($M = 4.37$, SD $= 1.24$, F (1,30) $= 112.71$, $p < 0.001$), as shown in Fig. 4. There was no significant effect of priming, nor was there an interaction between reliability and priming (see Fig. 5).

B. Eye Tracking Metrics

Eye tracking metrics were analyzed with a 2 (reliability: high versus low) *2 (priming: reliability information provided versus not) linear mixed model. The participant number was entered as a random effect. The significance level was set at 0.05.

The analysis revealed a significant main effect of automation reliability on only one of the two temporal eye tracking metric (see Table III). Longer total fixation durations were observed for low-reliability UAVs ($M = 817.75$, SD $= 168.65$), compared to high-reliability UAVs ($M = 650.06$, SD $= 160.82$; $F(1, 30) = 18.12$, $p < 0.001$). Except for mean saccade amplitude, all spatial metrics were significantly affected by system reliability. Participants' backtrack rate was significantly higher in the

TABLE III
EFFECTS OF SYSTEM RELIABILITY ON EYE TRACKING METRICS

Metric	High reliability	Low reliability	Main effects of reliability
TEMPORAL METRICS			
Total fixation duration(s)	650.06(160.82)	817.65(168.71)	F(1,30)=18.12, p<0.001
Average fixation duration (s)	0.30(0.06)	0.31(0.06)	Not significant
SPATIAL METRICS			
Mean saccade amplitude (pixel)	31.92(9.67)	32.09(11.20)	Not significant
Backtrack rate (/s)	0.05(0.04)	0.09(0.07)	F(1,30)=12.16, p=0.001
Rate of transitions (/s)	0.37(0.16)	0.46(0.17)	F(1,30)=5.06, p=0.028
Scanpath length per second (pixel/s)	522.07(221.79)	646.20(248.03)	F(1,30)=5.24, p=0.029
COUNT METRICS			
Total fixation count	2277.63(669.72)	2679.75(469.49)	F(1,30)=8.78, p=0.004
Transition count	781.91(330.17)	975.12(374.36)	F(1,30)=5.21, p=0.030

TABLE IV
EFFECTS OF PRIMING ON EYE TRACKING METRICS

Metric	Priming	No priming	Main effects of priming
TEMPORAL METRICS			
Total fixation duration(s)	744.32(225.41)	723.49(133.36)	Not significant
Average fixation duration (s)	0.33(0.08)	0.27(0.04)	F(1,30)=8.45, p=0.007
SPATIAL METRICS			
Mean saccade amplitude (pixel)	27.89(10.25)	36.12(8.88)	F(1,30)=6.56, p=0.016
Backtrack rate (/s)	0.07(0.05)	0.07(0.05)	Not significant
Rate of transitions (/s)	0.36(0.16)	0.47(0.16)	F(1,30)=7.24, p=0.009
Scanpath length per second (pixel/s)	518.36(244.38)	649.91(223.63)	F(1,30)=4.89, p=0.035
COUNT METRICS			
Total fixation count	2293.41(639.75)	2663.97(521.78)	F(1,30)=7.46, p=0.008
Transition count	799.44(383.83)	957.59(328.91)	Not significant

low-reliability condition ($M = 0.09$, SD $= 0.07$), compared to the high-reliability condition ($M = 0.05$, SD $= 0.04$, $F(1,30) = 12.16$, $p = 0.001$). The rate of transitions for highly reliable UAVs ($M = 0.37$, SD $= 0.16$) was significantly smaller than for low-reliability UAVs ($M = 0.46$, SD $= 0.17$, $F(1,30) = 5.06$, $p = 0.028$). And finally, participants' scanpath length per second was significantly shorter for highly reliable UAVs, compared to low-reliability UAVs ($F(1,30) = 5.24$, $p = 0.029$). Also, both count metrics changed significantly as a function of reliability: fixation counts were significantly higher for low-reliability UAVs ($M = 2277.63$, SD $= 669.72$), compared to high-reliability UAVs ($M = 2679.75$, SD $= 469.49$; $F(1,30) = 8.78$, $p = 0.004$), and transition counts were significantly larger for low-reliability UAVs ($M = 975.12$, SD $= 374.36$) than for highly reliable vehicles ($M = 781.91$, SD $= 330.17$, $F(1,30) = 5.21$, $p = 0.03$).

A significant main effect of priming was found for one of the two temporal metrics (see Table IV). The average fixation duration was longer for the priming group ($M = 0.33$, SD $= 0.08$) than for the no priming group ($M = 0.27$, SD $= 0.03$, $F(1, 30) = 8.45$, $p = 0.007$). A significant effect of priming was observed also for three spatial metrics. Mean saccade amplitude was significantly shorter for the priming condition ($M = 27.89$, SD $= 10.25$), compared to the no priming condition ($M = 36.12$, SD $= 8.88$, $F(1,30) = 6.56$, $p = 0.016$). When primed with reliability information before the experiment, participants' rate of transitions ($M = 0.36$, SD $= 0.16$) was significantly smaller compared with participants who had no information about UAV reliability ($M = 0.47$, SD $= 0.16$, $F(1,30) = 7.24$, $p = 0.009$). Finally, the scanpath length per second was significantly longer in the no priming group ($F(1,30) = 4.89$, $p = 0.035$). For count metrics, the priming group showed a smaller fixation count ($M = 2293.41$, SD $= 639.752$), compared with the no priming group ($M = 2663.97$, SD $= 521.78$, $F(1,30) = 7.46$, $p = 0.008$); however, the second count metric—transition counts—did not change as a function of priming.

The analysis also revealed a significant interaction between automation reliability and priming for total fixation duration ($F(1,60) = 7.50$, $p = 0.008$). A simple effect analysis showed that total fixation duration ($F(1,60) = 24.47$), $p < 0.001$) differed significantly as a function of automation reliability for the priming group but not for the no priming group (see Fig. 6).

1) Relationship Between Eye Tracking Metrics and Subjective Trust Ratings: To validate the eye tracking metrics, we calculated their correlations with the subjective trust ratings, both for all participants combined and separately for the priming and no priming groups (see Table V). The eye tracking metrics showed a significant negative correlation with subjective trust ratings, with two exceptions. There was no significant correlation between mean saccade amplitude and subjective trust ratings, and average fixation duration was not correlated with subjective ratings in the no priming group.

2) Changes in Eye Tracking Metrics as a Function of System Reliability: The overall reliability for the two groups of UAVs was 95% (group 1, high reliability) versus 50% (group 2, low

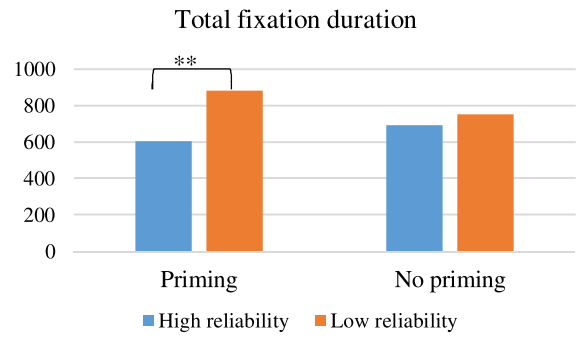


Fig. 6. Priming effect on attention allocation for high and low automation reliability (**: $p < 0.01$).

TABLE V
CORRELATIONS BETWEEN EYE TRACKING METRICS AND SUBJECTIVE RATINGS

	Subjective trust ratings (Overall)	Subjective trust ratings (Priming)	Subjective trust ratings (No priming)
Total fixation duration (s)	-0.918**	-0.928**	-0.799**
Average fixation duration (s)	-0.433*	-0.487**	-0.188
Mean saccade amplitude (pixel)	0.058	0.012	0.035
Backtrack rate (/s)	-0.858**	-0.839**	-0.687**
Rate of transitions (/s)	-0.821**	-0.852**	-0.565**
Scanpath per second (pixel/s)	-0.793**	-0.837**	-0.621**
Total fixation count	-0.796**	-0.849**	-0.556**
Transition count	-0.783**	-0.829**	-0.491**

*: $p < 0.05$, **: $p < 0.01$.

TABLE VI
ACTUAL RELIABILITY SETTINGS FOR HIGH (GROUP 1) AND LOW (GROUP 2) RELIABILITY AUTOMATION

	Interval1	Interval2	Interval3	Interval4	Interval5
Group 1	1	1	1	1	1
Group 2	0.5	0.4	0.375	0.5	0.625
	Interval6	Interval7	Interval8	Interval9	Interval10
Group 1	0.88	0.67	1	1	0.8
Group 2	0.67	0.4	0.75	0.286	0.6
	Interval11	Interval12	Interval13	Interval14	Interval15
Group 1	1	1	0.857	1	1
Group 2	0.57	0.625	0	0.4	0.56

reliability), respectively. However, it varied slightly for each vehicle throughout the experiment (every 2 minutes; range: 0–1) (see Table VI). A correlation analysis was conducted on the percentage differences for the various eye tracking metrics between the high- and low-reliability UAVs to determine whether these reliability variations were reflected in short-term changes in attention allocation. Only total fixation count was found to

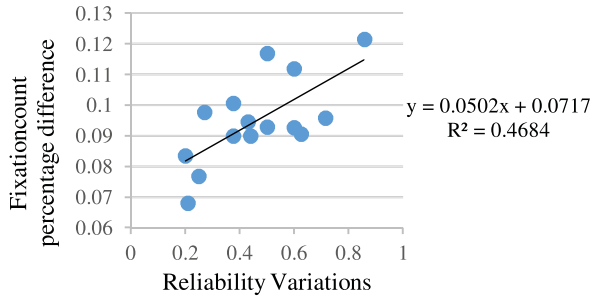


Fig. 7. Correlation between fixation count and actual reliability variations.

be correlated with reliability changes over time ($r = 0.685$, $p = 0.005$), as shown in Fig. 7.

C. Performance

1) Reaction Time: Response time was defined as the time between the first appearance of the target/nontarget on the screen and the participant's button press (confirm, reject, uncertain, miss). A 2 (reliability: high versus low) *2 (priming: reliability information provided versus not) mixed linear model showed a significant main effect for UAV reliability. Participants' response time for low-reliability UAVs ($M = 2.14$s, SD $= 0.17$) was significantly longer than for high-reliability UAVs ($M = 1.88$s, SD $= 0.18$; $F(1,30) = 205.07$, $p < 0.001$). Response time did not differ significantly between the priming and the no priming groups, and no interaction was observed between priming and reliability.

2) Overall Error Rate: The overall error rate was defined as the total number of errors, divided by the total number of trials in that condition. A 2 (reliability: high versus low) *2 (priming: reliability information provided versus not) mixed linear model was performed on these data. It reveals a significant main effect of reliability on error rate. Participants made more errors when interacting with the low-reliability UAVs ($M = 0.090$, SD $=$ 0.005), compared to the highly reliable vehicles ($M = 0.027$, SD $= 0.005$; $F(1,30) = 73.60$, $p < 0.001$). No significant effect of priming and no interaction effect were found.

3) Specific Error Rate: There were five possible types of errors that participants could make during the experiment, as listed in Table VII. A 2 (reliability)*2 (priming) mixed linear model was conducted on participants' error rates for each of these categories. Results indicated a significant main effect of reliability on error types 3 (missing a target that was missed by the UAV), 4 (false alarms) and 5 (failure to respond when the UAV window was highlighted). Participants showed a significantly higher type 3 error rate ($M = 0.125$, SD $= 0.072$) for low-reliability UAVs, compared to high-reliability UAVs where no such errors were observed. The type 4 error rate was also significantly higher ($M = 0.135$, SD $= 0.055$) when participants interacted with low-reliability UAVs, compared to the highly reliable UAVs ($M = 0.015$, SD $= 0.033$). Similar to type 3 errors, a significantly higher error rate for type 5 errors was observed ($M = 0.006$, SD $= 0.013$) with highly reliable UAVs, compared to the low-reliability conditions where no such errors occurred.

TABLE VII
ERROR DEFINITION AND EFFECT OF RELIABILITY

Type	Definition	Error rate_high reliability	Error rate_low reliability	Significance on reliability
Error 1	When the sub-screen was highlighted and there was a target, the participant clicked "reject" button.	M=0.031, SD=0.027	M=0.032, SD=0.036	Not significant
Error 2	When the sub-screen was highlighted and there was not a target, the participant clicked "confirm" button.	M=0.094, SD=0.177	M=0.070, SD=0.100	Not significant
Error 3	When the sub-screen was not highlighted and there was a target, the participant didn't click "miss" button.	M=0, SD=0	M=0.125, SD=0.072	F(1,30)=94.92, p<0.001
Error 4	When the sub-screen was not highlighted and there was not a target, the participant clicked "miss" button.	M=0.015, SD=0.033	M=0.135, SD=0.055	F(1,30)=147.71 p<0.001
Error 5	When the sub-screen was highlighted, the participant didn't click any button.	M=0.006, SD=0.013	M=0, SD=0	F(1,30)=6.39, p=0.017

No significant effect of priming and no interaction effect were found.

IV. DISCUSSION

The purpose of this study was to develop and validate an eye tracking-based method for inferring and tracing, in real time, changes in operator trust levels as a function of automation reliability and priming. Eye tracking metrics that fall into three different categories were calculated from raw eye movement data: temporal metrics, spatial metrics, and count metrics. To validate these eye tracking metrics, they were correlated with subjective trust ratings, the traditional means of assessing trust levels and variations. Observed differences in subjective trust ratings suggest that the reliability manipulation in this study was successful. Participants' performance on a UAV control task was recorded to determine whether and how it was affected by changes in attention allocation resulting from different levels of trust.

A. Effects of System Reliability on Eye Movements and Monitoring

The analysis of the eye tracking data revealed that low automation reliability was associated with longer total fixation durations, higher backtrack and transition rates, an increased scanpath length per second, and higher total fixation and transition counts.

Longer fixation durations were expected for low system reliability as participants would trust the automation less and therefore examine potential targets more carefully. This effect was observed for total, but not for average fixation duration. One possible explanation for this finding is that while participants visited low-reliability UAV windows more often (resulting in longer total fixation durations), they were able to examine potential targets equally quickly for both high and low-reliability UAVs (translating into comparable average fixation durations) due to their comparable low level of complexity which has been shown to affect temporal eye tracking metrics [38].

As predicted, most of the spatial metrics were significantly affected by system reliability, except for mean saccade amplitude. Backtrack rate, transition rate, and scanpath length per second can be considered indicators of the efficiency of visual search and scanning which, as predicted, suffered as a result of low automation reliability [46]. This may be explained by the high attentional load imposed and by a high level of uncertainty of where to look, leading to less systematic monitoring behavior. The fact that mean saccade length was not affected by reliability may be explained by earlier findings showing that this metric is particularly sensitive to mental effort [48]. All targets in this experiment for both high- and low-reliability UAVs were identical, and thus likely required the same amount of effort.

Finally, system reliability significantly affected both count metrics. This result confirms our expectations and is consistent with earlier research findings [49], [50]. When people trust automation to perform a task reliably, they will monitor the system less frequently (as expressed by fixation counts). The number of transitions between the low-reliability UAV windows was also higher, most likely because participants needed to scan these windows more frequently to make sure the UAVs did not miss any targets.

The above results indicate that it is useful to calculate and examine multiple eye tracking metrics as various aspects (temporal, spatial and count) of eye movements respond differently to variations in system reliability.

B. Effects of Priming on Eye Movements and Monitoring

Even though priming did not affect subjective trust ratings, it was associated with changes in participants' monitoring behavior. Specifically, priming seemed to contribute to smaller mean saccade amplitudes, transition rates, scanpath length per second, and total fixation counts, as well as longer average fixation durations.

Average fixation duration was the only temporal metric affected by priming. Receiving reliability information in advance likely resulted in participants allocating more effort and attention to the low-reliability UAVs, in a top-down fashion. In contrast, total fixation duration did not differ between the priming and the no priming group. One explanation could be that participants' attention allocation across the entire screen was decided mainly by the total task duration instead of other factors. Priming significantly affected all spatial metrics, except the backtrack rate. This confirms that participants' visual search and scan were indeed less efficient and systematic if participants were

not informed about system reliability. It is not clear why the backtrack rate was not affected by priming when it showed the expected change due to differences in UAV reliability.

Among the count metrics, significantly fewer fixation counts, but not fewer transition counts, were observed in the priming condition. It is possible that any difference in transition counts was masked when considered for the entire screen (including both high- and low-reliability UAVs).

C. Relationship Between Subjective Trust Ratings and Eye Tracking Metrics

For the most part, the observed differences in monitoring behavior for the high- and low-reliability UAVs aligned with participants' subjective trust ratings. The only exception was average fixation duration and mean saccade amplitude, which did not show a significant correlation with those ratings. These metrics were also not affected by system reliability but did change as a function of priming. This suggests that, overall, system reliability has a more pronounced effect on trust.

Another important difference between the eye tracking metrics and subjective trust ratings was that subjective ratings differed significantly between high- and low-reliability UAVs, independent of whether participants were informed about system reliability at the beginning of the experiment. In contrast, one eye tracking metric, total fixation duration, differed between the two reliability levels only in the priming condition. Participants in the no priming group explained that, even though they had noticed differences in automation reliability during the experiment, they still monitored all UAVs to the same extent because they were not sure whether observed reliability levels would remain constant.

A third difference between subjective ratings and the eye tracking metrics was the higher temporal resolution of total fixation counts which closely mirrored the changes in UAV reliability every 2 minutes. In other words, monitoring behavior changed even though participants' attitude towards the vehicle appeared unaffected. This finding is similar to the results from an earlier experiment on autonomous driving [51], where drivers intervened even when they expressed verbally that they expected the automation to be able to handle the emergency situation successfully.

The results suggest that automation reliability is more strongly associated with trust and monitoring behavior than priming. At the same time, the observed dissociation between subjective trust ratings and eye movements calls for more research to determine the nature and strength of causal relationships between system reliability, priming, trust, and eye movements.

D. Effects of System Reliability and Priming on Performance

In terms of performance, participants' reaction time was longer, and their overall error rate was higher for low-reliability UAVs. The high type 3 error rate (failure to notice a missed target when the window was not highlighted) for low-reliability UAVs may be explained by the lack of attention guidance in the form of highlighting. The high type 4 error rate (false alarms when the window was not highlighted) for low-reliability UAVs may be due to participants' response bias. They expected these vehicles to miss more targets and were therefore more willing to call an ambiguous object (due to the lack of highlighting) a target. And the increased type 5 error rate (failure to respond to highlighting of window) for highly reliable UAVs may be the result of participants focusing so much on the low-reliability vehicles that they did not react in time or totally missed in their peripheral vision to the highlighting of the high-reliability UAV windows. Priming did not affect participants' performance significantly which may be explained by a ceiling effect (above 90% accuracy rate).

V. Conclusion and Future Work

The findings from this experiment suggest that eye tracking is indeed an effective technique for inferring changes in operator trust levels in real time. It complements other trust measures, such as subjective ratings or behavioral data, by providing a continuous real-time trace of the temporal evolution of trust. Compared to other psychophysiological measures, eye tracking has the benefits of easier implementation, less intrusion, and a more fine-grained analysis of monitoring behavior. Given the dissociation between some trust-related measures in this study, eye tracking is ideally combined with other techniques to assess and study various facets of trust.

It is important to note two limitations of the present study. First, false alarms and misses were evenly distributed among UAVs of equal reliability and over each short time interval. Past studies have shown that these two failure types can have different effects on one's trust in automation [52]. Automation that mostly triggers false alarms may more strongly affect eye tracking metrics that relate to information processing; in contrast, automation that is prone to misses may lead to changes in eye tracking metrics that reflect search behavior. Second, in this experiment, high and low automation reliability were coupled, which means that, if a participant looked more at one group of UAVs, his/her monitoring of the other group was necessarily reduced, independent of whether or not his/her trust in those vehicles was different.

Additional studies are needed to address the above shortcomings and help achieve the ultimate goal of this research, namely to develop a real-time, eye tracking-based trust measurement that can be used to design adaptive interfaces and support better trust calibration in human–machine teams.

Acknowledgment

The preparation of this manuscript was supported by a UM Rackham Graduate Student Research. The authors would like to thank K. Lieberman, K. Zhao, and A. Mannari for their help with creating the UAV simulation.

References

[1] M. R. Endsley, "From here to autonomy," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 59, no. 1, pp. 5–27, 2016.

[2] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[3] K. A. Hoff and M. Bashir, "Trust in automation," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 3, pp. 407–434, 2014.

[4] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, nos. 5/6, pp. 527–539, 1987.

[5] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 39, no. 2, pp. 230–253, 1997.

[6] J. E. Bahner, A.-D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Human-Comput. Stud.*, vol. 66, no. 9, pp. 688–699, 2008.

[7] K. L. Mosier, L. J. Skitka, and K. J. Korte, "Cognitive and social psychological issues in flight crew/automation interaction," in *Human Performance in Automated Systems: Current Research and Trends*. Mahwah, NJ, USA: Erlbaum, 1994, pp. 191–197.

[8] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 58, no. 3, pp. 377–400, 2016.

[9] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[10] N. R. Bailey and M. W. Scerbo, "Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust," *Theor. Issues Ergonom. Sci.*, vol. 8, no. 4, pp. 321–348, 2007.

[11] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, 2003.

[12] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, "Automation failures on tasks easily performed by operators undermine trust in automated aids," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 48, no. 2, pp. 241–256, 2006.

[13] J. C. Walliser, E. J. D. Visser, and T. H. Shaw, "Application of a system-wide trust strategy when supervising multiple autonomous agents," *Proc. Human Factors Ergonom. Soc. Ann. Meeting*, vol. 60, no. 1, pp. 133–137, 2016.

[14] M. T. Dzindolet, L. G. Pierce, H. P. Beck, L. A. Dawe, and B. W. Anderson, "Predicting misuse and disuse of combat identification systems," *Mil. Psychol.*, vol. 13, no. 3, pp. 147–164, 2001.

[15] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *Proc. Amer. Control Conf.*, 2017, pp. 1542–1548.

[16] E. Tulving and D. Schacter, "Priming and human memory systems," *Science*, vol. 247, no. 4940, pp. 301–306, 1990.

[17] G. Abe and J. Richardson, "Alarm timing, trust and driver expectation for forward collision warning systems," *Appl. Ergonom.*, vol. 37, no. 5, pp. 577–586, 2006.

[18] N. Ezer, A. D. Fisk, and W. A. Rogers, "Reliance on automation as a function of expectation of reliability, cost of verification, and age," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 51, no. 1, pp. 6–10, 2007.

[19] F. C. Lacson, D. A. Wiegmann, and P. Madhavan, "Effects of attribute and goal framing on automation reliance and compliance," *Proc. Human Factors Ergonom. Soc. Ann. Meeting*, vol. 49, no. 3, pp. 482–486, 2005.

[20] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 4, pp. 545–556, 2014.

[21] M. Körber, E. Baseler, and K. Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Appl. Ergonom.*, vol. 66, pp. 18–31, 2018.

[22] J. Theeuwes, "Top-down and bottom-up control of visual selection," *Acta Psychol.*, vol. 135, no. 2, pp. 77–99, 2010.

[23] N. Bagheri and G. A. Jamieson, "Considering subjective trust and monitoring behavior in assessing automation-induced 'complacency'," in *Human Performance, Situation Awareness, and Automation: Current Research and Trends*. London, U.K.: Psychology Press, 2004, pp. 54–59.

[24] E. A. Bustamante, "A reexamination of the mediating effect of trust among alarm systems characteristics and human compliance and reliance," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 53, no. 4, pp. 249–253, 2009.

[25] A. S. Clare, M. L. Cummings, and N. P. Repenning, "Influencing trust for human–automation collaborative scheduling of multiple unmanned vehicles," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 7, pp. 1208–1218, 2015.

[26] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. 8th ACM/IEEE Int. Conf. Human-Robot Interact.*, 2013, pp. 251–258.

[27] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems," in *Proc. 21st Int. Conf. Intell. User Interfaces*, 2016, pp. 164–168.

[28] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2017, pp. 408–416.

[29] M. T. Dzindolet, L. G. Pierce, H. P. Beck, and L. A. Dawe, "The perceived utility of human and automated aids in a visual detection task," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 44, no. 1, pp. 79–94, 2002.

[30] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[31] C. M. Brown and Y. I. Noy, "Behavioural adaptation to in-vehicle safety measures," in *Traffic & Transport Psychology*. Amsterdam, The Netherlands: Elsevier, 2004, pp. 25–46.

[32] J. Meyer and J. D. Lee, *Trust, Reliance, and Compliance*. Oxford, U.K.: Oxford Handbooks Online, 2013.

[33] C. Boudreau, M. D. Mccubbins, and S. Coulson, "Knowing when to trust others: An ERP study of decision making after receiving information from unknown people," *Soc. Cogn. Affect. Neurosci.*, vol. 4, no. 1, pp. 23–34, 2008.

[34] Y. Long, X. Jiang, and X. Zhou, "To believe or not to believe: Trust choice modulates brain responses in outcome evaluation," *Neuroscience*, vol. 200, pp. 50–58, 2012.

[35] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-time sensing of trust in human-machine interactions," *IFAC-PapersOnLine*, vol. 49, no. 32, pp. 48–53, 2016.

[36] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using EEG and GSR," *ACM Trans. Interactive Intell. Syst.*, vol. 8, no. 4, pp. 1–20, 2018.

[37] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods Instrum. Comput.*, vol. 34, no. 4, pp. 455–470, 2002.

[38] M.-L. Lai *et al.*, "A review of using eye-tracking technology in exploring learning from 2000 to 2012," *Educ. Res. Rev.*, vol. 10, pp. 90–115, 2013.

[39] Z. Sharafi, Z. Soh, and Y.-G. Guéhéneuc, "A systematic literature review on the usage of eye-tracking in software engineering," *Inf. Softw. Technol.*, vol. 67, pp. 79–107, 2015.

[40] N. M. Moacdieh and N. B. Sarter, "Eye tracking metrics: A toolbox for assessing the effects of clutter on attention allocation," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 56, no. 1, pp. 1366–1370, 2012.

[41] J. C. Prinet and N. B. Sarter, "The effects of high stress on attention," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 59, no. 1, pp. 1530–1534, 2015.

[42] S. Hergeth, L. Lorenz, R. Vilimek, and J. F. Krems, "Keep your scanners peeled," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 58, no. 3, pp. 509–519, 2016.

[43] C. Geitner *et al.*, "A link between trust in technology and glance allocation in on-road driving," in *Proc. 9th Int. Driving Symp. Human Factors Driver Assessment Training Veh. Des.*, 2017, pp. 263–269.

[44] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 52, no. 3, pp. 381–410, 2010.

[45] C. Gold, M. Körber, C. Hohenberger, D. Lechner, and K. Bengler, "Trust in automation—Before and after the experience of take-over scenarios in a highly automated vehicle," *Procedia Manuf.*, vol. 3, pp. 3025–3032, 2015.

[46] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: Methods and constructs," *Int. J. Ind. Ergonom.*, vol. 24, no. 6, pp. 631–645, 1999.

[47] R. J. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research," in *The Mind's Eye*. Amsterdam, The Netherlands: Elsevier, 2003, pp. 573–605.

[48] S. Chen, J. Epps, N. Ruiz, and F. Chen, "Eye activity as a measure of human mental effort in HCI," in *Proc. 15th Int. Conf. Intell. User Interfaces*, 2011, pp. 315–318.

[49] N. Bagheri and G. Jamieson, "The impact of context-related reliability on automation failure detection and scanning behaviour," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 1, 2004, pp. 212–217.

[50] N. Moray and T. Inagaki, "Attention and complacency," *Theor. Issues Ergonom. Sci.*, vol. 1, no. 4, pp. 354–365, 2000.

[51] D. Miller *et al.*, "Behavioral measurement of trust in automation," *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 60, no. 1, pp. 1849–1853, 2016.

[52] C. Wickens, S. Dixon, J. Goh, and B. Hammer, *Pilot Dependence on Imperfect Diagnostic Automation in Simulated UAV Flights: An Attentional Visual Scanning Analysis*. Urbana-Champaign, IL, USA: Univ. Illinois, 2005.

The Statistical Saliency Model Can Choose Colors for Items on Maps

Joshua Shive, Sharra Rosichan, Sherika Davis, Christena Wade, James Ellison, and Santos Santoni-Sanchez

***Abstract*—We show how a model of visual salience that was originally developed to explain human visual search performance can suggest display design choices that reduce search time for items. The statistical saliency model proposes that the time to find an item on a visual display depends on the similarity between a target item's features and the statistical distribution of display features. In the present study, observers rated the amount of display clutter on a set of MapQuest maps containing colored pushpins. We identified a group of "high-clutter" maps and a group of "low-clutter" maps. Next, we used the statistical saliency model to choose colors for new pushpins placed on those maps. We show that the model's color assignments depend on the colors the display contains. Map designs produced using this method were tested in a visual search experiment. Search time decreased as a pushpin's predicted salience increased. In addition, choosing low salience colors led to slower search times for items on high-clutter displays than for items on low-clutter displays. The method we describe works with real images and does not require any parameter fitting. This study provides evidence that computational models of visual perception have potential as display design tools.**

***Index Terms*—Display clutter, display design, human-computer interface, human performance modeling, information visualization, visual search/scanning.**

I. Introduction

ELECTRONIC maps make it easy to display several layers of information about a geographical area on a single map. Internet mapping sites, such as Google Maps and MapQuest, allow users to display information about traffic, highway conditions, and local attractions and services. In addition, these sites make it possible for a user to customize their view of a map, switching between solid color map backgrounds that represent an area's terrain and digital satellite views of the area. While mapping technologies allow a user considerable flexibility in what is displayed, displays showing many layers of information may become cluttered [1]. The goal of this study is to examine whether a computational model of visual search can suggest colors for items on cluttered and uncluttered maps that reduce search time for those items.

In the current project, we use the statistical saliency model to choose colors for a new item that will be added to a map, given the current distribution of colors the display contains. We first describe cartographic principles of map design related to color use and review research examining visual salience and its impact on visual search. Next, we identify sets of high-clutter and low-clutter maps using observers' subjective clutter ratings in a preliminary experiment. Then, we show how to construct a statistical saliency model of visual search for a pushpin on a map. We use the statistical saliency model to identify the colors with maximum, median, and minimum predicted salience for each of the identified maps. Finally, we create versions of the maps using the model-chosen colors and conduct a visual search experiment to compare search times for items with minimum predicted salience, median predicted salience, and maximum predicted salience. We hypothesize that search time will decrease as model-predicted salience increases.

A. Map Design and Color

Color coding is useful on general purpose maps, which primarily give information about the spatial location of objects, and thematic maps, which illustrate a particular geographic pattern [2]. The same color can indicate regions that share a feature (e.g., counties with zero inches of rain in a month) or locations that are members of the same category (restaurants, for example). Likewise, variations in a particular aspect of color, such as saturation level, might indicate differences between regions representing ordered categories.

Cartography deals not only with mapmaking but with the perceptual and cognitive factors that affect map reading [3]. For example, the contrast between the colors that represent different categories of information on a thematic map affects a map reader's ability to distinguish between those categories. Online tools, such as ColorBrewer 2.0 (www.colorbrewer2.org), compare the relative color contrasts of families of map items to suggest contrasting color schemes for map displays [4], [5]. These tools typically calculate color contrasts as distances in a particular color space (e.g., the Munsell color space).

B. Visual Search

Before a reader can judge distances between points on a map or interpret the information the map presents, the reader must first find the location or an icon representing the location on the map. Laboratory studies of visual search typically require an

Manuscript received April 10, 2018; revised September 13, 2018 and December 9, 2018; accepted January 20, 2019. Date of publication March 19, 2019; date of current version November 21, 2019. This paper was recommended by Associate Editor M. L. Bolton. *(Corresponding author: Joshua Shive.)*

J. Shive, S. Davis, C. Wade, and S. Santoni-Sanchez are with the Department of Psychology, Tennessee State University, Nashville, TN 37209 USA (e-mail: jshive@tnstate.edu).

S. Rosichan is with the School of Information Sciences, University of Tennessee, Knoxville, TN 37996 USA.

J. Ellison is with the Department of Psychology, Middle Tennessee State University, Murfreesboro, TN 37132 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2901896

observer to search for a target on a display containing distracting items [6]. The target appears on half of the trials. Observers are instructed to make one response if the target is present and another response if the target is absent. Search time and accuracy are measured. Analysis of visual search data often relates set size (i.e., the number of search items) with search time or search accuracy. Search time is measured as the length of time it takes an observer to make a response after the display appears. When a search target shares no features with distracting items, set size does not affect search time. However, when the target is defined by a conjunction of features, search time increases as set size increases [7].

Research investigating visual search has characterized the distribution of search times across a variety of experimental search conditions [8] and identified factors that guide visual search. Guided searches tend to be faster and more accurate than unguided searches [9]. An observer's knowledge of a search target's identity or features can produce top-down guidance [10], whereas differences between a target item's features and the distracting items' features can produce bottom-up guidance. Bottom-up guidance can also arise when the distracting items share similar features with each other. Attributes of the scene, such as the common locations of an object, can also guide search. Additionally, an observer's prior history of searching the scene, as well as the perceived value of items in the search display, can guide search.

Several visual features, such as color, motion, orientation, and size, are known to guide search [11]. However, some features (e.g., object size, color value, and color hue) are more effective than others (e.g., orientation) at decreasing search time for map items and increasing the percentage of changes to maps that observers notice [12].

C. Visual Salience and Display Clutter

The influences of top-down and bottom-up guidance on a display location's visibility are sometimes collectively referred to as the salience of a display location or item. The concept of visual salience dates back at least to [13], which proposes a "saliency map" that the visual system uses to select where to move the eyes. The saliency map represents the activation of cells in the visual system that respond to changes in color, intensity, or orientation across locations in a display. Salience models attempt to predict the conspicuity of objects and locations in the world, given the distribution of display features. Most of these are bottom-up models [14]. That is, they do not incorporate information about a target's identity and use a rule for selecting the next object that should be fixated (but see [15]–[18] for examples of models that do incorporate top-down information).

Cartographers have used salience models to predict patterns of eye movements in an observer's search for items on a map [19]. In general, increasing the salience of a particular item tends to reduce search time for the item, especially when the item is relevant to the task at hand [20]. Novice map observers tend to spend more time looking at perceptually-salient items than they do looking at items that are not salient. However, after training, those observers spend more time looking at thematically-relevant items and less time on the perceptually-salient items [21].

An item's salience may be related to the amount of clutter a display contains [22]. Display clutter is an issue in a variety of domains including medicine [23] and aviation [24]. Studies of clutter on maps [25] distinguish between global clutter, which refers to the total amount of clutter on a display, and local clutter, which is the amount of clutter on a small region of the display. Both types of clutter can increase search times, with the longest search times for displays that are high in both global clutter and local clutter.

Ironically, the more information presented on a map, the slower search performance is [26]. The more layers or types of information the map contains, the more difficult it becomes to ignore irrelevant information [27]. Cartographers have identified methods for reducing the amount of clutter on electronic maps. For example, a designer can limit the items shown to those that match a search criteria or order the items that may be displayed on a map in terms of relative importance [28]. In addition, computational methods that estimate the amount of clutter on a map can identify map locations that are difficult to read [29].

D. Statistical Saliency Model

In the current study, we choose colors for items on map displays using predictions of the statistical saliency model, a computational model of visual search. The statistical saliency model [22], [30] proposes that the time to find an item on a visual display depends on the difference between a target item's features and the distribution of display features. The statistical saliency model represents the distinctiveness of an item's features on a display as the Mahalanobis distance between the item's features and the distribution of display features. The model calculates the salience S of a target item as

$$S = \sqrt{(T - \mu_D)\Sigma_D^{-1}(T - \mu_D)} \tag{1}$$

where T is a vector of the target's features and μ_D and Σ_D are the mean and covariance matrix of the distribution of display feature vectors, respectively.

According to this model, two factors determine the salience of an item on a display. The first is the difference between the target's features and the mean of the display's features, represented by the quantity $(T - \mu_D)$. The greater this distance, the more salient an item will be. Thus, a red item on a green background will be more salient than a light green item on a dark green background.

The second factor that determines an item's salience is the amount of variability in the display's features, represented by the quantity Σ_D. In the model, the covariance matrix of display features is in the denominator of the salience computation, so an item's predicted salience will decrease as the amount of variability in the display's features increases. For example, a colored item surrounded by distracting items of a different uniform color against a uniform background would have a higher predicted salience than the same item surrounded by multicolored distractors against a multicolored background.

In this study, we use the statistical saliency model as a model of global salience. That is, we use it to predict the salience of an item relative to the distribution of all features a particular display contains. However, the model was originally intended

as a measure of local salience, meant to capture the salience of an item relative to its immediate neighbors. This may have an effect on the precise colors the model selects for a given map in our study; however, we would expect that using the model to choose colors locally would generally produce color choices with higher salience than the ones our approach generates. Thus, if salient global color choices produce decreases in search time, locally salient color choices would result in even greater search time benefits.

II. Experiment 1: Identifying Cluttered and Uncluttered Maps

We conducted a clutter rating experiment to identify a group of low-clutter maps and a group of high-clutter maps. We obtained a group of online maps and added colored pushpins to each map. Next, we collected observers' subjective ratings of the amount of clutter on each map. We used the ratings to identify "high-clutter" and "low-clutter" maps that could be used in Experiment 2.

A. Participants

In total 30 Tennessee State University undergraduates who reported normal color vision and corrected to normal vision participated in this study. Participants received course credit for their participation.

B. Stimuli

We collected a group of 150 MapQuest maps showing areas of the United States. Each map measured 865 pixels by 718 pixels. The maps represented an area of the United States where 68 pixels equaled 9600 feet. The maps were chosen by scrolling around the map of the United States at the appropriate resolution and taking screenshots. After the screenshots were taken, the images were cropped to the map boundaries using a MATLAB script that loaded each map and cropped the image to include only the map. An effort was made to avoid including the names of cities in the screenshots. We left the MapQuest legend and interface elements (i.e., the zoom scale, the controls for moving north, south, east, or west on the map, the toggle buttons for Live Traffic and Satellite/Map) on the map, so the map would look as naturalistic as possible.

Next, we added colored pushpins to each map. We chose to mimic the shape and look of MapQuest pushpins that were in use at the time we downloaded the maps. For example, we gave each pushpin a black border. We first added an orange pushpin (RGB: 255,119,106) to the center of each map representing the observer's current location. Next, we wrote an algorithm that added groups of colored pushpins representing categories of map locations, with one color per category. This idea was based on a convention common in web-based maps, in which locations belonging to a category, such as hotels, are indicated using pushpins of the same color.

The number of categories of pushpins on a particular map was selected randomly from a value between 1 and 4. The minimum number of allowed pushpins per category was 1. The maximum was 5. Thus, it was possible for a map to contain as few as one pushpin and as many as 20 pushpins. The location of each pushpin was chosen randomly from a 30-by-74 grid of locations centered on the map, resulting in 2200 possible locations for each pushpin. Because of the large number of possible pushpin locations, we did not randomly jitter the location of each pushpin. In addition, we gave each pushpin a black border.

The color for each pushpin category was selected randomly from the list of 267 colors of the Inter-Society Color Council (ISCC) and National Bureau of Standards (NBS) color naming system [31]. This system provides a consistent method for describing an object's color. The ISCC-NBS system divides color space into 267 color blocks and provides consistent language for color names using modifiers such as "vivid," "strong," and "grayish" to distinguish saturation, and equivalent language for describing differences of value and differences of hue. The color at the approximate center of a color block is called the "centroid color."

C. Procedure

Participants were shown 150 MapQuest maps containing pushpins whose colors were selected from the ISCC-NBS list of standard colors. Stimuli were presented electronically using the E-Prime 2.0 software [32], [33] and displayed at the center of a 17 in Dell LCD monitor with a display resolution of 1280 $\times$ 1024 pixels. We asked participants to sit at a comfortable distance from the screen. At a distance of approximately 21 in, each map occupied approximately 24° by 22° of visual angle. For each map, participants rated subjective map clutter on a scale of 1–7 (1 = totally uncluttered; 7 = totally cluttered). The order of maps was randomized for each participant. Participants were given the opportunity to answer "not sure" if they could not determine the amount of clutter on a map. On average, "not sure" responses were given on 8% of the trials. Since only 36 trials out of 4500 received a "not sure" response, we plan to avoid this option in future studies to eliminate a potential source of noise in participants' responses.

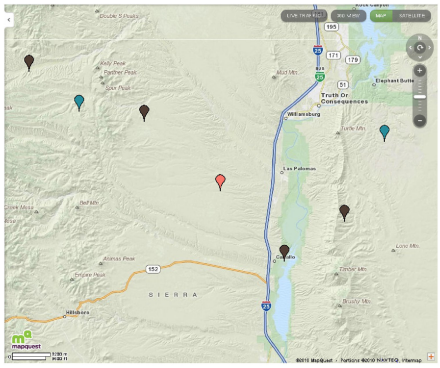
After data collection, we noticed that two of the Experiment 1 maps had appeared on more than one trial in Experiment 1. One of the maps had appeared twice, while another had appeared three times. Each time each map had appeared in Experiment 1, it had contained a different arrangement of pushpins. Because each observer saw each of these maps more than once, we eliminated these maps from further use in this study. The subjective clutter ratings for these maps were not included in the Experiment 1 analysis, and these maps were not used in the subsequent algorithm used to choose salient colors.

D. Results

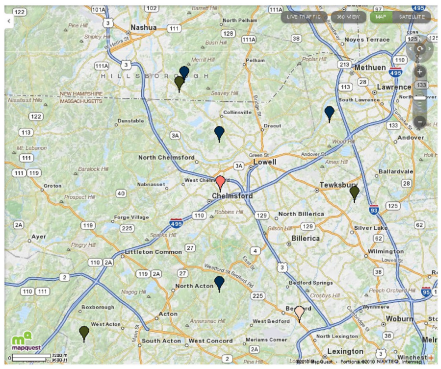
We averaged the clutter ratings for each of the 145 nonrepeated maps across observers. Average clutter ratings for the 145 maps were positively skewed (median = 3.33, skewness = .76) with mean (M) = 3.50, and standard deviation (SD) = 0.93.

III. Choosing Colors for Map Pushpins

The goal of this study is to evaluate the ability of computational models of visual perception to suggest features for map display items to minimize search time. We decided to consider



(a)



(b)

Fig. 1. (a) Example low-clutter map from Experiment 1. (b) Example high-clutter map from Experiment 2.

a very simple design goal: choosing the color of a new pushpin added to a map.

In this section, we first describe how we identified high-clutter and low-clutter maps in the set of maps used in Experiment 1. Next, we describe our implementation of the statistical saliency model and how we used it to choose colors for items on the high-clutter and low-clutter maps. Finally, we examine the model-chosen color assignments to explore the factors that determine model-predicted salience.

A. Identifying High-Clutter and Low-Clutter Maps

We wanted to test the statistical saliency model's ability to choose colors for items to minimize search time. However, the amount of clutter that a display contains affects search time, with displays with large amounts of clutter resulting in lengthier search times for items [25]. Thus, we decided to examine participants' subjective ratings of the Experiment 1 maps to identify maps rated as containing low levels of clutter and maps rated as containing high levels of clutter.

The appendix describes a clustering algorithm we designed to create groups of maps whose average clutter ratings differed as much as possible but whose standard deviations were as similar as possible. We identified a group of 25 "low-clutter" maps (M = 2.58, SD = .60) and a second group of 25 "high-clutter" maps (M = 5.05, SD = .60) using this method. Fig. 1 shows an example low-clutter map and an example high-clutter map this procedure produced. For the low-clutter map shown in Fig. 1(a), the mean clutter rating from Experiment 1 was 2.31. For the high-clutter map shown in Fig. 1(b), the mean clutter rating was 5.75.

B. Statistical Saliency Model

We tested the statistical saliency model's ability to choose unique colors for pushpins on each one of the 25 high-clutter and 25 low-clutter maps. Given a particular map display and a list of candidate colors from the ISCC-NBS color centroid list, we identified the model-predicted minimum-salience, median-salience, and maximum-salience pushpins using a brute-force algorithm.

We created an implementation of the statistical saliency model in MATLAB that analyzes the pixels in an image of an Experiment 1 map to predict the salience of a new colored pushpin added to the map. The model takes a digital color image of an Experment 1 map display and the RGB color value of a new pushpin as input and returns a predicted salience value for the new pushpin. The representation of the pushpin in the implementation of the model is very simple: it consists only of the RGB triplet that will be assigned to the pushpin.

For each map, we first converted the map, pixel-by-pixel, from the RGB color space to the CIELAB color space. We used the implementation of the CIELAB space included with the feature congestion toolbox in the function `RGB2Lab` [34]. The CIELAB color space is a perceptually-based color space in which the distance between two CIELAB color values is designed to represent perceived color difference [35]. Thus, the farther away two colors are in the CIELAB space, the more perceptually different the colors should be. The CIELAB space contains three dimensions. The L^* dimension corresponds to luminance, while the a^* and b^* contain color information. In the implementation of the CIELAB space used in the feature congestion toolbox, the values in the L^* dimension are in the range 0 to 10, while the values in the a^* and b^* dimensions are in the range $-$ 14 to 14.

Next, we calculated the Mahalanobis distance between the pixels in the CIELAB image and the CIELAB values of each of the 267 colors in the ISCC-NBS color centroid list using the built-in MATLAB function `mahal`. This produced a list of 267 salience values. Each value represented the predicted salience of a new pushpin of that ISCC-NBS color added to the map. Then, we identified the colors that produced the minimum, median, and maximum-salience values for that map.

We noticed that some of the model-predicted maximum-salience colors were similar to the colors of existing display items. This might seem counterintuitive, since the statistical saliency model analyzes the display colors in calculating the predicted salience of a new color and might be expected to pick colors that differ from existing display colors. However, the predicted salience is based on a comparison of a single color value with summary statistics of a display (i.e., the mean and covariance of display colors), rather than a pixel-by-pixel or item-by-item color comparison. Furthermore, as we mentioned earlier, each map contained a pushpin at its center that represented the current location of a hypothetical map observer. On a real map, other features (such as letters, numbers, or text) could be added to the pushpins to distinguish them from one another. However, we did not want to add additional features to

TABLE I
MODEL-SELECTED COLORS FOR MINIMUM-SALIENCE TARGETS ON EXPERIMENT 2 MAPS

Color	RGB	Low-Clutter Maps		High-Clutter Maps	
		Percent of Search Targets	Mean Salience (S)	Percent of Search Targets	Mean Salience (S)
Very Pale Green	216, 222, 186	76%	2.30	68%	0.92
Very Pale Blue	193, 202, 202	12%	2.89	32%	0.84
Light Greenish Gray	186, 175, 150	12%	5.81	-	-

the pushpins that we did not model in our implementation of the statistical saliency model.

Thus, before choosing minimum-salience, median-salience, and maximum-salience colors for a new pushpin on a particular map, we eliminated candidate colors that would not be discriminable from the colors of existing pushpins on that map. To do this, we first calculated the Euclidean distance (ΔE^*) between the CIELAB values of each of the existing pushpins on the map and the CIELAB values of each of the 267 ISCC-NBS color candidates. Next, we eliminated color candidates with (ΔE^*) values less than 2.2, in accordance with the finding that the just-noticeable color difference in the CIELAB space is approximately $\Delta E^* = 2.2$ [36], [37].

After eliminating the candidate colors that fell below the just-noticeable difference limit for the other pushpins on a particular map, we identified the minimum, median, and maximum predicted salience colors among the remaining candidate colors. We created three versions of each map, one with a new pushpin assigned the minimum-salience color, one with the median-salience color, and one with the maximum-salience color. The new pushpin's location was chosen randomly from one of the unused locations on the 30-by-74 grid from which the locations of the existing map pushpins were chosen. We also created a target image of the colored pushpin on a white background. We repeated this procedure for each of the 25 high-clutter maps and 25 low-clutter maps, resulting in 50 maps containing a new minimum-salience pushpin, 50 maps containing a new median-salience pushpin, and 50 maps containing a new maximum-salience pushpin.

In typical visual search experiments, half of the trials are "target-absent" trials, meaning that the observer is asked to search for a target that is not present on a display. Thus, we created a series of 150 target-absent maps for use in Experiment 2. We began with 50 Experiment 1 maps that had not been identified as either "high clutter" or "low clutter." For each of these 50 maps, we calculated the predicted salience of a new color added to the display for each of the ISCC-NBS colors. After eliminating colors that fell below the just-noticeable difference limit, we again identified the minimum-, median-, and maximum-salience colors that could be assigned to a new pushpin and created target images of each of these pushpins on a white background. However, we did not create a new map containing each of the targets but made three copies of the Experiment 1 map, which we used as the displays in Experiment 2 target-absent trials. Thus, we created 50 minimum-salience target-absent target/display pairs, 50 median-salience target/display pairs, and 50 maximum-salience target/display pairs.

Our choice to create target-absent maps using Experiment 1 maps that were not identified as high clutter or low clutter is not ideal, since it creates a situation where the distribution of clutter across the target-present and target-absent maps is not equivalent. If participants in Experiment 2 discovered that the target-present maps always contained an extremely high amount of clutter or an extremely low amount of clutter, they could provide the correct target-present/target-absent response without having to search for the target. However, as we will describe, we randomized the order of the target-present and target-absent trials in Experiment 2 and asked observers to search for pushpins of all three salience levels on both target-present and target-absent trials. Thus, we think the amount of clutter on the target-absent maps is unlikely to affect the results of Experiment 2.

C. Exploring the Model's Color Assignments

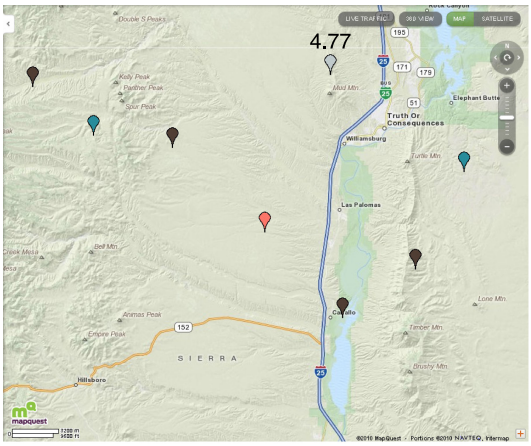
We created the three versions of each map to explore whether the statistical saliency model can choose colors that decrease search time for items on cluttered and uncluttered maps. However, comparing the model's color choices given a map's features also provides an opportunity to understand how the statistical saliency model predicts which colors will be salient and which will not, as well as how the amount of clutter on a display affects which colors the model chooses.

Tables I and II display information about the colors chosen for minimum-salience targets and maximum-salience targets. The first column of Table I lists the ISCC-NBS colors that were assigned most often to minimum-salience pushpins on low-clutter and high-clutter maps. The first column of Table II lists the colors assigned most often to maximum-salience pushpins. For the sake of space, we did not include a table for median-salience colors because there was considerable variability in the median-salience color assignments (24 different colors were assigned to median salience pushpins on low-clutter maps; 22 different colors were assigned to median salience pushpins on high-clutter maps).

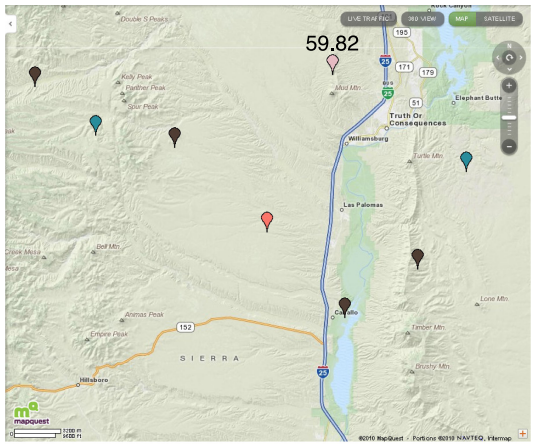
As both tables show, certain ISCC-NBS colors were chosen both for low-clutter maps and high-clutter maps. For example, as Table I shows, very pale green was the model-predicted worst color choice (i.e., the minimum-salience color) for 76% of low-clutter maps and 68% of high-clutter maps. The same pattern occured for maximum-salience pushpins (see Table II), with deep purplish pink being the model predicted best color choice for 76% of low-clutter maps and 72% of high-clutter maps. We urge caution in interpreting the model's color choices. They do not mean, for example, that very pale green should never be used on displays or that deep purplish pink should always be used on displays. Rather, the best and worst color choices depend on the distribution of display features. If we had chosen different displays with different distributions of features, the best and worst choices would have been different.

TABLE II
MODEL-SELECTED COLORS FOR MAXIMUM-SALIENCE TARGETS ON EXPERIMENT 2 MAPS

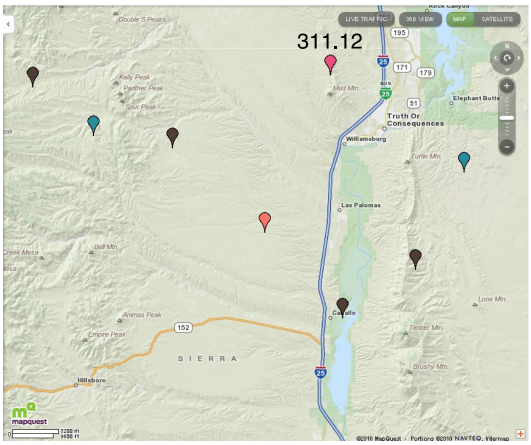
Color	RGB	Low-Clutter Maps		High-Clutter Maps	
		Percent of Search Targets	Mean Salience (S)	Percent of Search Targets	Mean Salience (§)
Deep Purplish Pink	235, 82, 132	76%	271.04	72%	121.97
Vivid Orange Yellow	255, 142, 0	12%	307.74	20%	112.62
Brilliant Purplish Pink	255, 151, 187	4%	430.60	8%	80.92
Brilliant Greenish Yellow	255, 220, 51	8%	212.94	-	-



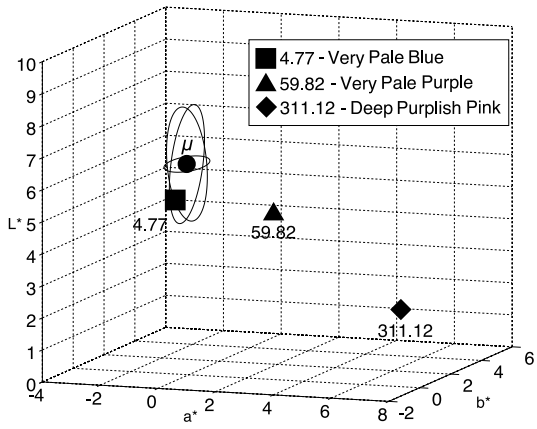
(a)



(b)



(c)



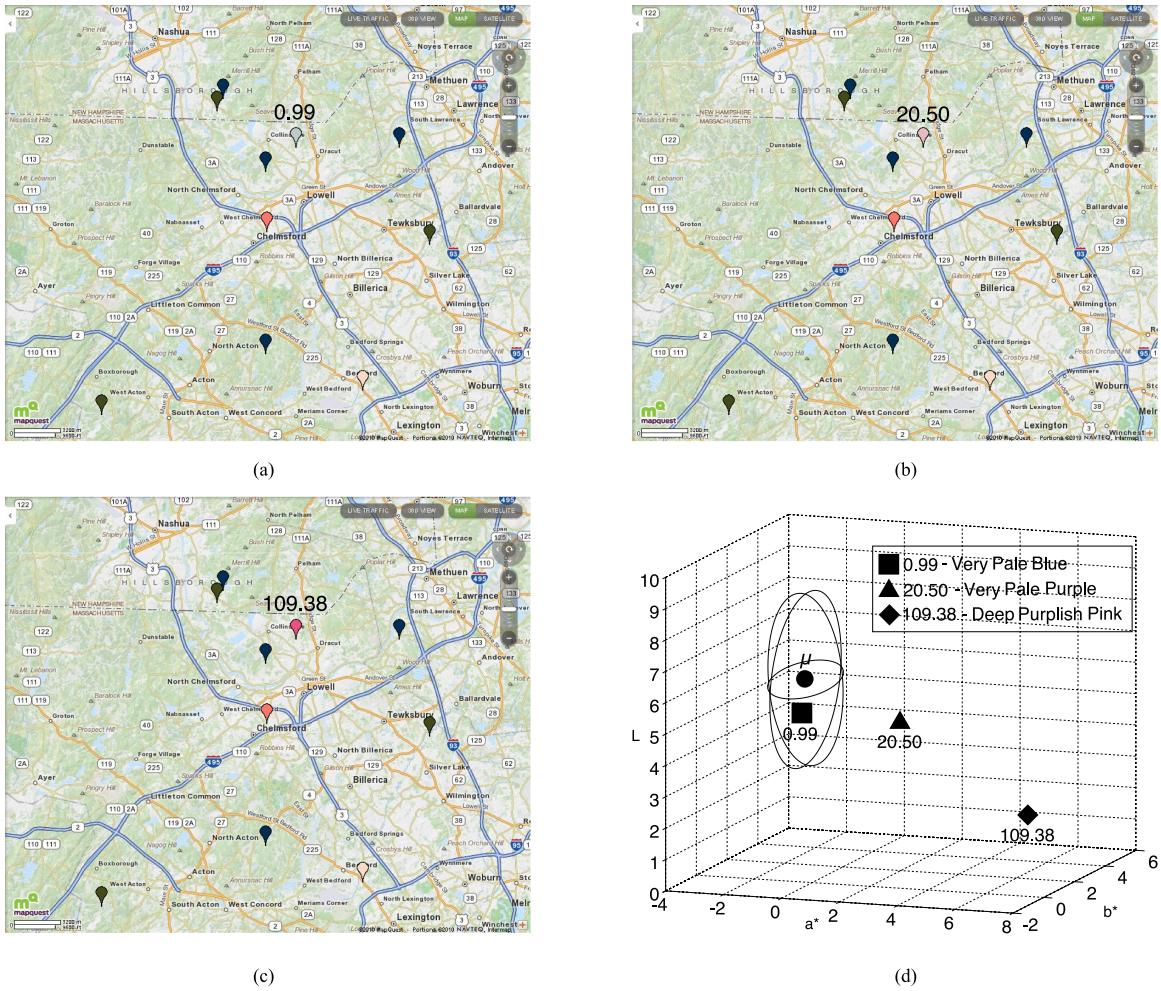
(d)

Fig. 2. (a) Example low-clutter map with the model-predicted minimum-salience color assigned to the new target, indicated by the pushpin with the number directly above it. The number indicates the salience value for the target item. (b) Example low-clutter map with the model-predicted median-salience color assigned to the new target. (c) Example low-clutter map with the model-predicted maximum-salience color added. (d) The locations of the minimum-salience, median-salience, and maximum-salience colors in CIELAB color space relative to the distribution of colored pixels on the map shown in Fig. 1(a).

Fig. 2 illustrates the model's color assignments for the low-clutter Experiment 1 map shown in Fig. 1(a). Fig. 2(a)–(c) shows the three model-constructed versions of the Experiment 1 map shown in Fig. 1(a). All three maps contain a new pushpin not contained on the Experiment 1 map. In Fig. 2(a), the map contains a pushpin whose color is the ISCC-NBS color with the minimum predicted salience ($S = 4.77$), given the display features. In Fig. 2(b), the new pushpin is the color with median predicted salience ($S = 59.82$). Fig. 2(c) contains a new pushpin with the maximum predicted salience ($S = 311.12$). The numeric values that appear directly above the new pushpin indicate the salience of the pushpin. These values did not appear on the maps during Experiment 2.

Fig. 2(d) shows a representation of the CIELAB color space for the maps shown in Fig. 2(a)–(c). The black square, triangle, and diamond represent the respective locations of the minimum, median, and maximum predicted salience colors relative to the display features. The rings represent the error ellipses for each dimension of the CIELAB space. The dimensions of these ellipses are defined by the covariance matrix of color features for the map. The black circle at the center of the ellipses represents the mean color for the display.

Fig. 2(d) illustrates the relative influences of two components of the model's optimal color choices for this map. First, in general, the farther a target color's features (T) are from the average color on the display μ_D, the greater the color's predicted



(a) (b) (c) (d)

Fig. 3. (a) Example high-clutter map with the model-predicted minimum-salience color assigned to the new target, indicated by the pushpin with the number directly above it. The number indicates the salience value for the target item. (b) Example high-clutter map with the model-predicted median salience color assigned to the new target. (c) Example high-clutter map with the model-predicted maximum-salience color added. (d) The locations of the minimum-salience, median-salience, and maximum-salience colors in CIELAB color space relative to the distribution of colored pixels on the map shown in Fig. 1(b).

salience S, as calculated in (1). Thus, deep purplish pink has a higher salience than very pale purple.

Second, the location of the maximum-salience color lies approximately at right angles with the dimensions that contain the most feature variability. For this display, the error ellipses are elongated along the L^* and b^* dimensions. Because of this, the farther a color is from the mean along the a^* dimension, the higher the color's predicted salience will be.

Fig. 3 shows the model's color assignments for the high-clutter Experiment 2 map shown in Fig. 1(b). For the maps in Figs. 2 and 3, the algorithm selected the same minimum-salience, median-salience, and maximum-salience colors. This was uncommon and occurred only twice in the algorithm. As Figs. 2(d) and 3(d) show, the same colors have higher predicted salience values on maps with lower levels of clutter.

A comparison of Figs. 2(d) and 3(d) illustrates the reason for the lower salience on the high-clutter maps. While their means are located in similar regions of CIELAB space, the ellipses in 3(d) are larger than the ellipses in 2(d). Thus, while the model might choose the color with the highest predicted salience given the features of a map, the same color will likely be more salient on a display with less clutter.

IV. Experiment 2: Testing the Model's Color Assignments

A. Participants

We conducted a visual search experiment to test whether model-predicted differences in salience produced differences in search time. In total 44 Tennessee State University students who reported normal or corrected-to-normal vision and normal color vision participated in the experiment and received course credit for their participation.

B. Stimuli

The experimental stimuli were the 300 maps created using the statistical saliency model's predictions. In total 50 of the maps contained a new maximum-salience pushpin, 50 contained a new

median-salience pushpin, and 50 contained a new minimum-salience pushpin. On half of the maps, the target was absent.

The experiment was run using the same computer and monitor from Experiment 1. The experiment was programmed and conducted using ePrime software.

C. Procedure

Each search trial began with a colored pushpin appearing at the center of the screen for 1500 ms. The pushpin measured 20 pixels by 40 pixels (approximately 0.57° by 1.14° of visual angle at a viewing distance of approximately 21 in). Then, the corresponding map display appeared at the center of the screen on a white background. Each map measured 865 pixels by 718 pixels (24.2° by 20.19°).

The display remained on the screen until the participant responded. Observers were instructed to search for the colored pushpin and press "p" if the target was present and "a" if it was absent. Observers were asked to respond as quickly as possible but make as few errors as possible. Next, a screen appeared that displayed feedback for the trial (i.e., whether the observer's response was correct, the speed of the observer's response, and the cumulative percentage of trials on which the observer had made correct responses). Finally, the observer pressed the SPACE bar to begin the next trial.

Observers first completed five practice trials, followed by the 300 experimental trials. The order of the 300 trials was randomized for each observer. The trials were presented in ten groups of 30 trials each. Between groups of trials, observers were instructed to take a break until they were ready to proceed to the next group. The experiment lasted approximately 30 min.

D. Results

We calculated the mean search time across observers for each of the 150 target-present trials where observers made a correct response. Observers made correct responses on 88% of trials.

We did not find evidence for a speed-accuracy tradeoff [26]. Mean reaction times were faster on trials where participants made an accurate response (M = 1316 ms, 95% CI [1303, 1328]) than on trials where participants made an inaccurate response (M = 1441 ms, 95% CI [1396, 1486], t(13198) = 6.60, $d = .18$).

We examined the effect of target salience on search time. Mauchley's test of sphericity indicated that the assumption of sphericity had been violated for the test of the main effect of salience ($\chi^2(2) = 23.61$, $p < 0.001$). A repeated measures ANOVA with a Greenhouse–Geisser correction revealed differences between mean search times for the different pushpin salience levels ($F(1.40,60.14) = 83.56$, $p < 0.001$, $\eta^2 = 0.66$). Pairwise comparisons showed that search times decreased significantly from minimum-salience pushpins (M = 1312 ms, 95% CI [1222, 1402]) to median-salience pushpins (M = 1095 ms, 95% CI [1024, 1165]) and again from median-salience pushpins to maximum-salience pushpins (M = 943 ms, 95% CI [883, 1002]). Thus, assigning minimum-, median-, or maximum-salience colors to items produced differences in the time it took observers to find those items.

We also tested the effects of clutter on search time. A repeated measures ANOVA showed a significant main effect of clutter ($F(1,43) = 16.85$, $p < 0.001$, $\eta^2 = .28$. Pairwise comparisons using the Bonferroni correction revealed that mean search times were longer for pushpins on high-clutter maps (M = 1160 ms, 95% CI [1085, 1235]) than on low-clutter maps (M = 1073 ms, 95% CI [1007, 1140]).

Finally, we tested the interaction of target salience and map clutter. Mauchley's test of sphericity revealed that the test of the salience × clutter interaction violated the assumption of sphericity ($\chi^2(2) = 8.43$, $p < 0.015$). A repeated measures ANOVA using the Greenhouse–Geisser correction revealed a significant interaction of target salience and map clutter ($F(1.69,72.77) = 9.24$, $p < 0.001$, $\eta^2 = 0.18$). Increased map clutter slowed search for low-salience targets. However, we did not find evidence for this effect of clutter for medium salience or high salience targets.

V. Discussion

Some maps may have a moderate amount of clutter, even when extraneous layers are not displayed. In these cases, the choice of distinctive colors is not obvious. In the present study, we have demonstrated that a computational model of visual search can choose highly-salient colors for items on those maps. The primary strength of the model we tested in this study is that it has zero free parameters. This is because the CIELAB color space used to represent the color features of the target and display is a standardized, perceptually-based space. Our implementation of the statistical saliency model is also extremely flexible, in that it works directly with the input display image. As we have shown, despite the simplicity of the model, its design suggestions produced significant decreases in search time in both low-clutter and high-clutter maps.

One may question whether the search time benefits of increased pushpin salience have practical value. After all, the difference between the mean search times for pushpins of different salience values is relatively small (i.e., approximately 370 ms for maximum salience pushpins and minimum salience pushpins). However, the Experiment 2 target stimuli were defined by a single feature. With more complex stimuli or search situations, the benefits of choosing salient colors may increase. Furthermore, small benefits of salience will accumulate across multiple searches with the same map.

While the method we describe reduces search time, it has limitations. As mentioned earlier, we chose the value of a single feature of an item. We did this because representing multiple features of an item would require fitting a parameter for each feature. In addition, the number of features an object possesses is not easy to quantify. Thus, restricting the model to a single feature simplifies the analysis. However, there are situations in which a designer might want to consider more than one feature in the brute-force algorithm. This would require a way of quantifying feature space for that feature and fitting the parameters for the weights for different feature values.

The statistical saliency model has its own limitations as well. For example, it does not represent eye movements or differences in acuity across the visual field, both of which are known to affect

search performance [15], [38]. In fact, recent work by Rosenholtz and colleagues has proposed a move away from feature-based theories of visual search in favor of theories proposing a summary statistic representation for information in peripheral vision [39]. Furthermore, high-salience items may not always be found quickly. Other display features may draw attention, or the observer's search may depend on the memory of previous searches. In addition, if the target's color is not known, the observer may rely on other display features when deciding where to search for the item.

Cartographers have developed a variety of quantitative and qualitative methods for choosing color schemes for items on maps. Furthermore, mapping services use existing color code conventions for display items, including pushpins. The approach we describe may have potential as a complement to those approaches. For example, it may be useful for choosing colors for novel items on maps that may not be members of a category with an existing color code convention. In future work, we intend to design search tasks and stimuli that provide a better approximation of how observers interact with maps. For example, the MapQuest stimuli we collected for use in this experiment were selected for convenience. In future work, we plan to test stimuli from national mapping agencies, such as the United States Geological Survey.

VI. Conclusion

Effective design of map displays requires a collaboration between cartographic science and cognitive science [12]. A research strategy that emphasizes thematic relevance and perceptual salience has proven effective in assessing and improving both static and dynamic displays [19], [40]. In this project, we have shown that the statistical saliency model, a measure of perceptual salience, can choose distinctive colors for items on low-clutter and high-clutter maps. Future work will examine the effect of local display clutter on the model's ability to assign colors to maximize item salience and consider searches for targets represented by more than one feature.

Appendix
Identifying High- and Low-Clutter Maps

Experiment 2 required that we analyze the subjective clutter ratings of the 145 Experiment 1 maps to identify a group of maps with low clutter ratings and a second group of maps with high clutter ratings. Because Experiment 2 compared search time for pushpins on low-clutter and high-clutter maps, we wanted the average clutter ratings of the maps in the two groups to differ as much as possible. However, because the Experiment 1 ratings were positively skewed, this was not a trivial task. Simply assigning the 25 maps with the lowest average ratings to the "low-clutter" group and the 25 maps with the highest average ratings to the "high-clutter" group would produce groups with unequal variances. The low-clutter group would contain maps whose clutter ratings ranged from 1.70 to 2.62, while the high-clutter group would have maps whose ratings ranged from 4.30 to 6.38.

We designed a brute-force algorithm to create groups of maps whose average clutter ratings differed as much as possible but whose standard deviations were as similar as possible. The algorithm considered all possibilities for assigning 25 maps to the low-clutter group and 25 maps to the high-clutter group. For each candidate assignment, the algorithm calculated the mean and standard deviation of clutter ratings in each group. Then, the algorithm assessed the quality of the group assignment by considering the difference in clutter ratings the assignment produced

$$\text{ClutterDifference} = \frac{(\mu_{\text{low}} - \mu_{\text{high}})^2}{(\sigma_{\text{low}} - \sigma_{\text{high}})^2 + 0.0001} \qquad (2)$$

where μ_{low} and μ_{high} are the mean clutter ratings for the 25 maps in the low-clutter group and the 25 maps in the high-clutter group, σ_{low} and σ_{high} are the standard deviations of the clutter ratings for the 25 low-clutter and 25 high-clutter maps, and 0.0001 is a constant that prevents ClutterDifference from going to infinity when the algorithm considers a set of map assignments with equal standard deviations. Adding 0.0001 to the result in the denominator allows the algorithm to differentiate between map assignments where the standard deviations are equal but the means are different. As (2) shows, the optimal assignment of maps to clutter groups is the one that maximizes the difference between the group means while minimizing the difference between the standard deviation of the clutter ratings of the groups.

Acknowledgment

The authors would like to thank G. Francis and J.-C. Pedjeu for helpful conversations regarding this project.

References

[1] M. Hegarty, "Cognition, metacognition, and the design of maps," *Current Directions Psychological Sci.*, vol. 22, no. 1, pp. 3–9, 2013.
[2] B. D. Dent, J. S. Torguson, and T. W. Hodler, *Cartography: Thematic Map Design*, 6th ed. Boston, MA, USA: McGraw-Hill, 1999.
[3] T. A. Slocum, R. M. McMaster, F. C. Kessler, H. H. Howard, and R. B. Mc Master, *Thematic Cartography and Geographic Visualization*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2008.
[4] M. Harrower and C. A. Brewer, "Colorbrewer.org: An online tool for selecting colour schemes for maps," *Cartographic J.*, vol. 40, no. 1, pp. 27–37, 2003.
[5] E. Chesneau, "A model for the automatic improvement of colour contrasts in maps: Application to risk maps," *Int. J. Geographical Inf. Sci.*, vol. 25, no. 1, pp. 89–111, 2011.
[6] M. P. Eckstein, "Visual search: A retrospective," *J. Vision*, vol. 11, no. 5, pp. 1–36, 2011.
[7] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[8] E. M. Palmer, T. S. Horowitz, A. Torralba, and J. M. Wolfe, "What are the shapes of response time distributions in visual search?" *J. Exp. Psychol., Human Percept. Perform.*, vol. 37, no. 1, pp. 58–71, 2011.
[9] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nature Human Behav.*, vol. 1, no. 3, pp. 1–8, 2017.
[10] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Rev.*, vol. 96, no. 3, pp. 433–458, 1989.
[11] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.
[12] S. Garlandini and S. I. Fabrikant, "Evaluating the effectiveness and efficiency of visual variables for geographic information visualization," in *Proc. Int. Conf. Spatial Inf. Theory*, Sep. 2009, vol. 5756, pp. 195–211.
[13] C. Koch and S. Ullman, " Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[14] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vision Res.*, vol. 116, pp. 95–112, 2015.

[15] G. J. Zelinsky, "A theory of eye movements during target acquisition," *Psychological Rev.*, vol. 115, no. 4, pp. 787–835, 2008.
[16] A. D. Hwang, E. C. Higgins, and M. Pomplun, "A model of top-down attentional control during visual search in complex scenes," *J. Vision*, vol. 9, no. 5, pp. 25–25, 2009.
[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 2106–2113.
[18] J. Shive and G. Francis, "Choosing colors for map display icons using models of visual search," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 55, no. 2, pp. 373–396, 2013.
[19] S. I. Fabrikant and K. Goldsberry, "Thematic relevance and perceptual salience of dynamic geovisualization displays," in *Proc. 22th Int. Cartographic Conf.*, 2005, pp. 11–16.
[20] M. Hegarty, M. S. Canham, and S. I. Fabrikant, "Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task," *J. Exp. Psychol., Learn., Memory, Cognition*, vol. 36, no. 1, pp. 37–53, 2010.
[21] S. I. Fabrikant, S. R. Hespanha, and M. Hegarty, "Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making," *Ann. Assoc. Amer. Geographers*, vol. 100, no. 1, pp. 13–29, 2010.
[22] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *J. Vision*, vol. 7, no. 2, pp. 1–22, 2007.
[23] N. Moacdieh and N. Sarter, "Clutter in electronic medical records examining its performance and attentional costs using eye tracking," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 4, pp. 591–606, 2015.
[24] R. F. Haines, E. Fischer, and T. A. Price, "Head-up transition behavior of pilots with and without head-up display in simulated low-visibility approaches," NASA Tech. Rep. 1720, 1980.
[25] M. R. Beck, M. C. Lohrenz, and J. G. Trafton, "Measuring search efficiency in complex visual search tasks: Global and local clutter," *J. Exp. Psychol., Appl.*, vol. 16, no. 3, pp. 238–250, 2010.
[26] J. Wilkening and S. I. Fabrikant, "How do decision time and realism affect map-based decision making?" in *Proc. Int. Conf. Spatial Inf. Theory*, 2011, vol. 6899, pp. 1–19.
[27] M. Yeh and C. D. Wickens, "Attentional filtering in the design of electronic map displays: A comparison of color coding, intensity coding, and decluttering techniques," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 43, no. 4, pp. 543–562, 2001.
[28] J. Korpi and P. Ahonen-Rainio, "Clutter reduction methods for point symbols in map mashups," *Cartographic J.*, vol. 50, no. 3, pp. 257–265, 2013.
[29] P. Olsson, K. Pippig, L. Harrie, and H. Stigmar, "Identifying areas of a map that are difficult to read," *Mapping Image Sci.*, no. 3, pp. 22–29, 2011.
[30] R. Rosenholtz, "Search asymmetries? What search asymmetries?" *Perception Psychophysics*, vol. 63, no. 3, pp. 476–489, 2001.
[31] K. L. Kelly and D. B. Judd, *Color: Universal Language and Dictionary of Names*. Washington, D.C., USA: U.S. Dept. Commerce, National Bureau of Standards, 1976.
[32] Psychology Software Tools Inc., 2000, E-prime 2.0. [Online]. Available: https://www.pstnet.com
[33] W. Schnieder, A. Eschman, and A. Zuccolotto, *E-Prime User's Guide*. Sharpsburg, PA, USA: Psychol. Softw. Tools Inc., 2002.
[34] R. Rosentholtz, Y. Li, and L. Nakano, "Feature congestion and subband entropy measures of visual clutter," Mar. 2018. [Online]. Available: http://dspace.mit.edu/handle/1721.1/37593
[35] M. D. Fairchild, *Color Appearance Models*. Reading, MA, USA: Addison-Wesley, 1998.
[36] M. Stokes, M. D. Fairchild, and R. S. Berns, "Precision requirements for digital color reproduction," *ACM Trans. Graph.*, vol. 11, no. 4, pp. 406–422, 1992.
[37] D. H. Brainard *et al.*, "Color appearance and color difference specification," *Sci. Color*, vol. 2, pp. 191–216, 2003.
[38] S. M. Anstis, "A chart demonstrating variations in acuity with retinal position," *Vision Res.*, vol. 14, no. 7, pp. 589–592, 1974.
[39] R. Rosenholtz, J. Huang, A. Raj, B. J. Balas, and L. Ilie, "A summary statistic representation in peripheral vision explains visual search," *J. Vision*, vol. 12, no. 4, pp. 14–14, 2012.
[40] G. Francis, K. Bias, and J. Shive, "The psychological four-color mapping problem," *J. Exp. Psychol., Appl.*, vol. 16, no. 2, pp. 109–123, Jun. 2010.

Joshua Shive received the Ph.D. degree in cognitive psychology from Purdue University, West Lafayette, IN, USA, in 2008.

He is an Associate Professor of psychology with Tennessee State University in Nashville, TN, USA. In 2018, he was a Summer Faculty Fellow at the NASA Marshall Space Flight Center. His research interests include human factors, display design, and visual perception.

Sharra Rosichan received the Bachelor of Science degree in psychology from the Tennessee State University, Nashville, TN, USA, in 2011. She is currently working toward the Master's degree in information sciences at the University of Tennessee, Knoxville, TN, USA.

Sherika Davis received the Bachelor of Science degree in psychology from Tennessee State University, Nashville, TN, USA, in 2012.

Christena Wade received the Bachelor of Science degree in psychology from Tennessee State University, Nashville, TN, USA, in 2012, and the Bachelor of Science in nursing from Belmont University, Nashville, TN, USA, in December 2017.

She has been a Registered Nurse with the state of Tennessee since February 2018 and currently with the Vanderbilt University Medical Center in Nashville, TN, USA.

James Ellison received the Bachelor of Science degree in psychology from Tennessee State University, Nashville, TN, USA, in 2015. He is currently working toward the master's degree in clinical psychology from Middle Tennessee State University, Murfreesboro, TN, USA.

His research interest include clinical psychology and cognitive psychology.

Santos Santoni-Sanchez was a student in the Bachelor of Science in Psychology program at Tennessee State University, Nashville, TN, USA, in 2014.

A Time-Efficient Approach for Decision-Making Style Recognition in Lane-Changing Behavior

Sen Yang, Wenshuo Wang, *Member, IEEE*, Chao Lu, Jianwei Gong, *Member, IEEE*, and Junqiang Xi

***Abstract*—Fast recognition of a driver's decision-making style when changing lanes plays a pivotal role in a safety-oriented and personalized vehicle control system design. This article presents a time-efficient recognition method by integrating k-means clustering (k-MC) with the K-nearest neighbor (KNN) algorithm, called kMC-KNN. Mathematical morphology is implemented to automatically label the decision-making data into three styles (moderate, vague, and aggressive), while the integration of k-MC and the KNN algorithm helps to improve the recognition speed and accuracy. Our developed mathematical-morphology-based clustering algorithm is then validated by a comparison with agglomerative hierarchical clustering. Experimental results demonstrate that the developed kMC-KNN method, in comparison with the traditional KNN algorithm, can shorten the recognition time by more than 72.67% with a recognition accuracy of 90–98%. In addition, our developed kMC-KNN method also outperforms a support vector machine in terms of recognition accuracy and stability. The developed time-efficient recognition approach would have great application potential for in-vehicle embedded solutions with restricted design specifications.**

***Index Terms*—Decision-making style classification and recognition, k-means-clustering-based K-nearest neighbor (kMC-KNN), lane change behaviors, mathematical morphology.**

I. Introduction

A. Motivation

MAKING a human-friendly decision when changing lanes is crucial to intelligent vehicle control [1], traffic efficiency and road safety [2], and human-like autonomous driving systems [3]. Human drivers generate and carry out various decision-making policies to determine if, when, and how to change lanes [4] by evaluating the current driving situation according to their internal model [5]. Modeling such various driver decision-making processes is a nontrivial task for applications. It is relatively easy to establish personalized models for a small group of drivers by using advanced learning methodologies [6], but this is not feasible for hundreds of millions of human drivers due to the excessive time cost and resources. Building a model for each group of drivers with similar driving characteristics, instead of a single model for each driver, will be a cost-effective solution for this issue. In other words, classifying decision-making styles into several distinguishable groups allows us to efficiently describe a large number of human drivers at low cost. Therefore, it is necessary to classify drivers into groups and then analyze their lane-changing behavior.

B. Related Research

In general, a complete lane change task consists of two parts: decision making and action execution. Much existing research concerning the operation style of drivers has been conducted, for instance, subjectively classifying and labeling drivers' steering signals of double-lane-changing maneuvers into several groups according to prior knowledge of the driving style [6], [7]. These labeled data were then used to train a classifier based on supervised learning methodologies, such as the support vector machine (SVM) [8] and fuzzy logic [9].

In terms of decision making, drivers will prefer different lane-changing strategies, depending on their mandatory and discretionary demands [10], [11] and the driving situations [12]. Sun and Elefteriadou [13] conducted an instrumented vehicle-based experiment and found that the urban arterial lane-changing decision-making process heavily depends on the driver characteristics. They also proposed a comprehensive framework for modeling the drivers' lane-changing maneuvers [14] by computing the lane-changing probability for each scenario considering different driver types [15]. To enrich lane-changing models in traffic simulation packages, Keyvan-Ekbatani *et al.* [11] categorized drivers' lane-changing strategies into different groups based on a two-stage framework consisting of online testing and offline reviewing, but this method could lead to subjective and empirical results. In addition, a driving style questionnaire was implemented by having distinguishing drivers provide scores to surveyed questions [16]. The aforementioned classification methods are supervised, but they are commonly time-consuming when manually labeling large amounts of driving data. In order to improve classification performance with a small labeling effort, a semisupervised SVM was developed to classify drivers into aggressive and moderate driving styles with few labeled

Manuscript received November 25, 2017; revised April 10, 2018, September 29, 2018, and February 17, 2019; accepted July 30, 2019. Date of publication September 13, 2019; date of current version November 21, 2019. This work was supported by the National Natural Science Foundation of China under Grant 91420203 and Grant 61703041. This article was recommended by Associate Editor M. Tanelli. *(Sen Yang and Wenshuo Wang contributed equally to this work.) (Corresponding author: Junqiang Xi.)*

S. Yang, J. Gong, and J. Xi are with the Department of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: yangsen1990@bit.edu.cn; gongjianwei@bit.edu.cn; xijunqiang@bit.edu.cn).

W. Wang is with the Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15232 USA (e-mail: wwsbit@gmail.com).

C. Lu is with the Department of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China, and also with the Advanced Vehicle Engineering Centre, Cranfield University, Cranfield MK43 0AL, U.K. (e-mail: chaolu@bit.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2938155

data among all the collected driving data [17]. However, in applications, it is intractable to prepare objective annotations for training data, since this approach does not fully rule out personal subjective impacts.

Numerous logic-based methods have been applied to improve the recognition performance, but they are computationally expensive and require prior knowledge of data. For example, fuzzy reasoning methods were used to infer a driver's lane-changing intent [12] and identify their driving style [9], which highly relied on prior knowledge and experience of data analysts and their observations, statistics, and analysis [18]. On the other hand, advanced machine-learning techniques have also been implemented to recognize driving style. For example, Enev *et al.* [19] applied four machine-learning algorithms, including the SVM, random forest, naive Bayes, and KNN algorithms, to recognize drivers' driving styles. Wang and Xi [20] proposed a pattern recognition algorithm by combining k-means clustering (k-MC) and an SVM together to shorten the recognition time and improve the recognition performance. In order to address the uncertainty of driver's behavior in driving style recognition, a statistical-based recognition method using Bayesian probability and kernel density estimation was proposed [21]. Zhang *et al.* [22] investigated three direct pattern recognition approaches to classify driver's steering operation skills in a double-lane-changing task, including multilayer perception artificial neural networks, a decision tree, and an SVM. For more state-of-the-art approaches related to driving style recognition, refer to the literature [1], [23], and [24]. The abovementioned algorithms greatly improve the recognition accuracy; however, they usually require a long time to train models to obtain satisfactory results [20], especially when dealing with big data.

The above-discussed literature mainly has two limitations: 1) requiring sufficient prior knowledge to manually label the training data, which is practically intractable for multidimensional large-scale driving data [25], [26]; and 2) learning classifiers and recognizing the driving style for a new observation takes a long time, which impedes these algorithms from being used in online applications.

C. Contributions

This article aims to develop a time-efficient way to improve the recognition performance of drivers' decision-making style in lane-changing scenarios with little subjective interference involved while labeling the training data. Our main contributions cover two aspects, which are listed as follows:

1) proposing an unsupervised method based on mathematical morphology to label the training data, which does not require prior knowledge of clusters or other parameters, thereby reducing the efforts of tagging data and excluding the subjective influences of data analysts;
2) developing a k-MC-based K-nearest neighbor (kMC-KNN) method to accelerate the recognition process and, thus, shorten the recognition time.

D. Article Organization

This article is organized as follows. Section II presents the mathematical morphology method and the kMC-KNN method. Section III describes the driving scenarios and data collection. Section IV presents the experimental results. Finally, conclusions are presented in Section V.

II. Classification and Recognition Methods

This section will present the approaches for training data autoclassification and new data recognition. First, we will describe in detail the mathematical morphology method, which can automatically label the collected data without any prior knowledge. Then, we will describe our proposed kMC-KNN method for driving style recognition.

A. Classification Method

Mathematical morphology was primarily constructed as a nonlinear processing and analysis tool [27] for image segmentation in order to obtain a good description and representation of the shapes of segments [28]. The expanding applications of mathematical morphology also cover, for instance, boundary detection, automatic image segmentation and reconstruction, pattern recognition, and signal and image decomposition [29]. Inspired by the its advantages and these applications, in this article, we employ the fundamental operators of mathematical morphology [30]—the *dilation* process and *erosion* process—to search data with the same characteristics and then cluster the data.

1) Dilation and Erosion: Given an original set $\mathbf{A} \subset \mathbb{Z}^d$ and a kernel set $\mathbf{B}(x) = \mathbf{B}_x^\vee = \{b - x : b \in \mathbf{B}\}$ with the point x as its origin [30], the morphological dilation and erosion of $\mathbf{A}$ by $\mathbf{B}$ are defined as (1) and (2), respectively, as follows:

$$\mathbf{A} \oplus \mathbf{B} \equiv \{x : \mathbf{B}_x^\vee \cap \mathbf{A} \neq \varnothing\} \tag{1}$$

$$\mathbf{A} \ominus \mathbf{B} \equiv \{x : \mathbf{B}_x^\vee \subseteq \mathbf{A}\}. \tag{2}$$

In order to understand the operations of dilation and erosion, a visualized example is shown in Fig. 1. The dilation process of $\mathbf{A}$ by $\mathbf{B}$ [see Fig. 1(c)] is achieved by adding the pixels of $\mathbf{B}$ into $\mathbf{A}$ when the origin pixel of $\mathbf{B}$ passes through $\mathbf{A}$, while the erosion process of $\mathbf{A}$ by $\mathbf{B}$ [see Fig. 1(d)] is achieved by removing these pixels of $\mathbf{A}$ in which $\mathbf{B}$ is not completely contained. From this example, it can be seen that the dilation operation merges the points around the target area ($\mathbf{A}$) to fill small holes in the area and small depressions at the edges of the area, while the erosion operation removes horns smaller than the kernel structure ($\mathbf{B}$).

2) Mathematical-Morphology-Based Clustering Algorithm: In order to objectively explore irregular clusters of the driving style, we develop a clustering algorithm by making full use of dilation and erosion, which can discover such clusters with an arbitrary shape and automatically determine the number of clusters by making full use of the underlying data information [31], [32]. Given a dataset $\{\mathbf{x}^{(n)}\}_{n=1}^N$, where $N \in \mathbb{N}^+$ is the length of the data and $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_I)$ is a vector with $I \in \mathbb{N}^+$

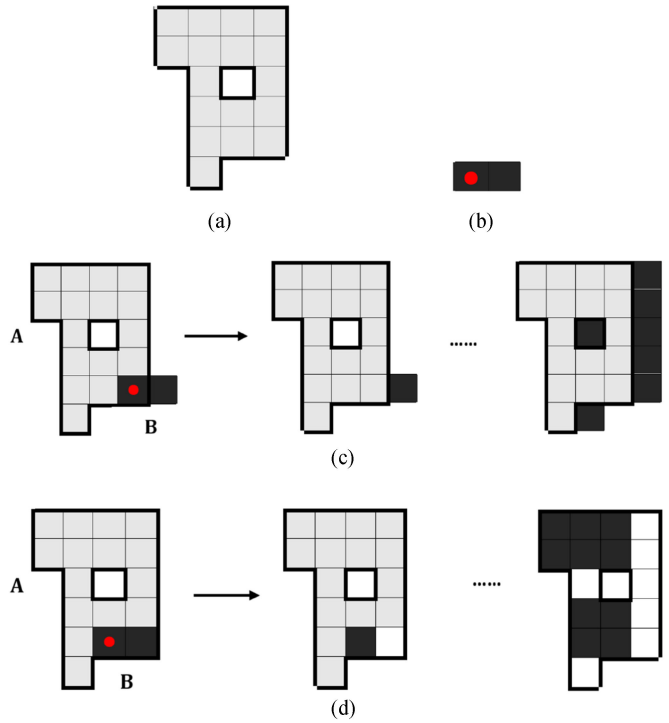


Fig. 1. Illustration of the dilation and erosion operations. (a) Original set **A**. (b) Kernel set **B**. The pixel marked with a red circle is the origin x. (c) Dilation process of **A** by **B** (i.e., $\mathbf{A} \oplus \mathbf{B}$). The pixels that were added by the dilation operations are marked in light black. (d) Erosion process of **A** by **B** (i.e., $\mathbf{A} \ominus \mathbf{B}$). The pixels retained by the erosion operation are marked in light black.

variables, the procedure of the mathematical-morphology-based clustering algorithm can be achieved with the following steps.

1) Given the dataset $\{\mathbf{x}^{(n)}\}$, we obtain the normalized dataset $\{\overline{\mathbf{x}}^{(n)}\}$ by (3) and then transform $\{\overline{\mathbf{x}}^{(n)}\}$ into a positive integer set $\{\widetilde{\mathbf{x}}^{(n)}\}$ with a value between 1 and $q+1$ by (4) as follows:

$$\overline{x}_i^{(n)} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

$$\widetilde{x}_i^{(n)} = \text{fix}\left(\overline{x}_i^{(n)} \times q_i\right) + 1 \tag{4}$$

where $n = 1, \ldots, N, i = 1, \ldots, I$, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum of x_i, respectively, $\overline{x}_i^{(n)}$ and $\widetilde{x}_i^{(n)}$ are the elements of $\overline{\mathbf{x}}^{(n)}$ and $\widetilde{\mathbf{x}}^{(n)}$, respectively, $\text{fix}(\cdot)$ in (4) is the truncated function, and q_i is a suitable integer for the parameter x_i.

2) Then, set matrix $\mathbf{A}_{q_1 \times q_2 \times \cdots \times q_I}$ with $\mathbf{A}(\widetilde{\mathbf{x}}^{(n)}) = 1$ if the element of **A** is equal to $\widetilde{\mathbf{x}}^{(n)}$, and otherwise set it to 0. Until now, the given data have been converted to the original matrix **A** filled by 0 or 1. For **B**, we choose the kernel matrix to be spherical with a suitable radius r, which is much smaller than the original matrix **A**.

3) The dilation result $\mathbf{A}_1$ of **A** by **B** is then obtained by (1), and $\mathbf{A}_2$ is the result of the erosion of $\mathbf{A}_1$ by $\mathbf{B}'$ whose radius is $r+1$ based on (2). These collected areas in $\mathbf{A}_2$ are clusters, and the number of collected areas, denoted as J, is the number of clusters. A cluster with a small amount of data is regarded as noise and then removed from $\mathbf{A}_2$.

4) The data $\mathbf{x}^{(n)}$ with the shortest Euclidean distance between the cluster centers C_j and $\mathbf{x}^{(n)}$ belong to the cluster j, computed as follows:

$$\hat{j}^{(n)} = \arg\min_j \left\|\mathbf{x}^{(n)} - C_j\right\| \tag{5}$$

with $j = 1, 2, \ldots, J$.

B. Recognition Method

Before introducing kMC-KNN, we shall present the preliminaries of the KNN and kMC methods.

1) KNN: The KNN algorithm is a nonparametric classification method that has been widely used in many research fields, such as text categorization and image processing. Given a labeled dataset $\mathcal{M} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$ and new data $\mathbf{x}^{(*)}$, where $y^{(n)} \in \{y_j\}_{j=1}^{J}$ are the labels of the training samples and $\{y_j\}$ is the set of labels. The KNN algorithm is described as follows.

a) Normalize the training samples $\mathbf{x}^{(n)}$ and data $\mathbf{x}^{(*)}$ using (3), and then, attain the normalized training sample set $\{\overline{\mathbf{x}}^{(n)}\}$ and the normalized new data $\overline{\mathbf{x}}^{(*)}$.

b) Evaluate the similarity levels between the training samples $\mathbf{x}^{(n)}$ and data $\mathbf{x}^{(*)}$ using (6). Then, choose $K = \sqrt{N}$ of the most similar samples as the KNN collection of $\mathbf{x}^{(*)}$

$$\text{SIM}\left(\mathbf{x}^{(n)}, \mathbf{x}^{(*)}\right) = \left\|\overline{\mathbf{x}}^{(n)} - \overline{\mathbf{x}}^{(*)}\right\|^2. \tag{6}$$

c) Decide the category of the given data $\mathbf{x}^{(*)}$ using

$$\hat{j}^{(*)} = \arg\max_j \sum_{n=1}^{K} \text{SIM}\left(\mathbf{x}^{(n)}, \mathbf{x}^{(*)}\right) \cdot \rho\left(\mathbf{x}^{(n)}, y_j\right) \tag{7}$$

with

$$\rho\left(\mathbf{x}^{(n)}, y_j\right) = \begin{cases} 1, & \text{if } y^{(n)} = y_j \\ 0, & \text{if } y^{(n)} \neq y_j \end{cases}$$

where $j = 1, 2, \ldots, J$ and $n = 1, 2, \ldots, K$.

The major limitation of the KNN method is that a large number of design vectors in the trained classifier will significantly increase the computational complexity in recognizing new data samples, which impedes its applications to vehicle dynamics, wherein high-dimensional variables are required for classification.

2) k-MC: MacQueen [33] first proposed the k-MC algorithm, which partitions the given n objects into k clusters with each object belonging to the cluster with the nearest mean. The k-MC algorithm includes four basic steps: initialization, assignment, update, and repeat. Given a dataset $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$, four steps should be taken as follows.

a) *Initialization*: Normalize $\{\mathbf{x}^{(n)}\}$ using (3), and obtain the normalized data $\{\overline{\mathbf{x}}^{(n)}\}_{n=1}^{N}$. Randomly choose k instances $\{\mathbf{m}_1^{(\eta)}\}_{\eta=1}^{k}$ from $\{\overline{\mathbf{x}}^{(n)}\}$ as the initial conditions.

b) *Assignment*: Assign each data point $\mathbf{x}^{(n)}$ to the nearest cluster according to $\hat{\eta}^{(n)}$

$$\hat{\eta}^{(n)} = \arg\min_\eta \left\|\mathbf{m}_t^{(\eta)} - \overline{\mathbf{x}}^{(n)}\right\|. \tag{8}$$

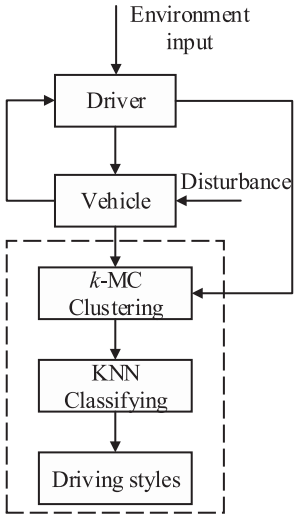


Fig. 2. Schematic diagram of our developed kMC-KNN method.

c) *Update*: Adjust the means $\mathbf{m}_t^{(\eta)}$ to match the sample means of the data points through the following formula:

$$\mathbf{m}_t^{(\eta)} = \frac{\sum_n \beta_\eta^{(n)} \overline{\mathbf{x}}^{(n)}}{\sum_n \beta_\eta^{(n)}}. \tag{9}$$

with

$$\beta_\eta^{(n)} = \begin{cases} 1, & \text{if } \hat{\eta}^{(n)} = \eta \\ 0, & \text{if } \hat{\eta}^{(n)} \neq \eta \end{cases}.$$

d) *Repeat*: Repeat the *assignment* step and *update* step until the assignments do not change

$$\mathbf{m}_t^{(\eta)} = \mathbf{m}_{t+1}^{(\eta)}. \tag{10}$$

3) kMC-KNN Algorithm: We developed a kMC-KNN algorithm (as shown in Fig. 2) to overcome the mentioned limitation of the KNN method in the classification procedure [34]. With this purpose in mind, we apply k-MC to search representatives of the whole training data to reduce the computational cost of the KNN algorithm. There are three main steps in this recognition method, including clustering, selecting, and classifying. Clustering is used to train the recognition model, and selecting and classifying are used to identify the pattern of the new data.

a) *Clustering:* The k-MC algorithm clusters the training samples of each category into k subclusters. As a consequence, we obtain the recognition model.

b) *Selecting:* For new input data, unlike the traditional KNN algorithm, which recognizes its patterns with all the training data in each category, the kMC-KNN algorithm selects the subset with the largest similarity between the given data and the center of each subset in each category as the training samples as follows:

$$\hat{\eta}^{(*)} = \arg\min_{\eta} \left\| \mathbf{x}^{(*)} - \hat{\mathbf{x}}^{(\eta)} \right\| \tag{11}$$

where $\hat{\mathbf{x}}^{(\eta)}$ is the mean of the data in subclusters $\{(\mathbf{x}^{(\eta)}, y_j)\}_{\eta=1}^{k}$.

Algorithm 1: Algorithm for kMC-KNN.

Training

Given the labeled dataset $\mathcal{M} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$, $y^{(n)} \in \{y_j\}_{j=1}^{J}$, obtain $\{(\overline{\mathbf{x}}^{(n)}, y^{(n)})\}_{n=1}^{N}$ using (3).
for $j = 1$ to J **do**
  Randomly choose k instances $\{(\mathbf{m}^{(\eta)}, y_j)\}_{\eta=1}^{k}$ from $\{(\overline{\mathbf{x}}^{(n)}, y_j)\}_{n}^{N_j}$.
  while $\mathbf{m}_t^{(\eta)} \neq \mathbf{m}_{t+1}^{(\eta)}$ **do**
    Update $\mathbf{m}_t^{(\eta)}$ using (9)
    Assign $\mathbf{x}^{(n)}$ to $\hat{\eta}^{(n)}$ using (8)
  end while
  Obtain the final k subclusters $\{(\mathbf{x}^{(\eta)}, y_j)\}_{\eta=1}^{k}$
end for
Obtain the $J \times k$ subclusters $\{(\mathbf{x}^{(\eta)}, y_j)\}_{\eta=1}^{k}$, $j = 1, 2, \ldots, J$

Testing

Given the new data $\mathbf{x}^{(*)}$, we obtain $\overline{\mathbf{x}}^{(*)}$ using (3).
for $j = 1$ to J **do**
  Select the nearest subcluster $\{(\mathbf{x}^{(\eta^*)}, y_j)\}$ using (11)
end for
Obtain the selected J subclusters $\{(\mathbf{x}^{(\eta^*)}, y_j)\}_{j=1}^{J}$
for $n = 1$ to N **do**
  Calculate $\mathrm{SIM}(\mathbf{x}^{(n)}, \mathbf{x}^{(*)})$ using (6)
end for
Choose the $K = \sqrt{N}$ samples with minimum similarity
Judge $\mathbf{x}^{(*)}$ to $\hat{j}^{(*)}$ using (7)
return $\hat{j}^{(*)}$

c) *Classifying:* Apply the KNN algorithm to classify the given data with the selected training samples.

The detailed procedure of kMC-KNN for driving style recognition is also shown in Algorithm 1. We shall learn a mapping between driving styles and driving data, formulated as $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \{\mathbf{x}^{(n)}\}$ is a set of all collected driving data and $\mathcal{Y} = \{y^{(n)}\}_{n=1}^{N}$ is the set of driving styles.

III. Experiments and Data Collection

A. Lane-Changing Scenarios

Among the various driver behaviors, the lane-changing maneuver occurs most frequently in real traffic [35]. Drivers should be fully aware of driving situation changes in order to make a safe decision and take correct action when changing lanes. Completing a lane change task mainly involves three stages [4]: *determining* whether a lane-changing maneuver is desirable, *selecting* the intervehicle traffic gap and initiation time, and *planning* the longitudinal and lateral trajectories. The dynamic environment with the surrounding vehicles is one of the main factors that influences a driver's decision making, including chance determination and selection for lane changes. Accordingly, in order to show the efficiency of our proposed method in classifying and recognizing drivers' decision-making styles,

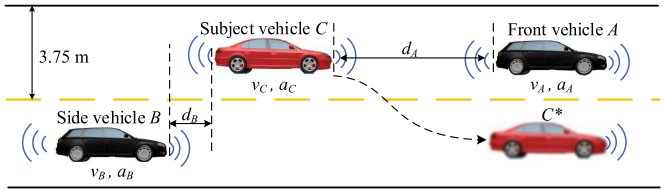


Fig. 3. Specified lane-changing scenarios.

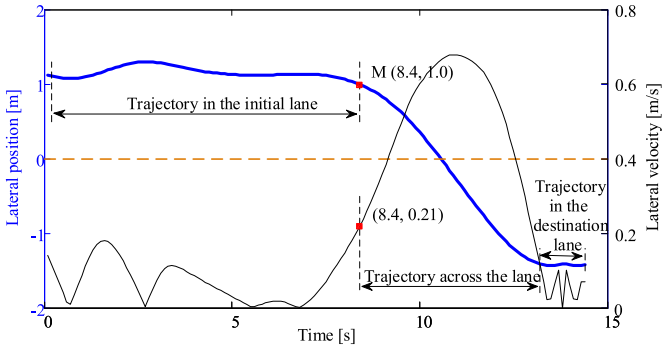


Fig. 4. Example of lane-changing behavior.

we conduct and analyze a typical lane-changing scenario with three vehicles involved, as shown in Fig. 3.

In the driving scenario, the surrounding vehicles A and B drive straight at a speed of 40–60 km/h over a distance ($d_A + d_B$) of 20–40 m, and then, the subject in vehicle C changes lanes between vehicles A and B. A dynamic traffic environment is designed, allowing vehicles A and B to accelerate and decelerate to maintain a distance of around 30 m. When the distance is greater than 30 m, the front vehicle A will brake slowly; meanwhile, the side vehicle B will accelerate. A complete lane change procedure is achieved when the driver steers the vehicle from the left lane to the center of the right lane, as shown from vehicle C to the position of vehicle C^* in Fig. 3. All the involved vehicles drive on a two-lane motorway with a sufficient length to ensure that the driver can complete the lane-changing task. The lane width is set to 3.75 m, according to the Chinese national standard. All vehicles are equipped with a vehicle-to-vehicle (V2V) capability in our simulation environment. The distance d_A (or d_B) between vehicle A (or B) and the subject vehicle C is recorded through V2V communication. The subject vehicle C also receives the speeds (denoted as v_A and v_B) and accelerations (denoted as a_A and a_B) of the surrounding vehicles A and B.

B. Feature Selection

Feature selection is very important for driving style classification and recognition, which should allow pattern vectors to belong to different categories, so that they occupy compact and disjoint regions as much as possible in a specified feature space [36]. From a geometric point of view, lane-changing behavior is a modification in the lateral position of a vehicle relative to the current-driving lanes and can be divided into three segments (see Fig. 4) [37]: a straight trajectory in the initial lane, a trajectory across the line, and a trajectory in the destination lane. In the first segment, the driver continues to observe the position, speed, and acceleration of the front vehicle A and the side vehicle B and then decides whether or not to change lanes.

Human drivers have different decision-making thresholds regarding when and whether to change lanes, which are essentially influenced by their surroundings, perceptible relative changes in the environment, and their internal models [38], [39]. Therefore, a relative change in information was selected to characterize the driver's decision to change lanes [40], including the distance between the front vehicle and the subject vehicle (d_A), the distance between the side vehicle and the subject vehicle (d_B), the speed difference between the front vehicle and the subject vehicle ($v_{AC} = v_A - v_C$), and the speed difference between the side vehicle and the subject vehicle ($v_{BC} = v_B - v_C$). Additionally, drivers also prefer different levels of acceleration and deceleration when changing lanes [41].

According to the above discussions, we select three relative pieces of information as feature parameters $\mathbf{x} = (\Delta d, \Delta v, \Delta a)$ to characterize the driver's decision-making style during the lane-changing procedure, including the distance difference Δd between d_A and d_B, the relative speed difference Δv between v_{AC} and v_{BC}, and the relative acceleration difference Δa between $a_{AC} = a_A - a_C$ and $a_{BC} = a_B - a_C$, as discussed as follows.

1) *Relative distance difference* ($\Delta d = |d_A - d_B|$)*:* A greater distance difference indicates that the subject vehicle is closer to the front vehicle or the side vehicle. Drivers who prefer a greater distance difference are more likely to make aggressive decisions when changing lanes.
2) *Relative speed difference* ($\Delta v = ||v_{AC}| - |v_{BC}||$)*:* A higher relative speed difference indicates that the subject vehicle is approaching the front vehicle or side vehicle with a higher speed. Drivers who prefer a large speed difference when changing lanes would be aggressive.
3) *Relative acceleration difference* ($\Delta a = ||a_{AC}| - |a_{BC}||$)*:* A larger acceleration difference indicates a more dangerous situation. Drivers who prefer a large acceleration would be treated as aggressive.

In order to intuitively understand the relationship between the selected features and the driving style, we visualize different typical cases of Δv when $v_A < v_B$ in Fig. 5.

1) If $v_C = (v_B + v_A)/2$ [see Fig. 5(a)], then $|v_{BC}| = |v_{AC}|$ and $\Delta v = 0$, which indicates that the front and side vehicles are both approaching the subject vehicle C equally.
2) If $v_C < (v_B + v_A)/2$ [see Fig. 5(b)], then $|v_{BC}| > |v_{AC}|$ and $\Delta v = |v_{BC}| - |v_{AC}|$, which indicates that the side vehicle B is approaching the subject vehicle C faster than the front vehicle A. Therefore, drivers with $\Delta v = |v_{BC}| - |v_{AC}|$ drive more aggressively than with $\Delta v = 0$. Analogously, the case with $\Delta v = |v_{AC}| - |v_{BC}|$ [see Fig. 5(c)] is also more dangerous than the case with $\Delta v = 0$, indicating that the driver behaves more aggressively.
3) If $|v_C| < |v_A|$ [see Fig. 5(d)], we have $|v_{BC}| \gg |v_{AC}|$ and $\Delta v = v_B - v_A$, which indicates that the side vehicle B is approaching the subject vehicle C faster than in the case of $\Delta v = |v_{BC}| - |v_{AC}|$. Therefore, drivers with $|v_C| < |v_A|$ drive more aggressively than with $v_C <$

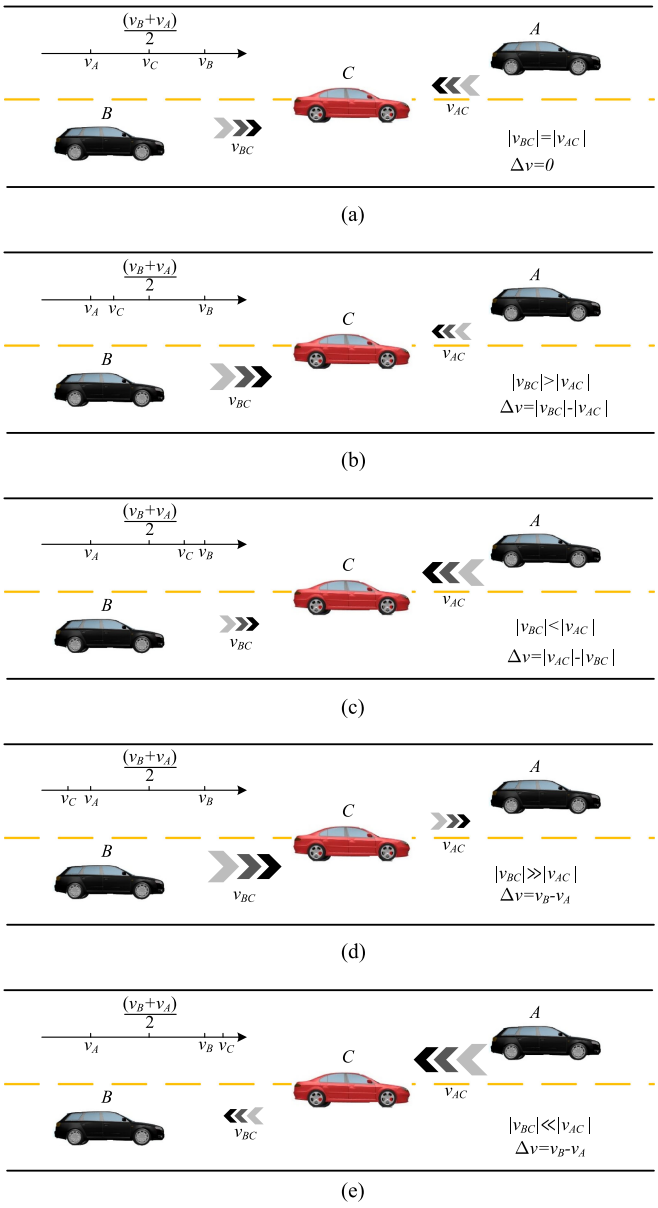


Fig. 5. Illustration of the relative speed difference when $v_A < v_B$. (a) $v_C = (v_B + v_A)/2$. (b) $v_C < (v_B + v_A)/2$. (c) $v_C > (v_B + v_A)/2$. (d) $v_C < v_A$. (e) $v_C > v_B$.

$(v_B + v_A)/2$. Analogously, the case with $|v_C| > |v_B|$ [see Fig. 5(e)] is also more dangerous than the case with $v_C > (v_B + v_A)/2$, indicating the driver is more aggressive.

The principles of the relative distance difference and the relative acceleration difference can be interpreted in the same way as in the case of the relative speed difference.

C. Driving Simulator and Data Collection

The training and testing data were collected in a driving simulator (see Fig. 6). The driving simulator consists of four main parts: a human driver, operation input equipment, a vehicle dynamics model, and the virtual environment display. Custom-built driving peripherals, including the steering wheel, brake/acceleration/clutch pedals, and gear shift handle, were utilized to collect the driver's operating signals, such as the steering wheel angle, brake pedal position, and throttle opening. A bicycle-vehicle dynamics model was built using MATLAB/Simulink. Virtual scenarios, including the vehicle, roads, and traffic facilities, were designed through the PreScan software.

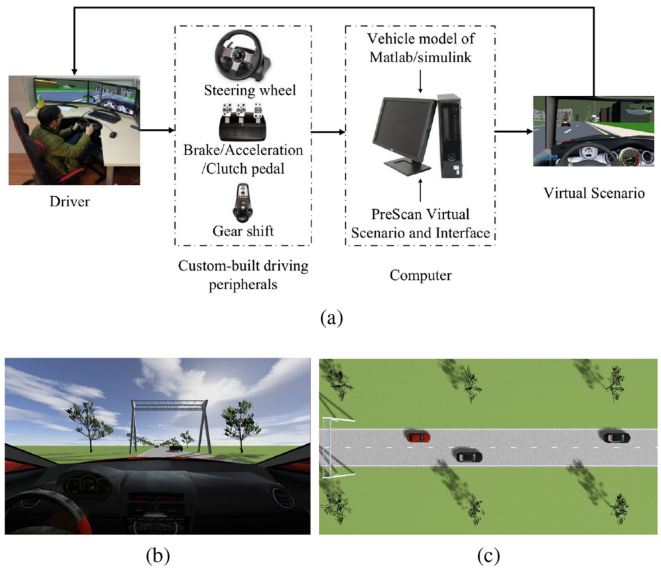


Fig. 6. (a) Schematic diagram of the driving simulator. (b) Driver view of scenario *A* in PreScan. (c) Top view of scenario *A* in PreScan.

In total, 16 subjects (12 males and 4 females) participated in our experiment as volunteers, with a minimum of 22 years in age and a maximum of 28 years in age. All of the participants held a driver license for a minimum of two years. Each driver executed 25 trials of changing lanes repeatedly. Each driver was familiarized with the test course and the driving simulator before the trials. During the trials, all the drivers followed the rules: secondary tasks, such as talking with others, making or answering telephone calls were forbidden; each participant rested 2 min before the next trial; all participants were in mentally and physically normal states; and all participants manipulated the subject vehicle in their own driving style without guidance.

All the collected data were time-series data. Therefore, we should define the specific decision-making moment of a lane change to obtain high-quality training and testing data. This moment is defined as when the lateral velocity of the host vehicle C is at least 0.21 m/s, signifying the start of a discretionary lane change [42], as illustrated by the point M in Fig. 4. Thus, the driving data at that moment were extracted as feature data to characterize the driver's decision-making style in lane-changing scenarios. Fig. 7(a) shows the extracted experimental data of point M. Fig. 7(b)–(d) shows the distributions of different features. We can see that 1) the relative acceleration difference is not strictly subject to a uniform distribution, and most data points gather around 0.05 or 0.1 m/s^2; 2) the data samples of the relative velocity difference approximately fall in [0, 1.2] m/s, and only a few data samples are greater than 1.2 m/s; and 3)

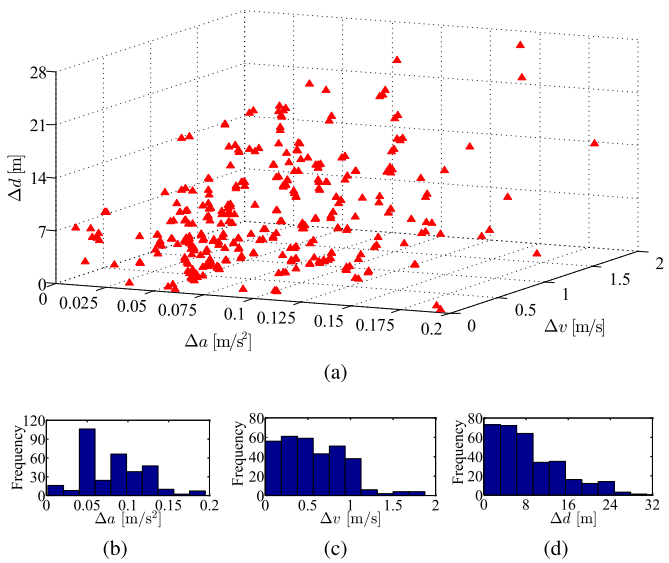


Fig. 7. Collected driving data from all drivers and their distribution.

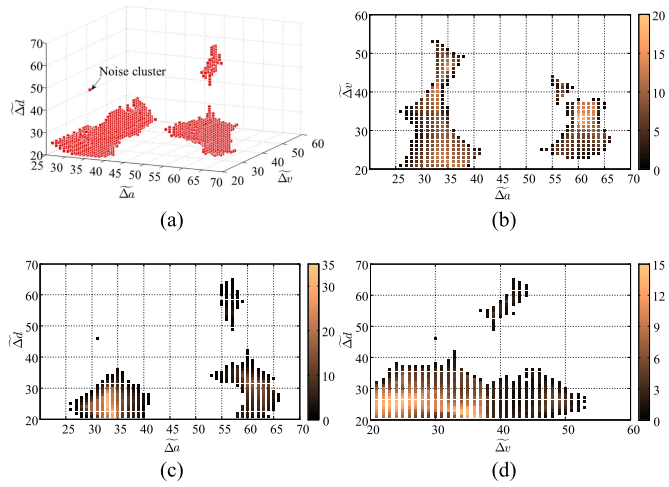


Fig. 8. Clustering results of our proposed mathematical-morphology-based method with the dilation and erosion procedures.

the data of the relative distance difference fall in the range of [0, 32] m.

IV. Analysis and Evaluation of the Experimental Result

This section analyzes and evaluates the experimental results of our developed kMC-KNN method by a comparison with traditional methods, including the KNN and SVM methods.

A. Analysis and Evaluation of the Clustering Result

1) Analysis: The collected data were finally clustered into three decision-making groups and one noise group (see Fig. 8) using the mathematical-morphology-based clustering method with $q_{\Delta d} = q_{\Delta v} = q_{\Delta a} = 100$ and $r = 10$. The centers $(C_{\mathrm{mod}}, C_{\mathrm{vag}}, C_{\mathrm{agg}})$ and ranges of each cluster for each decision-making style are shown in Table I. The raw driving data are assigned to these driving styles according to (5), as shown in Fig. 9. All the training data are automatically labeled using the mathematical-morphology-based approach with little effort and little subjective interference.

The points with different shapes in Fig. 9 represent different driving styles. The blue crosses represent drivers who prefer a low relative speed difference ($\leq$1 m/s), a small relative acceleration difference ($\leq$0.075 m/s^2), and a narrow relative distance difference ($\leq$10 m) when making a lane-change decision. We tag these kinds of drivers as *moderate* in style in their decision making. When the relative acceleration difference reaches a certain threshold (0.08 m/s^2), the moderate driver has less of a preference to change lanes than the other two types of drivers, indicating that a moderate driver is inclined to make a more conservative lane change. In addition, the moderate driver rarely drives the vehicle with a relative distance difference of more than 10 m.

The red squares represent drivers who prefer a large relative distance difference ($\geq$10 m) in most cases, covering only a few points with a small relative distance difference. This kind of driver is categorized as *aggressive*. When $\Delta d \geq 10$ m, the aggressive driver prefers a large relative acceleration difference ($\geq$0.05 m/s^2), which indicates that the aggressive driver prefers risky lane-changing maneuvers. In addition, the relative speed difference of the aggressive driver is in a large range of [0, 2] m/s.

The purple triangles represent drivers who prefer to change lanes with a relative acceleration difference in the range of [0.08, 0.2] m/s^2. We categorize these kinds of drivers as *vague*. When the relative acceleration difference is $\Delta a \in [0.08, 0.2]$ m/s^2, the vague driver prefers a small relative speed difference ($\leq$1 m/s) and a small relative distance difference ($\leq$10 m). When $\Delta d \leq 10$ m, the vague drivers prefer a larger relative acceleration difference than a moderate driver. When $\Delta a \in [0.08, 0.2]$ m/s^2, the vague driver prefers a smaller relative distance difference than an aggressive driver.

Comparing the centers of the three driving styles in Table I, it can be concluded that for each variable, a moderate driver obtains a smaller value than an aggressive driver. For vague drivers, only the relative distance difference is smaller than that of aggressive drivers, and only the relative acceleration difference is larger than that of moderate drivers.

2) Evaluation: To demonstrate the correctness of our proposed method, we compare the method with agglomerative hierarchical clustering (AHC), which is an important and well-established technique in unsupervised machine learning. AHC starts from a partition of the dataset into singleton nodes and merges the current pair of mutually closest nodes into a new node step-by-step until there is one final node left, which comprises the entire dataset.

Fig. 10 presents the results of the final four clusters using AHC. We can see that the training data are classified into three main clusters and one noise cluster. The center and range of each cluster are shown in Table II. Comparing the clustered centers in Table I with those in Table II, we can conclude that clusters

TABLE I
CLUSTERING CENTERS AND RANGES OF EACH DRIVING STYLE USING OUR MATHEMATICAL-MORPHOLOGY-BASED CLUSTERING METHOD

Driving style	Cluster centers	Ranges
Moderate driver	$C_{\text{mod}} = (0.0470, 0.4779, 4.6597)$	(0.0020∼0.0831, 0.0036∼1.1273, 0.0090∼12.4460)
Vague driver	$C_{\text{vag}} = (0.1153, 0.5335, 4.4702)$	(0.0850∼0.1951, 0.0360∼1.2958, 0.0090∼11.5193)
Aggressive diver	$C_{\text{agg}} = (0.1012, 0.6962, 15.727)$	(0.0444∼0.1854, 0.0223∼1.8764, 6.2681∼30.9796)

TABLE II
CENTERS AND RANGES OF EACH CLUSTER USING THE AHC METHOD

Cluster #	Centers	Ranges
Cluster 1	$C'_{\text{mod}} = (0.0478, 0.4837, 4.9682)$	(0.0020∼0.0767, 0.0123∼1.1284, 0.0903∼13.0704)
Cluster 2	$C'_{\text{vag}} = (0.1116, 0.6091, 5.0022)$	(0.0847∼0.1648, 0.0360∼1.3468, 0.2208∼11.7478)
Cluster 3	$C'_{\text{agg}} = (0.1002, 0.5437, 16.6271)$	(0.0446∼0.1516, 0.0266∼1.0668, 10.8425∼25.0622)

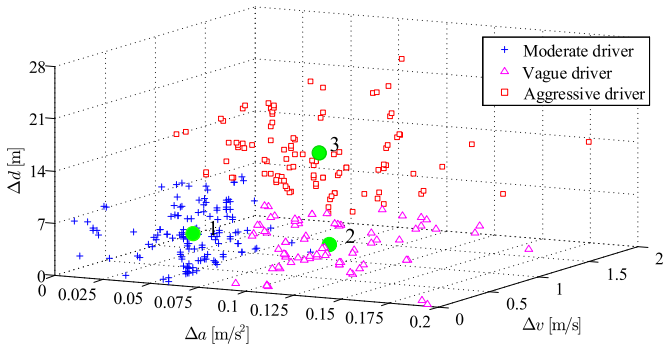


Fig. 9. Clustering results for the original data.

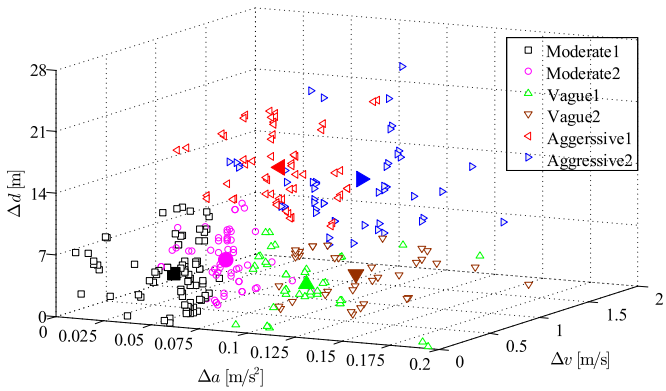


Fig. 11. Clustering results using k-MC.

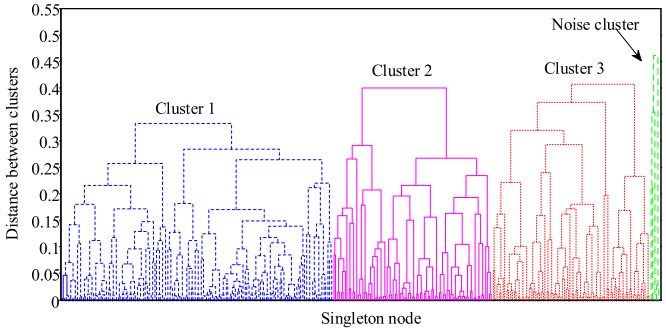


Fig. 10. Clustering results using the AHC method.

1–3 in Fig. 10 are associated with the moderate, vague, and aggressive decision-making styles in Fig. 9, respectively.

B. Recognition Performance Analysis and Evaluation

For the developed kMC-KNN recognition method, k-MC is used to partition the raw data in each driving style (J) into k subsets; thus, the raw data are divided into $J \times k$ subsets. For example, the driving data in each diving style ($J = 3$) were divided into k clusters ($k = 2$), as shown in Fig. 11. Given the test data $x^{(*)}$, kMC-KNN chooses one cluster from the two clusters based on the similarities between $x^{(*)}$ and the centers in each driving style as the training samples to reduce the computational cost of the KNN method.

1) Evaluation Metrics: In order to evaluate the recognition performance of kMC-KNN, the cross-validation procedure was utilized. For p-fold cross-validation, the original datasets were randomly and evenly partitioned into p folds. One single fold was retained as validation data for testing the model, and the remaining $p - 1$ folds were used as training data. The cross-validation process was then repeated p times, with each of the p folds being used exactly once as validation data. Totally, p results from all folds then were averaged (or otherwise combined) to obtain a single estimation result. Here, we randomly partitioned the original driving data ($N = 9936$) into $p = 4$ folds ($N_p = 2484$) to evaluate the performance of kMC-KNN. Then, the average accuracy was taken as the final results. The accuracy of the driving-pattern recognizer is computed by

$$\lambda_j = \frac{K_{\text{cor},j}}{\sum K_{\text{all},j}}, \quad j = \text{mod}, \text{vag}, \text{agg} \tag{12}$$

where λ_j is the accuracy of the j driving style. $K_{\text{cor},j}$ is the number of clustering points that are correctly recognized as the j driving style. $K_{\text{all},j}$ is the number of clustering points in the j driving style.

To show the time-saving performance of kMC-KNN, we conducted offline tests of kMC-KNN with different number

TABLE III
COMPARISON RESULTS FOR THE KNN, SVM, AND kMC-KNN METHODS

$K=\sqrt{N_p}$		KNN	kMC-KNN			SVM
			$k=2$	$k=3$	$k=4$	
Accuracy	λ_{mod}	99.06% ($^{+0.85\%}_{-0.58\%}$)	96.45% ($^{+2.73\%}_{-5.74\%}$)	96.89% ($^{+2.11\%}_{-1.12\%}$)	97.11% ($^{+1.30\%}_{-1.50\%}$)	91.45% ($^{+8.55\%}_{-16.69\%}$)
	λ_{vag}	97.08% ($^{+2.21\%}_{-2.91\%}$)	98.26% ($^{+0.89\%}_{-1.36\%}$)	95.25% ($^{+3.75\%}_{-4.30\%}$)	97.59% ($^{+0.99\%}_{-1.93\%}$)	87.42% ($^{+12.44\%}_{-11.30\%}$)
	λ_{agg}	94.93% ($^{+3.60\%}_{-6.25\%}$)	91.90% ($^{+4.51\%}_{-8.27\%}$)	90.39% ($^{+5.06\%}_{-9.22\%}$)	91.16% ($^{+5.49\%}_{-6.52\%}$)	86.55% ($^{+13.45\%}_{-12.90\%}$)
T [s]		852.75 ($^{+3.33}_{-1.36}$)	233.08 ($^{+10.47}_{-20.92}$)	131.92 ($^{+15.52}_{-10.60}$)	91.84 ($^{+10.09}_{-6.98}$)	—-
T_0 [ms]		343.30 ($^{+1.34}_{-0.56}$)	93.83 ($^{+4.25}_{-8.42}$)	53.11 ($^{+6.25}_{-4.27}$)	36.97 ($^{+4.06}_{-2.81}$)	—-

of clusters clustered by k-MC. The traditional KNN and SVM methods were chosen for comparative studies, and the same parameter ($K=\sqrt{N_p}$) was selected for both the KNN and kMC-KNN methods, including training data and testing data. The test results using kMC-KNN with $k=2,3,4$ and the KNN method are shown in Table III. T is the recognition time for the kMC-KNN and KNN methods, and T_0 is the recognition time of one data point.

2) Analysis Results: From Table III, it can be seen that the developed kMC-KNN outperforms the KNN method by significantly reducing the recognition time by more than 72.67%. With an increasing value of k, the recognition time of kMC-KNN gradually decreases since the computation load decreases. The accuracy of kMC-KNN for a vague driving style also outperforms the KNN algorithm; however, the accuracy for both aggressive and moderate driving styles is slightly lower than that of the KNN algorithm. The accuracy of kMC-KNN fluctuates slightly with increasing k.

To demonstrate the recognition performance of our proposed method, we also compare our method with an SVM, as shown in Table III. We found that kMC-KNN obtains a better performance than the SVM. More specifically, the SVM obtains an average recognition accuracy of 87.42% for a vague driving style, while kMC-KNN with $k=2$ achieves an accuracy of 98.26%. Additionally, the deviations in the recognition accuracy also demonstrate that kMC-KNN is more robust than the SVM. For example, the SVM obtains an average accuracy for a vague driving style varying from 76.12% to 99.86%, while kMC-KNN with $k=2$ achieves a more stable performance, varying from 96.90% to 99.15%.

V. CONCLUSION

This article developed a kMC-KNN method in order to improve recognition efficiency. An unsupervised clustering method was also proposed based on mathematical morphology in order to reduce the efforts associated with labeling training data and to exclude subjective interference from humans. The mathematical-morphology-based clustering method can classify drivers' decision-making styles of lane-changing behavior into three categories with little labeling effort. The experiment results show that our proposed kMC-KNN method can shorten the recognition time greatly without degrading the recognition accuracy. We also found that the developed kMC-KNN method outperforms the SVM method in terms of recognition accuracy and stability.

REFERENCES

[1] C. M. Martinez, M. Heucke, F.-Y. Wang, B. Gao, and D. Cao, "Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 666–676, Mar. 2018.

[2] X. Li and J.-Q. Sun, "Studies of vehicle lane-changing dynamics and its effect on traffic efficiency, safety and environmental impact," *Phys. A: Statist. Mech. Appl.*, vol. 467, pp. 41–58, 2017.

[3] J. Nilsson, M. Brännström, E. Coelingh, and J. Fredriksson, "Lane change maneuvers for automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1087–1096, May 2017.

[4] J. Nilsson, J. Silvlin, M. Brannstrom, E. Coelingh, and J. Fredriksson, "If, when, and how to perform lane change maneuvers on highways," *IEEE Intell. Transp. Syst. Mag.*, vol. 8, no. 4, pp. 68–78, Winter 2016.

[5] H. Zhou and M. Itoh, "How does a driver perceive risk when making decision of lane-changing?" *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 60–65, 2016.

[6] S. Schnelle, J. Wang, H.-J. Su, and R. Jagacinski, "A personalizable driver steering model capable of predicting driver behaviors in vehicle collision avoidance maneuvers," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 5, pp. 625–635, Oct. 2017.

[7] W. Wang, J. Xi, C. Liu, and X. Li, "Human-centered feed-forward control of a vehicle steering system based on a driver's path-following characteristics," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1440–1453, Jun. 2017.

[8] M. V. Ly, S. Martin, and M. M. Trivedi, "Driver classification and driving style recognition using inertial sensors," in *Proc. Intell. Vehicles Symp.*, 2013, pp. 1040–1045.

[9] A. Aljaafreh, N. Alshabatat, and M. S. N. Al-Din, "Driving style recognition using fuzzy logic," in *Proc. IEEE Int. Conf. Veh. Electron. Safety*, 2012, pp. 460–463.

[10] T. Pan, W. H. Lam, A. Sumalee, and R. Zhong, "Modeling the impacts of mandatory and discretionary lane-changing maneuvers," *Transp. Res. C: Emer. Technol.*, vol. 68, pp. 403–424, 2016.

[11] M. Keyvan-Ekbatani, V. L. Knoop, and W. Daamen, "Categorization of the lane change decision process on freeways," *Transp. Res. C: Emer. Technol.*, vol. 69, pp. 515–526, 2016.

[12] E. Balal, R. L. Cheu, and T. Sarkodie-Gyan, "A binary decision model for discretionary lane changing move based on fuzzy inference system," *Transp. Res. C: Emer. Technol.*, vol. 67, pp. 47–61, 2016.

[13] D. Sun and L. Elefteriadou, "Lane-changing behavior on urban streets: An in-vehicle field experiment-based study," *Comput.-Aided Civ. Inf. Eng.*, vol. 27, no. 7, pp. 525–542, 2012.

[14] D. Sun and L. Elefteriadou, "Research and implementation of lane-changing model based on driver behavior," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2161, no. 1, pp. 1–10, 2010.

[15] D. Sun and L. Elefteriadou, "A driver behavior-based lane-changing model for urban arterial streets," *Transp. Sci.*, vol. 48, no. 2, pp. 184–205, 2014.

[16] M. Ishibashi, M. Okuwa, S. Doi, and M. Akamatsu, "Indices for characterizing driving style and their relevance to car following behavior," in *Proc. Soc. Instrum. Control Eng. Conf.*, 2008, pp. 1132–1137.

[17] W. Wang, J. Xi, A. Chong, and L. Li, "Driving style classification using a semi-supervised support vector machine," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 5, pp. 650–660, Oct. 2017.

[18] Y. Lei, K. Liu, Y. Fu, X. Li, Z. Liu, and S. Sun, "Research on driving style recognition method based on drivers dynamic demand," *Adv. Mech. Eng.*, vol. 8, no. 9, pp. 1–14, 2016.

[19] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proc. Privacy Enhancing Technol.*, vol. 2016, no. 1, pp. 34–50, 2016.

[20] W. Wang and J. Xi, "A rapid pattern-recognition method for driving styles using clustering-based support vector machines," in *Proc. Amer. Control Conf.*, Jul. 2016, pp. 5270–5275.
[21] W. Han, W. Wang, X. Li, and J. Xi, "Statistical-based approach for driving style recognition using Bayesian probability with kernel density estimation," *IET Intell. Transp. Syst.*, vol. 13, no. 1, pp. 22–30, 2019.
[22] Y. Zhang, W. C. Lin, and Y. K. S. Chin, "A pattern-recognition approach for driving skill characterization," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 905–916, Dec. 2010.
[23] W. Wang, J. Xi, and H. Chen, "Modeling and recognizing driver behavior based on driving data: A survey," *Math. Probl. Eng.*, vol. 2014, 2014, Art. no. 245641.
[24] W. Wang, J. Xi, and D. Zhao, "Driving style analysis using primitive driving patterns with Bayesian nonparametric approaches," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2986–2998, Aug. 2019.
[25] T. Appenzeller, "The scientists' apprentice," *Science*, vol. 357, no. 6346, pp. 16–17, 2017.
[26] W. Wang, C. Liu, and D. Zhao, "How much data are enough? A statistical approach with case study on longitudinal driving behavior," *IEEE Trans. Intell. Veh.*, vol. 2, no. 2, pp. 85–98, Jun. 2017.
[27] J. Serra and P. Soille, *Mathematical Morphology and Its Applications to Image Processing*, vol. 2. Berlin, Germany: Springer, 2012.
[28] Management Association, Information Resources, *Ophthalmology: Breakthroughs in Research and Practice*. Hershey, PA, USA: IGI Global, 2018. [Online]. Available: https://books.google.com.hk/books?id=eblDDwAAQBAJ
[29] M. E. Valle and R. A. Valente, "Mathematical morphology on the spherical cielab quantale with an application in color image boundary detection," *J. Math. Imag. Vis.*, vol. 57, pp. 1–19, 2017.
[30] G. Agam and I. Dinstein, "Regulated morphological operations," *Pattern Recognit.*, vol. 32, no. 6, pp. 947–971, 1999.
[31] H. L. Luo, F. S. Kong, X. B. Yang, and B. H. Liu, "Cluster analysis based on mathematical morphology," *Pattern Recognit. Artif. Intell.*, vol. 19, pp. 727–733, 2006.
[32] H. L. Luo, F. University, and Shanghai, "Clustering ensemble algorithm based on mathematical morphology," *Comput. Sci.*, vol. 37, no. 8, pp. 214–218, 2010.
[33] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[34] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," *Lecture Notes Comput. Sci.*, vol. 2888, pp. 986–996, 2003.
[35] J. D. Chovan, L. Tijerina, G. Alexander, and D. L. Hendricks, "Examination of lane change crashes and potential IVHS countermeasures," Nat. Highway Traffic Saf. Admin., Tech. Rep. HS-808 071, Mar. 1994.
[36] R. A. Dunne, *A Statistical Approach to Neural Networks for Pattern Recognition* (Wiley Series in Computational Statistics). Hoboken, NJ, USA: Wiley-Interscience, 2007.
[37] R. D. Worrall and A. G. R. Bullen, "An empirical analysis of lane changing on multilane highways," Highway Research Board, Highway Research Rec. no. 303, 1970.
[38] G. Asaithambi and G. Shravani, "Overtaking behaviour of vehicles on undivided roads in non-lane based mixed traffic conditions," *J. Traffic Transp. Eng.*, vol. 4, no. 3, pp. 252–261, Jun. 2017.
[39] Y. Hou, P. Edara, and C. Sun, "A genetic fuzzy system for modeling mandatory lane changing," in *Proc. Int. IEEE Conf. Intell. Transp. Syst.*, 2012, pp. 1044–1048.
[40] Y. Hou, P. Edara, and C. Sun, "Modeling mandatory lane changing using Bayes classifier and decision trees," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 647–655, Apr. 2014.
[41] S. Moridpour, M. Sarvi, and G. Rose, "Modelling lane changing behaviour of heavy commercial vehicles," in *Proc. 30th Australas. Transp. Res. Forum*, 2007, pp. 1–15.
[42] J. Nie, J. Zhang, X. Wan, W. Ding, and B. Ran, "Modeling of decision-making behavior for discretionary lane-changing execution," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2016, pp. 707–712.

Electroencephalographic Phase–Amplitude Coupling in Simulated Driving With Varying Modality-Specific Attentional Demand

Ernesto Gonzalez-Trejo, Hannes Mögele, Norbert Pfleger, Ronny Hannemann, and Daniel J. Strauss

***Abstract*—The quantification of attention during driving can help identify situations in which the driver is not completely aware of the situation. By using the principle of phase–amplitude coupling (PAC) in electroencephalographic (EEG) signals, we aimed to test if PAC might be eligible as a biomarker of attention in multimodal tasks such as driving. Surface EEG was measured simultaneously in drivers and copilots while participating in simulated driving scenarios with varying multimodal attentional demands. The PAC between Theta-band phase and Gamma-band amplitude from the EEG was obtained and evaluated. Results showed significant PAC differences between drivers and copilots in areas related to multimodal attention (prefrontal cortex, frontal eye fields, primary motor cortex, and visual cortex). The results were confirmed by behavioral data acquired during the test (detection task). We conclude that PAC does function as a biomarker for attentional demand by detecting cortical areas being activated through specific multimodal (in this case, driving) tasks.**

***Index Terms*—Attention, driving, electroencephalography (EEG), phase–amplitude coupling (PAC).**

I. Introduction

DRIVING is a complex cognitive task, involving audiovisual cues, mechanical coordination, decision making, and information retention, among others, which have to be processed simultaneously [1]. Attention during driving is critical; even a slight distraction can lead to a dangerous situation. Current technological advances, such as navigation systems, hands-free systems, smart cockpits, and smartphone interfaces, aim to support driving tasks, but might actually act as a distraction [2]–[5]. Moreover, these interfaces might overlap additional distraction sources, such as passengers, fatigue effects, weather and/or traffic conditions, all inherent to driving [6]–[8].

The current challenge for physicians and researchers lies in the assessment of perceptual and cognitive workload and attention during a driving situation. Both functional magnetic resonance imaging (fMRI) [1], [9]–[12] and magnetoencephalography (MEG) [13]–[15] have been able to describe biomarkers such as Theta-band power increase and decrease correlated to active and passive driving [15], occipital brain activation during driving [1], [9], correlation of parietal–occipital and frontal lobe activation with steering operations [10], and decrease of parietal activity when distractions appear while driving (fMRI: [11], [12], MEG: [13], [14]); they have also shown a decrease in driving performance when engaged in tasks such as language comprehension [11] and conversation [13], which suggests a difference between external (driving task) and internal (cognitive task) attentional effort. However, these biomarkers have only been observed in a virtual environment due to the equipment required in order to acquire either fMRI or MEG. Portable solutions to study brain activity include electroencephalography (EEG) [16]–[22] and functional near infrared spectroscopy (fNIRS) [23]–[25]. These noninvasive methods allow for an easier setup and measurement of neural activity (e.g., using band power as a biomarker for attention) both in virtual and in real driving situations; although the information obtained is limited by the constraints of measuring neural activity at the scalp, both techniques provide a way to confirm results from imaging studies and allow for more flexible situations and novel study designs, which may even combine both EEG and fNIRS [26].

The analysis of EEG frequency band power has been commonly reported in the literature as a biomarker of neural activity related to attention while driving [22], [27]. For example, Gamma activity has been shown to correlate with sensory processing and both attentional and memory tasks [28]–[31], whereas Theta has been related with error processing, working memory, and information encoding [32]–[35], and Delta/Beta power spectra have been related to driving fatigue [22]. However, the reliability of band power alone has been questioned as it might reflect decreases in attention due to monotony, and not due to attentional resources being assigned to different cognitive tasks, as shown by [21]; their results suggest that the synchronization of oscillatory activity in the brain might offer a sounder alternative to quantify attention and cognitive effort.

Manuscript received June 13, 2018; revised December 28, 2018 and April 25, 2019; accepted June 30, 2019. Date of publication September 5, 2019; date of current version November 21, 2019. This work was supported in part by the Federal Ministry of Education and Research (BMBF) Germany under Grant 03FH036I3 and in part by the German Research Foundation (DFG) under Grant STR 994/1-1. This article was recommended by Associate Editor Y. Li. *(Corresponding author: Daniel J. Strauss.)*

E. Gonzalez-Trejo and D. J. Strauss are with the Systems Neuroscience and Neurotechnology Unit, Saarland University and htw saar, 66421 Homburg/Saar, Germany (e-mail: ernesto.gonzaleztrejo@uni-saarland.de; daniel.strauss@uni-saarland.de).

H. Mögele is with the Audi Electronics Venture GmbH, 85080 Gaimersheim, Germany (e-mail: hannes.moegele@audi.de).

N. Pfleger is with the paragon semvox GmbH, 66459 Kirkel-Limbach, Germany (e-mail: pfleger@semvox.de).

R. Hannemann is with Sivantos GmbH, 91058 Erlangen, Germany (e-mail: ronny.hannemann@sivantos.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2931011

Neuronal oscillations are the base of cognitive activity [36], [37], but the synchronization between different frequency bands is still being studied [38]. Cross-frequency coupling has received a lot of attention in recent years, by proving to be a biomarker of cognitive processing in humans [39]–[41]. Especially Theta–Gamma band coupling [42], [43], which has been related to sensory integration, working memory [44], [45], and visual perception [46], [47], which are essential for driving tasks [19]. Tort *et al.* proposed a new form of cross-frequency coupling, called phase–amplitude coupling (PAC) [42]. This coupling adapts the Kullback–Leiber (KL) distance measure [48] to calculate the deviation between the uniform distribution and an empirical amplitude distributionlike function projected over phase bins (in their original work, amplitude and phase information from *in vivo* hippocampal recordings) and assigns a modulation index to it (representing the amplitude–phase coupling intensity).

In this article, we obtained both the amplitude and phase information of surface EEG acquired during simulated driving (from both driver and copilot) and applied the framework established by Tort *et al.* to measure PAC. We aimed to test the feasibility of PAC as a biomarker of attention-related cortical activation in multimodal tasks (in our case, an audiovisual task while driving), both for active (driving) and passive (copilot) scenarios. We hypothesized that PAC would allow us to identify the cortical regions most active during attentional tasks. A behavioral task was also performed in order to confirm our main hypothesis. Moreover, we hypothesized that PAC could differentiate between auditory and visual attention based on the cortical activation observed. If confirmed, PAC, as obtained from surface EEG, could provide benefits such as portability, applicability in real-life situations, and noninvasiveness. Not only would PAC offer an alternative to analyze neural activity, but it might support all established biomarkers previously discussed as well.

II. Methods

A. Participants

Twenty two healthy subjects (three females), ages between 20 and 29 years (mean $= 24.4 \pm 2.7$ years), took part in the study. All subjects were right handed, German native speakers, and were required to have a valid driver's license (mean driving experience 6.8 ± 2.3 years); subjects were recruited from the social environment of the authors. The subjects had no hearing/visual impairments (contact lenses were allowed if they were required according to their driver's license). Before each measurement, the subjects were informed of the objectives and methodology of the measurement and asked to provide their written consent. No history of neurophysiological diseases/impairments was present in any of the subjects. The measurements were performed at the MINDSCAN Laboratory from the Systems Neuroscience and Neurotechnology Unit, Saarbruecken, Germany. The study was designed following ethic guidelines and the declaration of Helsinki; each participant was given the choice to abandon the measurement at any given time.

B. Driving Simulator

The indoor driving simulator consists of a real commercial vehicle cockpit (Audi A3, 2013) with an interface connected to a personal computer; the cockpit transfers information from the steering wheel and pedals to the simulation software (simulator interface and software from Simutech GmbH, Bremen, Germany). The software displays the virtual environment along three monitors. Driving in the simulation requires the same actions as real-life driving; the simulator is equipped with a manual, 5-gear gearbox transmission; all pedals (gas, brakes, and clutch) are functional. Speedometer and tachometer (RPM) are also connected to the software and display the speed/RPM in real time. The software simulation is able to render a small city, rural roads and/or highways, or a combination of them, according to the selected task. Pedestrians, traffic, and randomized events, such as pedestrian cross walks, wild animals, or construction sites, appear during driving and can be turned ON or OFF. Road navigation systems are also integrated (NAVI) and guide the subjects from the beginning to the end of the task through audiovisual cues (arrows displayed between speedometer and tachometer and auditory playback of spoken directions). Weather conditions (day/night, rain, snow, fog) can also be modified according to the task chosen.

The auditory output from the simulation software was rerouted to a separate personal computer, which was also used to playback additional auditory stimuli/tasks. A 24-loudspeaker array (JBL Control 1 Pro Loudspeakers, JBL, California, CA, USA) was mounted around the cockpit allowing for a directional control of auditory stimuli/distractions during driving. Each loudspeaker was controlled independently through audio-processing software (PreSonus Audio Electronics, Louisiana, LA, USA). Loudspeakers were arranged in three circular arrays, with a 45° angular resolution at floor (8 speakers, 40-cm high), ear (8 speakers, 115 cm) and over-the-head (8 speakers, 190 cm) heights. Both the audio coming from the driving simulator software as well as the audio used for auditory tasks were first converted to analog signals using a digital/analog interface (RME Digital/Analog Interface M-32 DA, RME Audio AG, Haimhausen, Germany). The loudspeakers were driven using three 8-channel audio amplifiers (Apart PA8250, Apart Audio NV, Antwerp, Belgium). Additionally, two subwoofer systems were embedded under the cockpit in order to elicit low-frequency vibration akin to the one present while driving. The subwoofers were driven using two additional audio amplifiers (the t.amp S-100 mk2, Thomann GmbH, Burgebrach, Germany; BKA 1000 N, The Guitammer Company, Ohio, OH, USA). Fig. 1 shows the layout of the driving simulator.

All audio tracks used were fine tuned and calibrated (intensity and position) to allow for a realistic spatial sound localization from the seat of the driver. This included varying amplitudes of combined speakers and cross mixing two or more speakers in order to enhance the directionality of the sound (e.g., sounds perceived as coming from the back seat or cockpit seat).

C. Driving Task

The driving task within all modalities of the study was to drive correctly (following traffic laws) around the selected environment, following the directions of the NAVI system. If a participant was seen deliberately disobeying traffic laws (e.g., speeding, driving on the wrong lane, running under red lights),

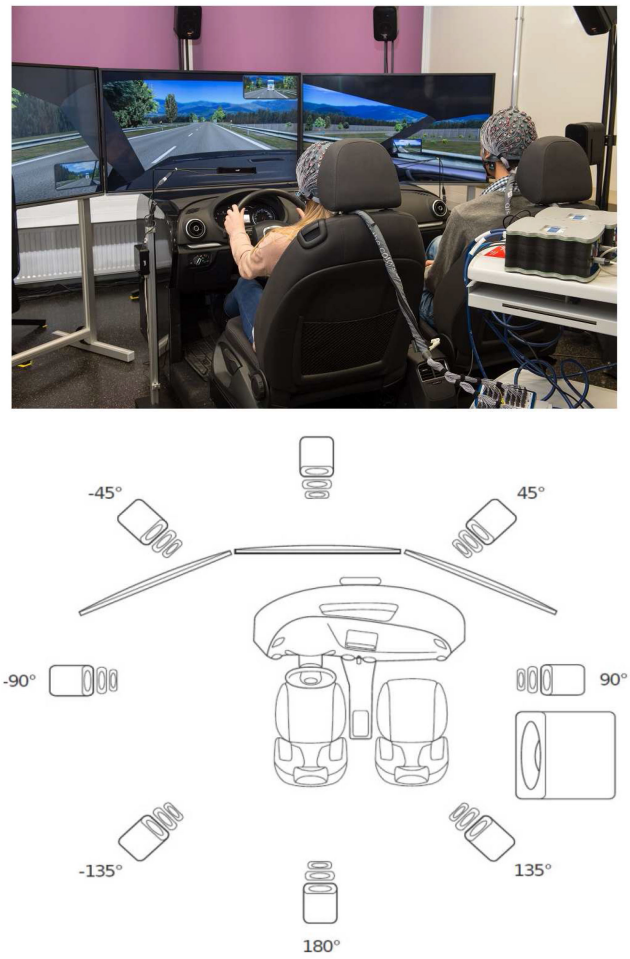


Fig. 1. Layout of driving simulator: (Top) A real automobile cockpit is used, together with software rendering a virtual environment on three monitors in front of the driver and a multidirectional speaker array, which provides auditory feedback, tasks, and distractions according to the task being presented. (Bottom) Speaker array: 24 speakers are arranged around the cockpit and allow for a flexible spatial localization of sounds. Several speakers can be linked together in order to simulate spatial localization more accurately (e.g., back seat or cockpit seat), or can be used independently. The subwoofer system under the cockpit is not shown.

the measurement was stopped and restarted after reminding the subject of the correct driving behavior expected. According to the selected task, the driving conditions changed (more/less traffic, bad weather) to increase or decrease the difficulty of the driving task. Additionally, eye-tracking hardware (Tobii Pro X2 60, Tobii AB, Sweden) was used to monitor the gaze of the subjects in real time and ensure that they were looking at the road during the measurement; both driver and copilot were reminded before each measurement to keep their eyes on the road during driving tasks.

D. Behavioral (Auditory) Task and Stimuli

The main auditory stimuli used for the behavioral tasks was a radio broadcast (in German) played while driving. According to the modality (scenario) of the study, the subjects were asked to detect a specific word within the broadcast ("und") and press a button, or do nothing and keep driving with the broadcast playback still ongoing. The word was chosen due to the number of occurrences in the selected radio broadcasts and due to being one of the most used words in German language [49]. This constituted the behavioral task of the study and was used to rate the difficulty of the driving situations; more difficult scenarios were expected to require more attention from the participants.

In order to increase the difficulty of the task, background noise was added. The main component of the background noise was the so-called "Fastl noise" [50], which simulates spectral and temporal features of human speech. This noise was implemented in the same loudspeakers as the radio broadcast. Additional auditory distracters were implemented, coming from different directions—recordings of babies crying and kids arguing were played from the loudspeakers simulating the back seats, while mobile-phone ringtones were played from the side loudspeakers to appear coming from the seat next to the driver. The sounds were evenly distributed along the task length, and special care was taken to ensure that none of the distracters overlapped the target words. All sounds were calibrated in order to attain safe hearing amplitude levels and were strictly kept under 80 dB sound pressure level (SPL). The maximum amplitude recorded (driving at maximum speed, with all auditory stimuli being played at the same time) was 72.3 dB SPL. The Fastl noise was set at 66.5 dB SPL in order to obtain a signal-to-noise ratio (SNR) between the radio broadcast and the Fastl noise of -1.5. The radio broadcast consisted of conversations between several radio presenters (no music) and the amplitude varied between 64.8 and 65.6 dB SPL (playback amplitude set to 65 dB SPL); therefore, the actual SNR measured varied between -0.9 and -1.7. This was deemed as adequate for the given tasks. The additional noises (kids, baby, and mobile phone) were all played at 64 dB SPL.

The beginning of each behavioral task, the task word, and the button presses generated nonaudible trigger signals, which were sent to the data acquisition system through a trigger box (g.Trigbox, g.Tec, Austria) for *post hoc* signal processing.

E. Task Scenarios

The three main tasks were named "A," "B," and "C," each 4 min in length, involving different degrees of auditory and visual or nonauditory attentional effort while driving. Driving in task "A" was done on a highway without traffic, without speed limit, with clear weather, and without randomized traffic events. Driving in tasks "B" and "C" started on a rural road and crossed through an urban environment; for both B and C the weather was set to rain, together with fog and a sight distance limited to 40 m ahead, forcing an increased driving effort. The tasks can be summarized as follows:

1) Task "A": Driving effort—low; auditory effort—high. Driving on the highway without traffic and with a clear weather, while solving the behavioral task (pressing a button when "und" is heard) in the presence of auditory distracters.
2) Task "B": Driving effort—high; auditory effort—low. Driving on different roads with bad weather conditions;

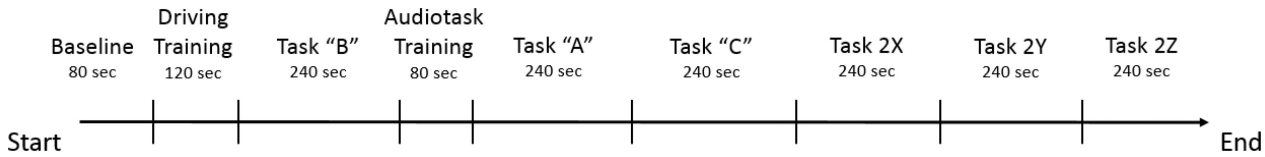


Fig. 2. Measurement plan. A baseline, two training stages, and three different tasks were performed. The three tasks (A, B, C) were repeated in a randomized order (shown here as X, Y, and Z) for retest purposes.

no auditory distracters, radio broadcast playing but button should NOT be pressed.

3) Task "C": Driving and auditory effort—high. Driving on different roads with bad weather conditions, while solving the behavioral task (pressing a button when "und" is heard) in the presence of auditory distracters.

Besides the main tasks, two training stages and an initial baseline measurement were performed.

1) Baseline: No tasks, driving simulator OFF, 80 s.
2) Driving training: Driving freely in an urban environment, 2 min.
3) Auditory training: Broadcast was presented with occurrences of the target word; participants are instructed to press the button if they detect the target word. Driving simulator OFF, 80 s.

After the main tasks were performed, tasks A, B, and C were repeated in a randomized order, as a retest to enhance the robustness of the results (referred from now on as 2A, 2B, and 2C). The scenarios and their distribution along the measurement are shown in Fig. 2.

Two subjects were measured simultaneously in each measurement (driver and copilot). Once the measurement was finished, the current driver and the copilot switched places and the measurement was restarted. Due to the subjects not being naive anymore to the measurement, the second measurement was classified differently from the first—subjects from the first measurement were classified as "round 1" (Driver and Copilot R1), and subjects from the second measurement (nonnaive) were classified as "round 2" (Driver and Copilot R2).

F. Psychophysiological Measurements

Electroencephalogram (EEG) was acquired using a high-resolution EEG system (g.HiAmp, g.Tec, Austria) with 128 active electrodes, at a sampling rate of 512 Hz. Impedances for the active electrodes were kept under 50 kΩ. No online filtering was used and all electrodes were referenced to CPz and rereferenced to the average reference postmeasurement. Additionally, two passive Ag/AgCl electrodes were placed around the left eye in order to acquire an electrooculogram (EOG), used to detect blinking.

All biosignals (EEG, EOG), together with the trigger signals from the auditory stimuli and button, and eye position tracking were acquired through the same biosignal amplifier (g.HiAmp) and controlled through a SIMULINK interface (The Mathworks, Massachusetts, MA, USA). The output file was processed using MATLAB R2013a (The Mathworks, Massachusetts, MA, USA).

G. Measurement Protocol

After the subjects read and signed the consent forms, the Ag/AgCl electrodes for EOG were placed. Afterward, the subjects were asked to sit in the driving simulator and the EEG cap was placed on their heads. The subjects received a button for solving the behavioral task; in the case of the driver, the button was fixed to the side of the index finger of the right hand. This position allowed the driver to press the button with the thumb without leaving the steering wheel.

Once all sensors were in place, the subjects were informed of the course of the measurement. They were not informed of the auditory task until the auditory training took place. Between each task, there was a small pause (approximately 1 min) where the subjects were asked if the measurement may continue. If so, the operator informed them of the upcoming task and checked that all sensors were still working correctly (e.g., impedances from the EEG and EOG electrodes remain within acceptable ranges).

At the end of the measurement, the subjects answered a small questionnaire regarding their perception of the difficulty of the tasks, their performance, the level of distraction attained through the auditory stimuli, and the quality of the simulation overall. As a safety measurement, subjects were not allowed to drive their own vehicle 30 min postmeasurement.

III. Data Processing

A. Electroencephalography

1) Signal Conditioning: The first step in the processing of the acquired (raw) EEG signals was the removal of the blinking artifacts by using independent component (IC) analysis (ICA)-based blind-source separation. This was implemented using the EEGLAB Toolbox (SCCN, San Diego, CA, USA) in MATLAB and the FastICA [51] algorithm. ICA was performed for the 128 EEG channels and the ICs were individually analyzed to search for recognizable blinking artifact patterns. Once found, these ICs were removed from the original signal. Care was taken to remove only the ICs corresponding to eye artifacts, even if in some cases the eye artifacts were not completely removed. This compromise was chosen in order to keep as much signal integrity while removing as much artifacts as possible. The output signal post-ICA was manually compared to the original EEG raw signal and to the EOG raw signal in order to confirm that the artifacts removed were indeed caused by blinking. After removing the eye artifacts, the EEG data were rereferenced to the average reference.

2) Phase–Amplitude Coupling: The PAC analysis was done following the framework established by Tort *et al.* [42]; what follows is a summary of the procedure.

In order to analyze the PAC, the EEG signal was split into three frequency bands—Theta (4–8 Hz), low Gamma (30–50 Hz), and high Gamma (50–80 Hz), using an FIR filter with a Hanning window in MATLAB. It has been argued that the useful spectrum of EEG that can be measured through scalp EEG is limited to 80 Hz [39], [43]. We split the Gamma-band analysis based on this assumption, and studied both conditions simultaneously. The mean of the filtered signals was subtracted afterward.

The Hilbert transform was applied to both Theta and Gamma acquired signals; the instantaneous phase was then extracted from the Hilbert transform of the Theta-filtered signal and the amplitude envelope was extracted from the Hilbert transform of the Gamma-filtered signal. This allows for a representation of the amplitude of the Gamma oscillations at any given phase value of the Theta oscillations.

The phase information was then split in 18 bins (each accounting for 20-degree changes) and the mean amplitude (acquired from the Gamma oscillations) over each phase bin was calculated. Finally, the mean amplitude values for each bin were normalized; the normalized amplitude acts a discrete probability density function. For no PAC, the amplitude distribution (P) over the phase bins would be uniform; therefore, a deviation from a uniform distribution denotes coupling. The deviation was calculated using the KL distance [48], adapted in order to obtain deviation values between 0 and 1. The KL distance is defined as

$$D_{KL}(P,Q) = \sum_{j=1}^{N} P(j) \log \left[\frac{P(j)}{Q(j)} \right] \tag{1}$$

where *D* is the KL distance between a discrete distribution *P* and a distribution *Q* for each element *j* in a signal of length *N*. Moreover, the KL distance can be expressed in terms of the Shannon entropy *H* of the distribution *P*, as

$$D_{KL}(P,U) = \log(N) - H(P) \tag{2}$$

where the KL distance is now expressed as the distance between the distribution *P* and the uniform distribution U. Here, $\log(N)$ corresponds to the maximal possible entropy value (uniform distribution). A modulation index (MI) is then calculated, by dividing the KL distance as expressed in (2) by $\log(N)$ as

$$MI = \frac{D_{KL}(P,U)}{\log(N)}. \tag{3}$$

For a uniform distribution of the amplitude over the phase values, the MI is zero; a value of one would correspond to a distribution centred on a specific bin (a Dirac distribution) representing a Gamma oscillation that only occurs in a single phase bin of Theta.

The MI presents very low values (as seen in [42], a "locked" Theta-phase Gamma-amplitude modulation would reach an MI value around 0.015 in ideal, simulated conditions). It is also important to note that the value is sensitive to the noise within and the length of the measurement (the measurement should have more than one cycle of the phase information). The training tasks (baseline, driving, and auditory training) are shorter (80 s) than the main tasks (240 s), and therefore, present larger MI values (longer measurements present a smaller MI variation, allowing for a stabilization of the values); the EEG information from training stages was not used for the PAC analysis due to this reason.

The MI was calculated for each channel, for each task, and for each subject, separately. Once finished, the mean for each subject group was calculated for each of the tasks (e.g., all drivers, drivers from first round separately, all copilots, and so on). The MI values were then plotted over the scalp using EEGLAB-based scripts. If an electrode had an MI value three times bigger than the average value of all electrodes, it was considered a faulty electrode; its value was then assigned based on its nearest neighbors. Measurements with more than ten faulty electrodes were not taken into account. For the final analysis, 18 of the original 22 subjects were considered, based on the quality of the measurement.

TABLE I
MEAN TASK RESULTS (IN %) FOR THE AUDITORY TASK

Group	Audio Training	A	2A	C	2C
Drivers R1	83.5	48.9	50.2	39.7	37.6
Drivers R2	82.1	56	57.3	47.6	46
D R1 + R2	82.8	52.4	53.8	43.7	41.8
Copilots R1	78.7	67.1	67.1	55.6	56.6
Copilots R2	88.3	66.7	65.8	54.5	57.7
C R1 + R2	83.5	66.9	66.4	55	57.1

IV. RESULTS

A. Behavioral Data (Detection Task)

The number of targets for each task were as follows: Training (7); Task A (25); Task C (21). Task B did not require any button presses. The task results (percentage of target words detected) are shown in Table I and represent the behavioral data of the study.

Scores for task A were higher than task C for both test and retest situations (scores for the audio training without driving were higher than both tasks A and C as expected). A repeated-measures analysis of variance (ANOVA) was performed to compare the results (in % of correct button pressings per task) between audio training, task A and C, and their retests. For each group (drivers R1, drivers R2, copilots R1, and copilots R2), the difference between tasks was significant (all $p < 0.001$ except copilot R2 $p = 0.0079$). Additionally, the difference between drivers and copilots was studied for significance as well (drivers R1 against copilots R1 and drivers R2 against copilots R2). The group difference for R1 was significant ($p = 0.0366$); the group difference for R2 did not reach significance ($p = 0.2176$).

The questionnaires filled out by the participants after the measurement contained questions rated on a five-point scale. The participants described on average the difficulty to solve the behavioral task as very easy during task A (highway) and easy/mild during task C (mixed road with bad weather). Of the

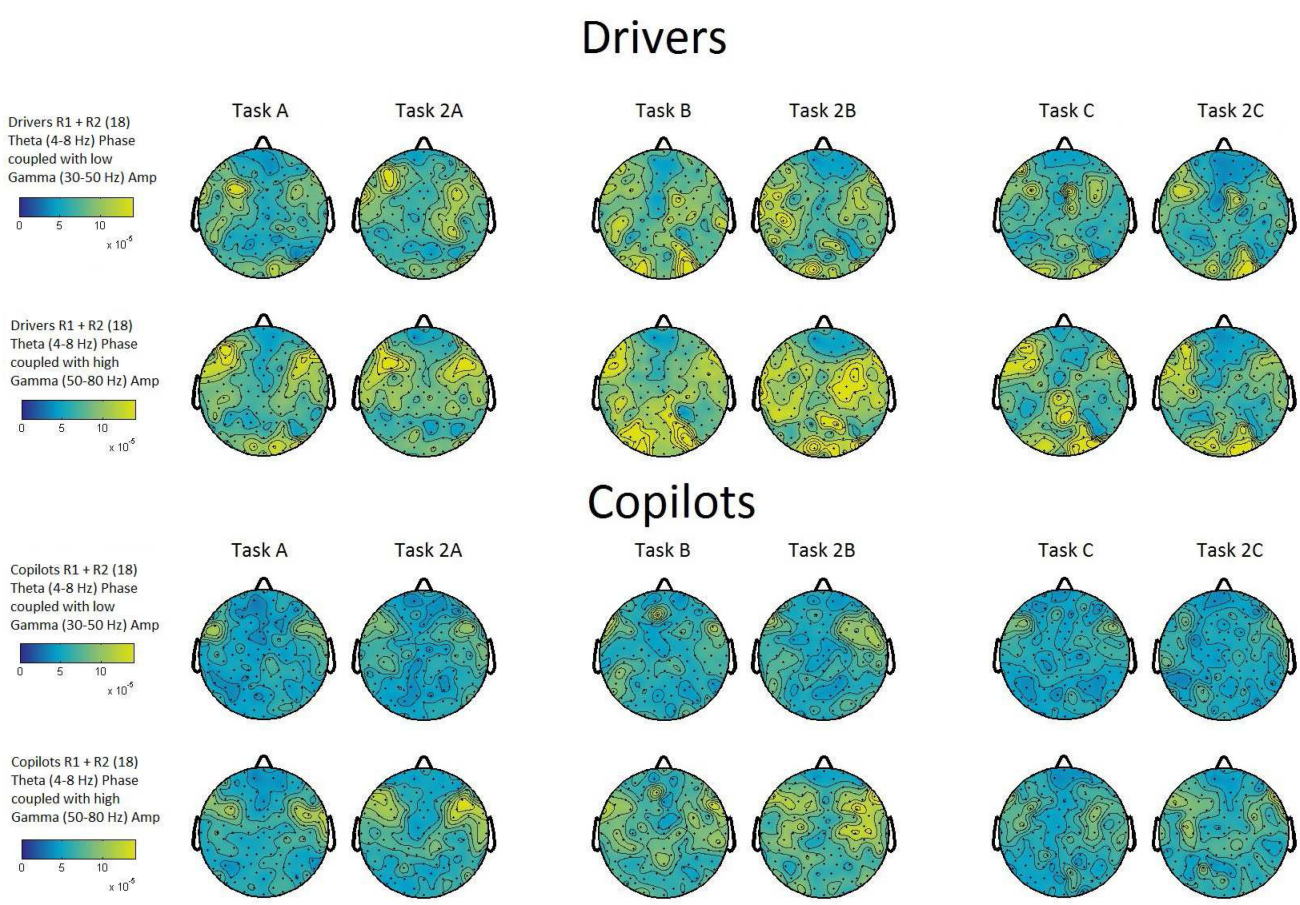


Fig. 3. PAC results for all driver and copilots, main tasks. MI is truncated to 0.00014 for clarity. Low (above) and high (below) Gamma amplitudes are shown, for all tasks.

22 subjects, 19 rated the weather as having a low influence on their driving performance (1/5) while three subjects rated it as having a strong influence (5/5). Compared to the behavioral task results, subjects overestimated their performance and level of distraction.

B. Phase–Amplitude Coupling

Fig. 3 shows the PAC results for all drivers (R1 and R2 combined), and all copilots (R1 and R2 combined), for all tasks. Fig. 4 shows the results for drivers only, comparing R1 and R2 rounds.

C. PAC Statistical Analysis

To assess if there was indeed a significant difference between drivers and copilots in general, a mixed ANOVA test using the PAC value of each single electrode as the dependent variable was used (between-subjects factor—driver/copilot; within-subjects-factor—tasks). The difference between drivers "R1" and "R2" was again assessed using a mixed-ANOVA test (between-subjects factor—drivers R1/R2; within-subjects factor—tasks). In order to assess significant differences between single tasks, one-way repeated measures ANOVA was used between all tasks (again using the PAC value of each electrode as the dependent variable); drivers and copilots were analyzed separately.

Nine regions of interest (ROIs) were defined to present the results—dorsolateral prefrontal cortex left (DLPFC L) and right (DLPFC R), frontal eye fields left (FEF L) and right (FEF R), superior frontal gyrus (SFG), primary motor cortex left (PMC L) and right (PMC R), and primary visual cortex left (PVC L) and right (PVC R). Electrodes that presented significant differences in the statistical analysis were grouped into each ROI and accounted for. The ROIs are shown in Fig. 5 and the number of electrodes with significant differences are shown in Fig. 6.

Given the results from the repeated measures ANOVA test, significant differences between the first and second round of each task were additionally investigated using a paired-T test (comparison between test and retest). Results are shown in Figs. 7 (drivers) and 8 (copilots).

V. Discussion

The behavioral data (results on the detection task) showed a significant difference between the pure auditory task (training, without driving) and the multimodal tasks (A and C), which allows to confirm that the tasks were distinct enough and that task

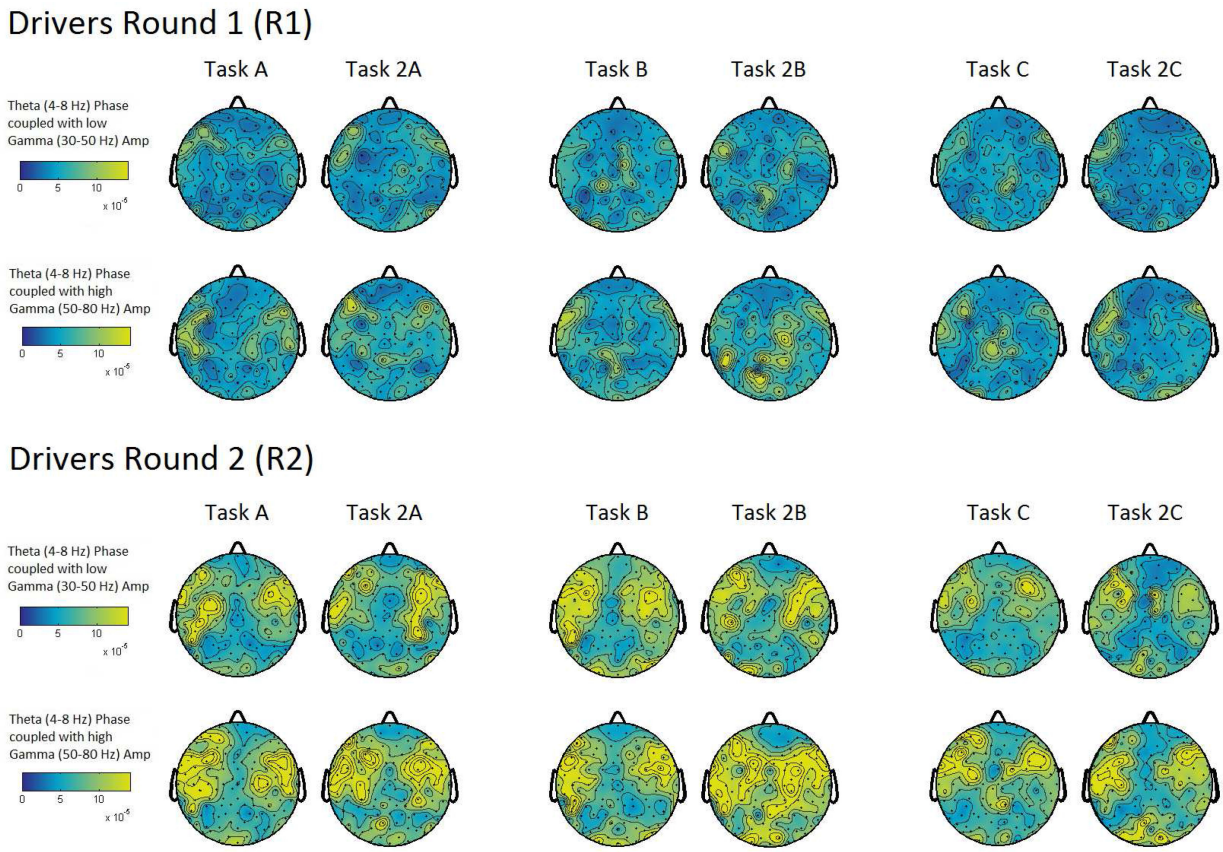


Fig. 4. PAC results for drivers from rounds 1 and 2 separately. Modulation index is truncated to 0.00014 for clarity. Low (above) and high (below) Gamma amplitudes are shown, for all tasks.

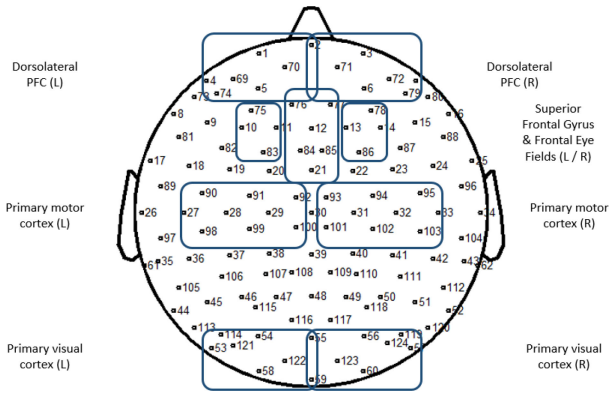


Fig. 5. Regions of interest defined for the statistical analysis.

C was more difficult than A (as seen in the scores). It also showed a high retest reliability in both tasks A and C (see Table I). A significant difference was also found between drivers and copilots for the first round.

Regarding PAC indices and focusing on the Theta/high Gamma band, an overall higher activation was seen in drivers (both R1 and R2) compared to copilots, especially in occipital areas; activation on prefrontal and parietal areas was observed on both, but reduced in copilots (see Fig. 3). Electrodes showing significant PAC differences were found in prefrontal, primary motor cortex, frontal eye fields, and occipital areas, confirming the main hypothesis of this article, which suggested that PAC might be a suitable biomarker for attention-related cortical activation. From the Gamma subdivisions chosen, the high-Gamma domain (50–80 Hz) offered the most recognizable differences overall, as seen by the number of electrodes with significant differences (see Fig. 6). For example, the PMC R region presented six electrodes out of ten with significantly higher PAC when observed in the high-Gamma domain, compared to one single electrode out of ten in the low-Gamma band for the same region. Significant differences between drivers and copilots were found in the same regions both for low and high Gamma (motor area, left prefrontal cortex, and occipital area); however, as previously discussed and as seen in Fig. 6, a higher number of electrodes within the high-Gamma band presented a significant PAC difference.

The observation regarding the high-Gamma band agrees with the literature regarding cross-frequency coupling, where usually electrocorticography (ECOG) is preferred and higher Gamma-band components are taken into account; it has been discussed that surface EEG cannot provide reliable signal components above 80 Hz [39], [43]. Our results suggest that the Gamma range between 50 and 80 Hz might provide enough information for cross-frequency coupling, even when using surface EEG.

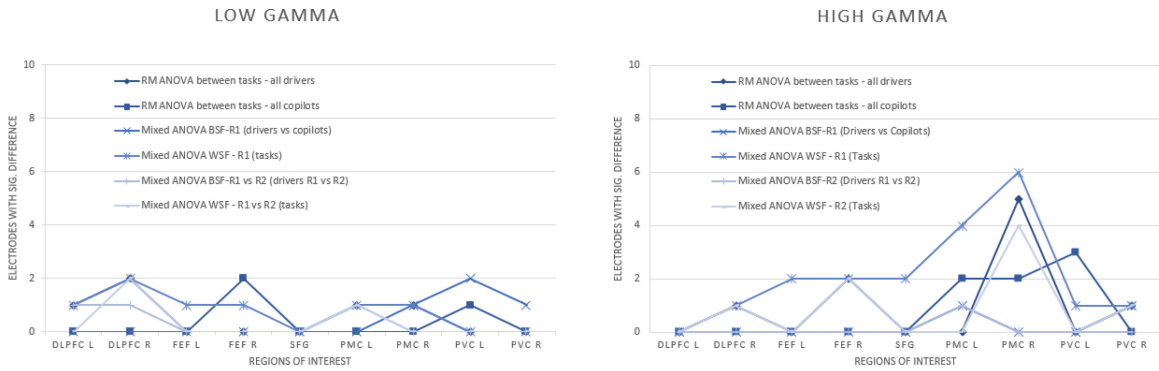


Fig. 6. Number of electrodes presenting significant differences in PAC, for each ROI. The ROIs are DLPFC L, DLPFC R, FEF L, FEF R, SFG, PMC L, PMC R, PVC L, and PVC R. (Right) A higher number of electrodes presented significant differences in PAC within the high-Gamma band (6) compared to the same region when limited to the low-Gamma band (1), thus, suggesting that a stronger PAC is seen in the high-Gamma (50–80 Hz) range.

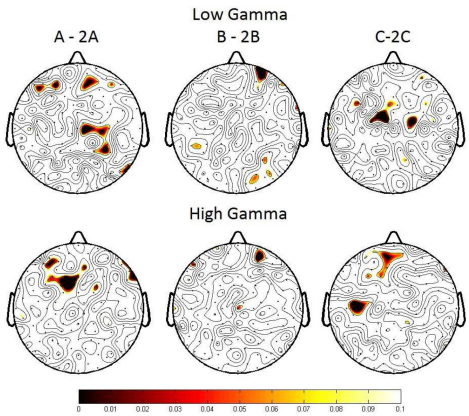


Fig. 7. (Drivers) Paired-T test results between test and retest of each task, for low (above) and high (below) Gamma.

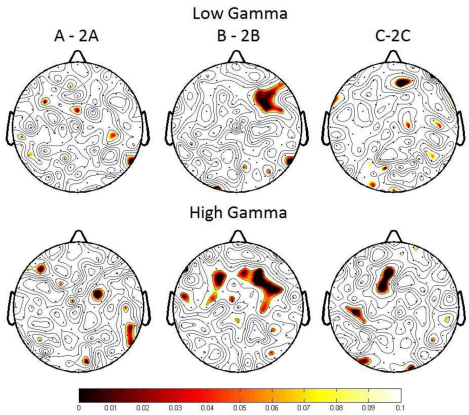


Fig. 8. (Copilots) Paired-T test results between test and retest of each task, for low (above) and high (below) Gamma.

When analyzing drivers R1 against drivers R2, R2 showed overall higher PAC (see Fig. 4) in prefrontal and parietal areas. Since these participants knew the tasks beforehand, this finding might perhaps indicate an increased motivation/effort to solve the tasks correctly. This agrees with the behavioral data; the scores were reduced for the multimodal tasks suggesting an increased effort to solve the task. However, there was no significant difference in the PAC in the occipital areas when comparing R1 and R2 drivers (mixed ANOVA, between subject factor), which may indicate that a similar effort is required for the driving task as compared to R1, and reduced (but present) significant differences in the prefrontal cortex and left parietal cortex. The PAC difference between tasks between R1 and R2 was significant on the right prefrontal cortex, right frontal eye fields, right primary motor cortex, and right primary visual cortex as well (mixed ANOVA, within subject factor).

Significant differences between the test and retest of each task (see Figs. 7 and 8) were found, albeit strictly localized. For drivers, significant differences were found mostly in the prefrontal cortex and motor area. For copilots, significant differences were found in the frontal eye field region for task B and C in the high-Gamma band. This might suggest a learning effect in the retest stages. Moreover, this might explain why the difference in behavioral data between drivers and copilots was not significant for the second round. Prefrontal areas presented activity during most of the tasks, both for drivers and copilots. This agrees with the idea of assessing working memory-involved tasks through PAC, which include executive, integrative, and storage functions simultaneously [19], [45]. Left dorsolateral prefrontal cortex showed a significant difference between all tasks for the drivers only; this might point to different resource allocations during a multimodal task (trying to drive in taxing weather conditions while solving an auditory task).

Opposed to our secondary hypothesis, we did not find an increased PAC around the auditory cortex areas in the low-driving effort/high-auditory effort tasks. A reason might be either a prioritization of visual attention or the lack of a "pure" auditory stimulation, instead being replaced by multimodal stimulation. The main objective of the auditory task was to provide a distraction to the driving task and to act as a behavioral task to rate the difficulty of scenarios. There might also not have been enough contrast between the auditory task and the distraction, rendering the task easy to solve. However, this contradicts the behavioral task results, which show a lower success rate for tasks A and C compared to pure auditory tasks (see Table I). The designed

auditory task corresponds more strictly to a detection task, which might not elicit the same effort compared to an auditory scene analysis task [52].

Additional drawbacks of the current study are given by the spatial distribution of the sensors. Whereas a 128-electrode system offers good spatial resolution, the measured PAC can be influenced by volume conduction and correct placement of electrodes for each measurement and their performance during the complete length of the study. Both factors increase the variability of the results. The obtained results provide an insight of areas showing increased PAC, but they would greatly benefit from comparison to studies using more in-depth (e.g., invasive or combined with MEG) measurements. Moreover, as explained by Tort *et al.* in [42] and as seen in the PAC results for the baseline and training stages, the PAC scale is sensitive to time resolution, i.e., shorter measurements present higher PAC indices due to less cycles available for each frequency band and an increased SNR (although the ratio of activation between coupled and noncoupled areas seems to be conserved). This is due to the MI's dependence on the length of the signal (in this case, the length of the EEG measurement). Tort *et al.* suggest either a minimum of 30 s, or more than 200 cycles for Theta-band analysis. While our baseline and training stages (80 and 120 s) fulfill the criteria, longer measurements (such as the main tasks with 240 s) provided more accurate results.

VI. Conclusion

A higher PAC index was observed in the EEG from drivers, for prefrontal (DLPFC, FEF), parietal (PMC), and occipital (PVC) areas, especially on tasks B and 2B (pure driving tasks). The high-Gamma range (50–80 Hz) seems to be more sensitive to phase locking analysis than low Gamma as judged by the number of electrodes with significantly different PAC indices. Copilots presented an overall lower cortical activation, specifically in the occipital area compared to the pilots; this might indicate a reduced cognitive load for visual tasks and a focus on the context of the behavioral task (prefrontal activity).

The hypothesis of PAC as a biomarker of attention-related cortical activation was confirmed by the region-locked results, especially within the high-Gamma band. Additional significance measures were able to selectively detect areas of interest such as the frontal eye fields, the primary motor cortex, and the primary visual cortex. The study allowed the identification of areas in which PAC can be measured and might be easily implemented in real driving situations; this information might be used in driver-assistance systems to enhance safety. The nonnaive drivers (Round 2) presented an overall higher PAC, which might also suggest PAC as a tool to isolate different multimodal cognitive effort; however, the expected activity in the auditory cortex was not found.

Future work should be focused on increasing the reliability of PAC as a biomarker for attention in multimodal tasks. Coupling between different brain areas, which has been shown to have effect in processes such as memory encoding [53] should also be studied. Future work might also include the analysis of PAC in nondriving modality-specific tasks (e.g., pure auditory stimulation) and their comparison to multimodal tasks.

References

[1] T. A. Schweizer, K. Kan, Y. Hung, F. Tam, G. Naglie, and S. J. Graham, "Brain activity during driving with distraction: An immersive fMRI study," *Frontiers Human Neurosci.*, vol. 7, no. 53, pp. 1–11, 2013.

[2] L. Tijerina, S. Johnston, E. Parmer, and M. D. Winterbottom, "Driver distraction with wireless telecommunications and route guidance systems," Nat. Highway Traffic Saf. Admin., U. S. Dept. Transp., Washington, DC, USA, Tech. Rep. DOT HS 809-0692, 2000.

[3] W. J. Horrey and C. D. Wickens, "Driving and side task performance: The effects of display clutter, separation, and modality," *Human Factors*, vol. 46, no. 4, pp. 611–624, 2004.

[4] J. M. Owens, S. B. McLaughlin, and J. Sudweeks, "Driver performance while text messaging using handheld and in-vehicle systems," *Accident Anal. Prevention*, vol. 43, pp. 939–947, 2011.

[5] D. V. McGehee, "Visual and cognitive distraction metrics in the age of the smart phone: A basic review," *Ann. Adv. Automot. Med.*, vol. 58, pp. 15–23, 2014.

[6] K. Young and M. Regan, "Driver distraction: A review of the literature," in *Distracted Driving*, I. J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter, and J. Irwin, Eds. Sydney, NSW, Australia: Australas. College Road Saf., 2007, pp. 379–405.

[7] P. Sagaspe *et al.*, "Extended driving impairs nocturnal driving performances," *PLoS One*, vol. 3, no. 10, pp. 1–6, 2008.

[8] C. T. Lin *et al.*, "Mind wandering tends to occur under low perceptual demands during driving," *Sci. Rep.*, vol. 17, no. 6, 2016, Art. no. 21353.

[9] V. D. Calhoun, V. B. McGinty, and G. D. Pearlson, "Driving and the brain: An imaging study," in *Proceedings of the 1st Human-Centered Transportation Simulation Conference, The University of Iowa, Iowa City, Iowa, November 4–7, 2011*.

[10] V. D. Calhoun and G. D. Pearlson, "A selective review of simulated driving studies: Combining naturalistic and hybrid paradigms, analysis approaches, and future directions," *NeuroImage*, vol. 59, pp. 25–35, 2012.

[11] M. A. Just, T. A. Keller, and J. Cynkar, "A decrease in brain activation associated with driving when listening to someone speak," *Brain Res.*, vol. 1205, pp. 70–80, 2008.

[12] S. C. Chung *et al.*, "Effects of distraction task on driving: A functional magnetic resonance imaging study," *Bio-Med. Mater. Eng.*, vol. 24, pp. 2971–2977, 2014.

[13] S. M. Bowyer *et al.*, "Conversation effects on neural mechanisms underlying reaction time to visual events while viewing a driving scene using MEG," *Brain Res.*, vol. 1251, pp. 151–161, 2009.

[14] A. Fort *et al.*, "Attentional demand and processing of relevant visual information during simulated driving: A MEG study," *Brain Res.*, no. 1363, pp. 117–127, 2010.

[15] K. Sakihara *et al.*, "Cerebral oscillatory activity during simulated driving using MEG," *Frontiers Human Neurosci.*, vol. 8, no. 975, pp. 1–9, 2014.

[16] B. T. Jap, S. Lal, P. Fischer, and E. Bekiaris, "Using EEG spectral components to assess algorithms for detecting fatigue," *Expert Syst. Appl.*, vol. 36, pp. 2352–2359, 2009.

[17] C. T. Lin, S. A. Chen, T. T. Chiu, H. Z. Lin, and L. W. Ko, "Spatial and temporal EEG dynamics of dual-task driving performance," *J. NeuroEng. Rehabil.*, vol. 8, no. 11, pp. 1–13, 2011.

[18] C. T. Lin, R. C. Wu, T. P. Jung, S. F. Liang, and T. Y. Huang, "Estimating driving performance based on EEG spectrum analysis," *EURASIP J. Appl. Signal Process.*, vol. 19, pp. 3165–3174, 2005.

[19] S. Lei and M. Roetting, "Influence of task combination on EEG spectrum modulation for driver workload estimation," *Human Factors*, vol. 53, no. 2, pp. 168–179, 2011.

[20] G. Li and W. Y. Chung, "A context-aware EEG headset system for early detection of driver drowsiness," *Sensors*, vol. 15, pp. 20873–20893, 2015.

[21] E. Wascher, S. Getzmann, and M. Karthaus, "Driver state examination—Treading new paths," *Accident Anal. Prevention*, vol. 91, pp. 157–165, 2016.

[22] J. M. Morales *et al.*, "Monitoring driver fatigue using a single-channel electroencephalographic device: A validation study by gaze-based, driving performance, and subjective data," *Accident Anal. Prevention*, vol. 109, pp. 62–69, 2017.

[23] K. Yoshino, N. Oka, K. Yamamoto, H. Takahashi, and T. Kato, "Functional brain imaging using near-infrared spectroscopy during actual driving on an expressway," *Frontiers Human Neurosci.*, vol. 7, no. 882, pp. 1–16, 2013.

[24] N. Oka *et al.*, "Greater activity in the frontal cortex on left curves: A vector-based fNIRS study of left and right curve driving," *PLoS One*, vol. 10, no. 5, 2015, Art. no. e0127594.
[25] R. Nosrati, K. Vesely, T. Schweizer, and V. Toronov, "Event-related changes of the prefrontal cortex oxygen delivery and metabolism during driving measured by hyperspectral fNIRS," *Biomed. Opt. Exp.*, vol. 7, no. 4, pp. 1323–1335, 2016.
[26] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, "Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG and fNIRS data," *Frontiers Human Neurosci.*, vol. 10, no. 219, pp. 1–14, 2016.
[27] A. K. Engel, P. Fries, and W. Singer, "Dynamic predictions: Oscillations and synchrony in top–down processing," *Nature Rev. Neurosci.*, vol. 2, pp. 704–718, 2001.
[28] C. Tallon-Baudry, O. Bertrand, C. Delpuech, and J. Pernier, "Oscillatory gamma-band (30–70 Hz) activity induced by a visual search task in humans," *J. Neurosci.*, vol. 17, no. 2, pp. 722–734, 1997.
[29] T. Gruber, M. M. Müller, A. Keil, and T. Elbert, "Selective visual-spatial attention alters induced gamma band responses in the human EEG," *Clin. Neurophysiol.*, vol. 110, pp. 2074–2085, 1999.
[30] C. S. Herrmann, M. H. K. Munk, and A. K. Engel, "Cognitive functions of gamma-band activity: Memory match and utilization," *Trends Cogn. Sci.*, vol. 8, no. 8, pp. 347–355, 2004.
[31] O. Jensen, J. Kaiser, and J. P. Lachaux, "Human gamma-frequency oscillations associated with attention and memory," *Trends Neurosci.*, vol. 30, no. 7, pp. 317–324, 2007.
[32] W. Klimesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis," *Brain Res. Rev.*, vol. 29, pp. 169–195, 1999.
[33] M. X. Cohen, "Error-related medial frontal theta activity predicts cingulate-related structural connectivity," *NeuroImage*, vol. 55, pp. 1373–1383, 2011.
[34] M. Gärtner, L. Rohde-Liebenau, S. Grimm, and M. Bajbouj, "Working memory-related frontal theta activity is decreased under acute stress," *Psychoneuroendocrinology*, vol. 43, pp. 105–113, 2014.
[35] P. Arrighi *et al.*, "EEG theta dynamics within frontal and parietal cortices for error processing during reaching movements in a prism adaptation study altering visuo-motor predictive planning," *PLoS One*, vol. 11, no. 3, pp. 1–27, 2016.
[36] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2013.
[37] P. Fries, "A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence," *Trends Cogn. Sci.*, vol. 9, no. 10, pp. 474–480, 2005.
[38] P. J. Uhlhaas, F. Roux, E. Rodriguez, A. Rotarska-Jagiela, and W. Singer, "Neural synchrony and the development of cortical networks," *Trends Cogn. Sci.*, vol. 14, no. 2, pp. 72–80, 2009.
[39] O. Jensen and L. L. Colgin, "Cross-frequency coupling between neuronal oscillations," *Trends Cogn. Sci.*, vol. 11, no. 7, pp. 267–269, 2007.
[40] M. X. Cohen, "Assessing transient cross-frequency coupling in EEG data," *J. Neurosci. Methods*, vol. 168, pp. 494–499, 2008.
[41] V. Jirsa and V. Müller, "Cross-frequency coupling in real and virtual brain networks," *Frontiers Comput. Neurosci.*, vol. 7, no. 78, pp. 1–25, 2013.
[42] A. B. L. Tort, R. Komorowski, H. Eichenbaum, and N. Kopell, "Measuring phase-amplitude coupling between neuronal oscillations of different frequencies," *J. Neurophysiol.*, vol. 104, pp. 1195–1210, 2010.
[43] R. T. Canolty *et al.*, "High gamma power is phase-locked to theta oscillations in human neurocortex," *Science*, vol. 313, pp. 1626–1628, 2006.
[44] J. Lisman, "The theta/gamma discrete phase code occuring during the hippocampal phase precession may be a more general brain coding scheme," *Hippocampus*, vol. 15, pp. 913–922, 2005.
[45] S. I. Dimitriadis, Y. Sun, K. Kwok, N. Laskaris, N. Thakor, and A. Bezerianos, "Cognitive workload assessment based on the tensorial treatment of EEG estimates of cross-frequency phase interactions," *Ann. Biomed. Eng.*, vol. 43, no. 4, pp. 977–989, 2015.
[46] A. Bruns and R. Eckhorn, "Task-related coupling from high- to low-frequency signals among visual cortical areas in human subdural recordings," *Int. J. Psychophysiol.*, vol. 51, pp. 97–116, 2004.
[47] T. Demiralp *et al.*, "Gamma amplitudes are coupled to theta phase in human EEG during visual perception," *Int. J. Psychophysiol.*, vol. 64, pp. 24–30, 2007.
[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
[49] A. Ruoff, *Häufigkeitswörterbuch Gesprochener Sprache: Gesondert Nach WOrtarten, Alphabetisch, Rückläufig Alphabetisch und Nach Häufigkeit Geordnet*. Berlin, Germany: de Gruyter (in German), 1981.
[50] H. Fastl, "A background noise for speech audiometry," *Audiol. Acoust.*, vol. 26, pp. 2–13, 1987.
[51] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, pp. 411–430, 2000.
[52] C. Bernarding, D. J. Strauss, R. Hannemann, H. Seidler, and F. I. Corona-Strauss, "Neurodynamic evaluation of hearing aid features using EEG correlates of listening effort," *Cogn. Neurodyn.*, vol. 11, no. 3, pp. 203–215, 2017.
[53] U. Friese, M. Köster, U. Hassler, U. Martens, N. Trujillo-Barreto, and T. Gruber, "Successful memory encoding is associated with increased cross-frequency coupling between frontal theta and posterior gamma oscillations in human scalp-recorded EEG," *Neuroimage*, vol. 66, pp. 642–647, 2013.

Using EEG for Mental Fatigue Assessment: A Comprehensive Look Into the Current State of the Art

Thiago Gabriel Monteiro, Charlotte Skourup, and Houxiang Zhang, *Senior Member, IEEE*

***Abstract*—This paper provides a brief survey of recent developments on the use of electroencephalogram (EEG) sensors for detecting mental fatigue (MF) in human operators during tasks involving human–machine interaction. This research topic has received much attention since there is a consensus among experts on the increasing relation between human failure and accidents in safety-critical tasks. MF is one of the most influential aspects leading to human failure and the most reliable way to assess it is using operator's physiological data, especially EEG. In the past few decades, hundreds of publications have explored the use of EEG alone or together with other objective and subjective measures for assessing MF, drowsiness, and tiredness in human operators. With recent improvements in data preprocessing, feature extraction, and classification algorithms, the monitoring and mitigation of MF in real time has become a reality. This trend is mainly due to the increasing use of machine learning techniques. This paper provides a comprehensive look at the current state of the art in the field of MF detection using EEG, identifying the currently used technique, algorithms, and methods and possible trends and promising areas for further research. The paper is concluded by suggesting a kernel partial least squares discrete-output linear regression based model as an all-around good option for an MF assessment system.**

***Index Terms*—Electroencephalogram (EEG), human factors, human–machine systems, mental fatigue (MF) assessment, risk assessment, sensor fusion.**

I. Introduction

In RECENT decades, the greater role human operators play in life-threatening accidents than systems and equipment malfunction or failure has become increasingly clear [1], [2]. Research has demonstrated this in relation to driving [3], public transportation [4], [5], commercial air transportation [6], air traffic control [7], nuclear power plants [8], maritime operations [9], etc. This has prompted explorations into assessments of operator functional state (OFS) as a key means of lowering risk.

Manuscript received October 22, 2018; revised February 19, 2019; accepted July 30, 2019. Date of publication September 9, 2019; date of current version November 21, 2019. This work was supported in part by the project "SFI Offshore Mechatronics" funded by Norway Research Council, Norway (Project 237896) and in part by the project "SFI Marine Operations" funded by Norway Research Council, Norway (Project 237929). This article was recommended by Associate Editor L. Contreras-Vidal. *(Corresponding author: Thiago Gabriel Monteiro.)*

T. G. Monteiro and H. Zhang are with the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Aalesund 6009 Norway (e-mail: thiago.g.monteiro@ntnu.no; hozh@ntnu.no).

C. Skourup is with the ABB AS, Products and Services R&D; Oil, Gas and Chemicals, Oslo 0603, Norway (e-mail: charlotte.skourup@no.abb.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2938156

OFS can be characterized as how well a human operator can react to the demands of an operation considering internal and external factors, according to the operator's cognitive and physiological capabilities [10]. OFS is a broad concept, but it can be evaluated under the perspective of three main areas: situation awareness, mental workload, and mental fatigue (MF). MF builds up as an operation progresses and can drastically reduce operators ability to understand, react, and solve problems imposed by the operation quickly, including both common procedures and unexpected situations.

Various methods have emerged of assessing MF. Subjective evaluations include the NASA Task Load Index [11], the Karolinska Sleepiness Scale [12], the Epworth sleepiness scale [13], and the Chalder fatigue scale [14]. While these subjective approaches can achieve good results in assessing MF states, they rely on self-report and are thus subject to bias. More objective methods include monitoring operators performance, such as tracking steering wheel movements or pressure on the acceleration pedal during a drive task [15] and monitoring the operator's behavior, including head position, blinking frequency, and yawning [16]. But monitoring the operator's physiological signals is considered the most reliable way to assess MF, since physiological signals start to change long before any external signs of MF manifest [17]. The physiological signals researchers have used include respiration, electrocardiogram (ECG), electromyogram (EMG), electrooculogram (EOG), and electroencephalogram (EEG).

Although it can achieve high accuracy level, the use of physiological signal for assessing MF can provide some challenges. First, measurement generally requires physical contact between operators and sensors, which can make operators uncomfortable and affect the measured signals [18]. Second, analysis accuracy is very sensitive to the quality of the measured signals. In most cases, these signals are very susceptible to noise and need to be preprocessed in order to provide any useful information. EEG is the most prominent signal in the field today due to its low intrusiveness [19] and its clear relation between the power spectrum characteristics in different frequency bands and MF levels [20]. EEG signals also have other applications in the medical field such as seizure detection and engineering field such as brain–computer interfaces.

The main goal of this survey paper is to provide a comprehensive understanding of the current state-of-the-art techniques regarding the use of EEG to assess MF and how to apply these

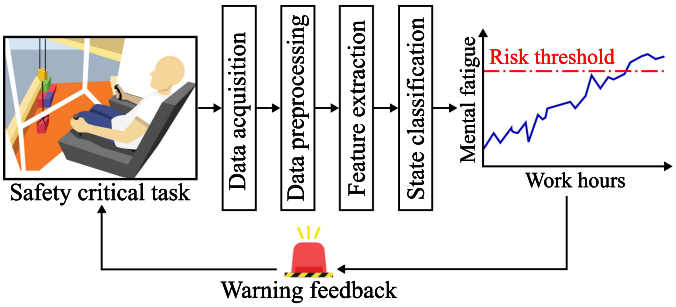


Fig. 1. Closed-loop EEG-based MF assessment framework.

techniques in real-life situations. We seek to understand the current trends in the field in order to present to researchers the newest techniques available and the new knowledge built in this area in the past years.

The rest of the paper unfolds as follows. Section II provides a comprehensive summary of MF assessment pipeline using EEG. Section III describes the methodology used for the literature survey and analysis. Section IV describes the state-of-the-art methods currently available for assessing MF using EEG. In Section V, we discuss the current picture of the EEG for MF assessment research field and provide our vision of a feasible EEG-based MF detection system for implementation in real-life applications. Section VI concludes the paper.

II. EEG for Metal Fatigue Assessment—Background

A. Structure of an EEG-Based MF Assessment Method

An EEG-based algorithm that can be used to assess MF must be structured to provide data acquisition, data preprocessing, feature extraction, and MF state classification [21]. Fig. 1 provides an example of the use of this structure in the MF detection framework we propose. EEG data is collected from a crane operator working in an oil rig and preprocessed; features are extracted, and the MF state is classified. The state assessment is used in a threshold algorithm, which closes the loop with the operator with a warning feedback when the MF level exceeds a critical level.

Beyond the basic structure, the methods to handle and classify EEG data need to be specially tailored for specific applications. In addition to the basic underlying structure, new elements can be added to the workflow according to the special needs of each algorithm. For example, applying a dimension reduction technique on the features vector will reduce the classification algorithm input to its more meaningful components [22].

The next sections describe the basic processes present in EEG-based MF assessment methods.

B. Data Acquisition

During the data acquisition phase, two main aspects govern the acquisition of EEG signals, as described below.

1) Electrodes Placement: Jasper [23] created the International 10–20 system in order to standardize the placement of

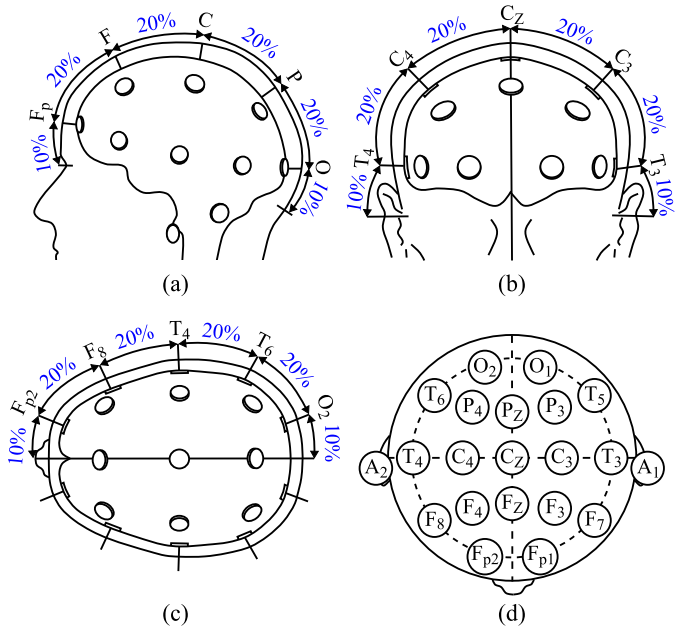


Fig. 2. International 10–20 system definition. (a) Lateral specification. (b) Frontal specification. (c) Superior specification. (d) Electrodes placement and naming. Adapted from [24].

electrodes on the scalp to make the direct comparison of results from different research groups possible.

The naming convention for each electrode is as follows: fronto-polar area (Fp), frontal area (F), central area (C), parietal area (P), occipital area (O), temporal area (T), and auricular electrodes (A). Besides the area names, there is also a numbering convention to help distinguish between left and right homologous regions, using odd numbers for the left hemisphere, even numbers for the right hemisphere, and "z" (standing for zero) for the vertex electrodes [24]. The full electrode arrangement is shown in Fig. 2.

The 10–20 system offers 21 positions to place electrodes, which is not enough when using newer EEG hardware that supports 32, 64, 128, or 256 data channels [25]. To account for the increase of data channels, the 10–20 system was expanded to define more standardized position in the scalp for electrode placement. The extended 10–20 system (also called 10–10 system) [26] increases the number of standard positions from 21 to 74. Other extensions have been proposed such as a 10–5 system supporting up to 345 standard positions [27].

2) Frequency Bands: EEG signals are composed of a wide range of frequency components. When evaluating EEG signals for MF detection, the range of interest is limited to between 0.3 and 30 Hz. This frequency interval is divided into five main frequency bands (also called rhythms): delta, theta, alpha, beta, and gamma [28].

1) Delta band (δ) corresponds to the interval of 0.3–4 Hz. It represents very slow brain activity, identified in infants up to 1 year or in deep sleep stage in health adults. It is usually not used for MF detection, since it is mostly present in a physiological state outside the interest of MF detection studies.

2) Theta band (θ) corresponds to the interval of 4–8 Hz. It can be found in healthy, alert infants, and children as well as during drowsiness and sleep in adults. Awake, healthy adults have low θ activity. The frontal θ activity is likely to increase as a person fatigues [29].
3) Alpha band (α) corresponds to the interval of 8–13 Hz. It can be found in healthy awake adults, when relaxed or mentally inactive. The occipital and parietal α activities are likely to increase as a person fatigues [29].
4) Beta band (β) corresponds to the interval of 13–30 Hz. It signifies tension and anticipation and can be found in alert and anxious subjects. The changes in β activity as a person fatigues still unclear [29].
5) Gamma band (γ) consists of frequencies above 30 Hz. Usually does not have impact on MF detection, being filtered out of the EEG data [30].

C. Data Preprocessing

EEG signals have very small amplitudes and are highly sensitive to noise. These noises are called artifacts. They are undesirable electrical potentials that come from sources other than the brain [28], such as EOG, EMG, ECG, and power line and amplifier noises, poor electrodes' contact with the scalp and current drift [31]. When present in the data, these noise components make the analysis of the desired phenomena nearly impossible [32], since they may have amplitude in the order of hundreds of μV [33] and EEG signals are in the order of tens of μV. Therefore, artifacts need to be detected and removed from the data. Understanding the advantages and limitations of each preprocessing technique is very important in order to choose the right method for each EEG application.

In recent years, the most commonly used preprocessing methods in the field of MF detection using EEG include digital filtering, independent component analysis (ICA), and discrete wavelet transform (DWT).

Digital filtering is very useful to remove noise and artifacts that are frequency-specific, such as body movement and power line noises. Among these filters, we can consider low-pass, high-pass, band-pass, and notch filtering [31]. In ICA, the EEG signal is seen as a linear combination of independent signals. The ICA decomposes the multichannel data into temporal independent and spatially fixed components [34].

DWT is the subset of wavelet transforms that discretely samples the wavelets. The DWT is capable of decomposing a signal in the time domain in a series of wave-like oscillations (wavelets) in different frequency bands. The biggest advantage of wavelet transform approaches for handling time-series data is the fact that they preserve the signal temporal information together with frequency, allowing the analysis of nonstationary data, which Fourier transforms cannot achieve.

D. Feature Extraction

After preprocessing, the data are more suitable for use in an MF assessment method, but work remains to make it a favorable format to allow classification algorithms to fully explain the represented phenomenon. In order to make the data contained in the EEG signal more meaningful and manageable, relevant features can be extracted. These features basically represent important characteristics of the dataset in a format more compact and easier to handle. Extracting meaningful features from EEG signals is complicated due to its complex, unstable, nonstationary, and nonlinear nature.

In past years, the most commonly used feature extraction methods in the field of MF detection using EEG include power spectral density (PSD), statistics, and entropy measures.

The PSD of a time series describes the power distribution in the signal as a function of frequency [35]. It is especially relevant for EEG classification, since the power in different frequencies can be related to the brain activity in different subbands of interest, making it possible to evaluate changes in the mental state of a subject by tracking changes in the signal PSD. The calculation of PSD is usually preceded by the application of a Fourier transformation in order to change the EEG signal from time to frequency domain.

Statistics can have poor performance when applied as a feature extraction method for EEG signals, since this kind of data is nonstationary by nature. A common way to avoid this problem is to divide the EEG signal in shorter segments and assume the signal is stationary in each of these segments. In this way, statistical analysis can be applied to EEG signals and parameters such as mean, standard deviation, skewness, and kurtosis can be calculated.

Various entropy measure methods have been used to analyze EEG signals [36]–[40] due to its robustness in evaluating the regularity and predictability of complex systems. Entropy was originally used in thermodynamics to assess the degree of disorder in a system, and now it is also used in information theory as a way to measure the uncertainty of systems [41]. As a person gets fatigued, we can expect a decrease in the entropy level of its EEG signals, indicating a decrease and weakening of brain synapses. Recently, the most commonly used entropy measures are sample entropy (SampEn), fuzzy entropy (FuzEn), approximate entropy (AppEn), and spectral entropy (SpecEn).

When the number of features obtained is too big to be directly used in the classification algorithm or when an improvement in the algorithm performance is needed (especially for online application), dimension reduction techniques such as principal component analysis (PCA) and ICA can be applied to obtain an optimized set of features.

E. MF State Classification

The classification algorithm can classify the input features in any number of classes, depending on how the algorithm is trained or designed to handle the input data. Most of the works use two MF states, but some works consider the existence of intermediate states. These states indicate the transition between the normal and fatigue states.

Among the EEG-based MF assessment methods, no classification algorithm clearly dominates the field. Classification algorithmwise, the state of the art is very heterogeneous, presenting just a few algorithms, which were applied in more than one publication [e.g., Bayesian neural network (BNN),

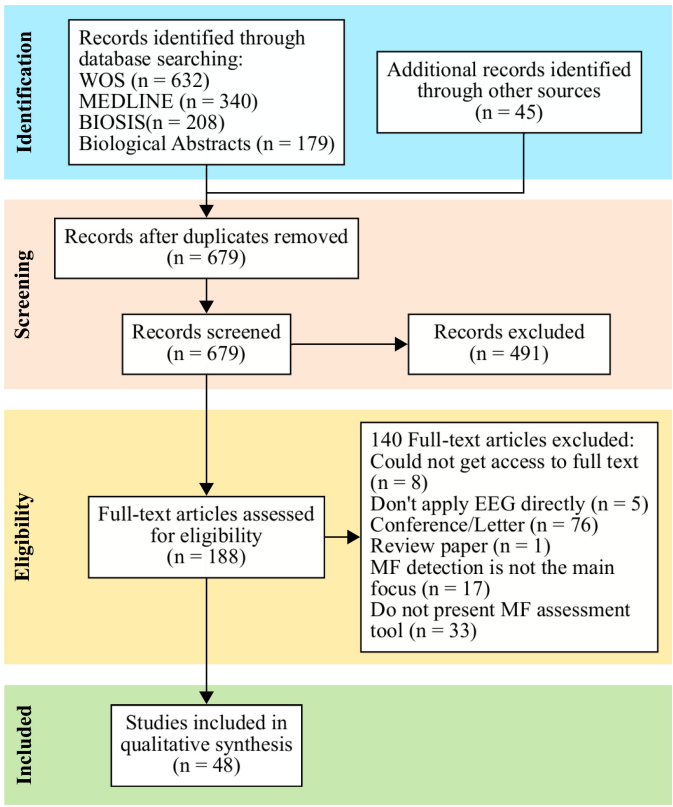


Fig. 3. PRISMA flow diagram.

k-nearest neighbor (kNN), support vector machine (SVM)]. This heterogeneity is due to the fact that no classifier fits all kinds of problems. The best classification tool for each case should be selected based on the particular characteristics of each dataset [42]. The different state-of-the-art classification algorithm will be briefly presented in Section IV.

III. Methods

In this section, we describe our survey methodology. We first present the literature search methodology using the PRISMA statement and follow up by presenting our classification strategy for the surveyed literature.

A. Literature Search Approach

This survey paper is structured following the guidelines presented on the PRISMA statement [43]. The process behind the selection of papers for this survey follows the PRISMA flow diagram, presented in Fig. 3. Papers included in the identification phase were those published in English between January 2013 and December 2017 with a title, abstract, or body containing results for the following Boolean search statement: (MF or fatigue) and EEG. The remaining steps of the selection process are depicted in Fig. 3, including all criteria used in the eligibility phase.

B. Categorization Approach

When describing a method to assess MF using EEG signals, there are some very important characteristics that are relevant to the way it will perform and to how it can be extended to case studies other than the original application. Having a taxonomy describing these characteristics can help researchers to select methods that meet the requirements of their specific problems. Hereafter, we present a taxonomy that highlights the characteristics we consider the most relevant for this kind of application. These characteristics are nonlinearity, nature, dynamics, implementation, and cross subject.

Nonlinear (N): Specifies which kind of discriminant algorithms was used in the current MF assessment method to distinguish between different MF states: linear or nonlinear classifier.

Generative (G): This refers to how the model handles different classes in the dataset. The method is discriminative if it is only capable of learning how to discriminating between the data in different classes, without explaining its structure. If a method is also capable of explaining how the data in each class is structured, being able to model the data, it is called a generative method.

Dynamic (D): Specifies if the MF assessment method uses temporal information during classification. This requires that the method is capable of storing a significant amount of past information to use in the classification task at each instant. If the method can keep track of temporal information, it is said to be dynamic; if not is it said to be static.

Online (O): A method may be implemented offline or online. Offline implies that the method stores information for later classification (i.e., not real time). Online means the MF assessment is made in real time.

Cross Subject (C). We say a method is considered cross-subject influence, if it explicitly took into consideration the individual physiological characteristics of each subject in the classification process. This is an important factor to consider, since MF limits and their representation in physiological signals can differ across subjects.

Besides the taxonomic decomposition of the state of the art described above, we also classified the surveyed papers according to the data processing pipeline presented in Section II by identifying the chosen approaches for data preprocessing, feature extraction, and classification algorithm. The results of this classification approach are presented in Section IV and are summarized in Tables I–III.

All the surveyed papers are reported according to the data domain, where the EEG data were analyzed during the MF assessment task. In each data domain table, papers are compared regarding the main characteristics of the MF assessment pipeline described in Section II, namely preprocessing, feature extraction, and classification algorithm, and the taxonomy presented in Section III. For the sake of a cleaner presentation, the taxonomy terms are represented by their initials in Tables I–III. The presence of a check mark in a taxonomy field indicates that the paper presents that desired characteristic.

Additionally, although the classification accuracy obtained by the authors in each paper is presented in these tables, they

TABLE I
TIME DOMAIN METHODS

Preprocessing	Feature extraction	Classification algorithm	N	G	D	O	C	Acc.	Ref.
Band-pass filter	Statistics	SDAR	✓	✓	✓	✓		95%	[58]
	PCA	OwARR, OwARR-SDS	✓			✓	✓	-	[56]
	SamEn, FuzEn, AppEn, SpecEn	AdaBoost	✓					97.5%	[55]
		DT	✓					95.7%	[54]
	SamEn, FuzEn, AppEn, SpecEn / Fisher distance	SVM	✓					98.8%	[53]
	Fuzzy Entropy / Fisher distance	SVM						85%	[52]
Band-pass filter, Notch filter	SamEn, FuzEn, AppEn, SpecEn	MLP	✓					98%	[46]
		RF	✓			✓	✓	96.6%	[47]
Band-pass filter, Visual inspection	DNTF	SVM	✓				✓	98%	[51]
ICA	AR modeling	BNN	✓					88.2%	[48]
		Sparse-DBN	✓	✓				93.1%	[49]
	Statistics	Dynamic-BNN	✓		✓	✓		95%	[50]
-	Meditation and Attention EEG	k-NN	✓					83.6%	[57]

TABLE II
FREQUENCY DOMAIN METHODS

Preprocessing	Feature extraction	Classification algorithm	N	G	D	O	C	Acc.	Ref.
Band-pass filter	PSD	SVM				✓		98.2%	[74]
		SVM				✓		96.2%	[75]
		RBF-SVR	✓			✓		-	[78]
		RSEFNN	✓		✓	✓	✓	-	[79]
Band-pass filter, Coherence method	PSD	ABSVM	✓			✓	✓	82.2%	[77]
Band-pass filter, Coherence method, ICA	PSD / Statistics	SDBN	✓	✓		✓	✓	77%	[59]
Band-pass filter, Notch filter, ICA	PSD	FLDA			✓		✓	80%	[60]
Band-pass filter, Visual inspection, ICA	PSD	Thresholding				✓	✓	76%	[61]
Band-pass filter, ICA	PSD	SDAE	✓	✓		✓	✓	85.6%	[62]
	PDC	SVM	✓					81.5%	[63]
ICA	FFT / NWFE	SVM	✓					88.7%	[64]
		RBFNN	✓			✓	✓	91.6%	[65]
	SPD / Statistics / SVM	SVM	✓				✓	75%	[66]
Low-pass filter, PCA	PSD	SVM	✓			✓	✓	80.4%	[67]
PCA	PSD	SVM						94%	[68]
De-noising wavelet	PSD	KPLS-DLR			✓	✓	✓	97%	[69]
		Sparse-KSVD				✓		95%	[70]
DWT	PSD	SVM	✓			✓		90.7%	[71]
Fisher bilateral test	PSD	Thresholding						98.3%	[72]
-	PSD	SVM	✓					86%	[76]
		temporal aggregation SVM	✓		✓	✓	✓	87%	[73]

should be analyzed with caution. Since most researchers in the field do not make use of benchmark datasets, it is not possible to compare the accuracy of different studies reliably. As an example, a case study that considers cross-subject classification is expected to have a lower classification accuracy than one that only considers single-subject classification. Publications presented in this survey without accuracy value use regression methods, and performance is evaluated by root-mean-square error instead of accuracy.

IV. STATE OF THE ART OF MF ASSESSMENT USING EEG

EEG data can be processed in different domains, including PCA and ICA [44]. In the present work, we divided the MF assessment method according to the three main data domains, namely time, frequency, and time–frequency.

EEG signals are obtained as time series, which are noisy, high dimensional, and nonstationary, have an explicit dependency on the time variable, and require the extraction of features from them to be invariant to translations in time [45]. So, although the natural domain of EEG is time, all the previously cited signal characteristics can make the analysis of data in the original domain quite challenging. Also, the EEG signals have clinical significance in different frequency bands (δ, θ, α, and β), which cannot be observed directly in the time domain.

To overcome these issues, some researchers opt to evaluate the EEG data in the frequency domain. This approach makes it possible to visualize the important frequency bands and avoid the problematic characteristics of time-series signals. The

TABLE III
TIME–FREQUENCY DOMAIN METHODS

Preprocessing	Feature extraction	Classification algorithm	N	G	D	O	C	Acc.	Ref.
Band-pass filter	Raw data, ICA	CCNN, CCNN-R	✓			✓	✓	76.7%	[90]
	WPT / Statistics / MI	SVM	✓			✓		98.6%	[91]
	DEn	SVR	✓		✓			85%	[89]
Band-pass filter, Notch filter	EMD	MLP	✓				✓	84.5%	[92]
	DWT	Thresholding				✓		85%	[93]
Band-pass filter, Visual inspection	Statistics, SamEn, PSD	SVM	✓				✓	80%	[81]
	KC / AppEn / PCA	SVM	✓					85%	[82]
Band-pass filter, Adaptive filter	Statistics / LDA	MLP	✓			✓		85.7%	[80]
Visual inspection	SamEn, AppEn, RenyiEn, RQA	ELM	✓			✓		97.3%	[83]
DWT	WEn	PCNN	✓			✓		97%	[86]
	Best m-term approximation	Thresholding				✓		98.7%	[87]
	Statistics	deep-LSTM	✓		✓			93%	[88]
	WEnS, PP-ApEnS, PP-SampEnS	MLP	✓			✓		96.5%	[84]
WPT	PSD	Thresholding				✓		-	[85]

time–frequency domain is another option that merges time and frequency domain characteristics by decomposing the original time series into one time series for each desired frequency band. This approach conserves the temporal characteristics of the original EEG signal, which can be valuable for certain assessment methods. The following sections discuss EEG-based MF assessment methods in each of these three domains.

A. Time Domain Methods

1) Preprocessing: Most time domain methods rely heavily on digital filtering as the main preprocessing approach. Band-pass filters are used to restrict the EEG signals to the frequency intervals of interest for MF analysis. Notch filters are applied to remove specific noise from the data, such as powerline noise [46], [47]. In order to remove blink, heart, and muscle artifacts from EEG data, ICA [48]–[50] and visual inspection [51] are also applied.

2) Feature Extraction: Several different types of entropy measures have been applied for feature extraction in the time domain. Entropy measures are popular due to the ability to measure the degree of uncertainty in unstable and nonlinear time series such as EEG signals. These measures can be used to compare normal and unsettled brain states. In some cases, different entropy measures are used in isolation [47], [52], but they can also be combined in order to improve the quality of extracted features [46], [53]–[55].

Complex EEG data from several channels can be represented in a simpler way by means of PCA, transforming the original set of inputs to a new set of coordinate systems that encapsulates the greatest amount of the original variance in the least number of new components as possible [56]. Alternatively, the complex input data can be factored using nonnegative decomposition methods, which factor the input data in meaningful components without applying any transformations to it [51].

An alternative for spectral analysis in the time domain is autoregressive (AR) modeling. It has been applied due to its ability to model the peak spectra of EEG signals and reportedly provides a better set of features then fast Fourier transform (FFT)-based methods [48], [49].

3) Classification: SVM is used as classifier for MF assessment problems, since it can group input elements in different classes. It is a kernel-based method, meaning that it can perform different linear [52] and nonlinear classifications [51], [53] by just changing its kernel function. Another algorithm that makes use of clustering is kNN. When kNN is used for a classification task, an input element is classified by a majority vote of its kNNs, where its class is assigned as the most common one among these neighbors [57].

Decision tree (DT) is a predictive model, where the branches and leaves structure represent, respectively, the classes features and labels. Thus, a label is represented by the conjunction of features that lead to it. It is a simple model capable of modeling large datasets with little data preparation well. When combined with boosting, it can perform well on noise datasets [54].

The main disadvantage of DT is its tendency to overfit to the training data. Random forest overcomes this flaw. It is an ensemble method that considers several DTs for the classification task. This approach provides a set of more robust features for training the classifier [47]. AdaBoost is another ensemble classification method, which weighs the classification results of multiple other classifiers and makes a decision based on the majority voting criteria. Hu [55] applied AdaBoost with SVM, DT, and Naive Bayes subclassifiers.

Neural networks are extensively used for MF assessment in time series, since different network structures can provide interesting properties to the classification algorithm. Good classification results can be obtained using all kinds of neural networks, depending on the specific necessities of each dataset. Even a simple and not very robust model, such as the multilayer perceptron (MLP), can achieve good classification accuracies if trained properly [46].

One of the crucial issues when using a neural network approach is to ensure that the learned structure can make a good generalization when data never seen before is presented to it. Chai *et al.* [48] propose the use of BNN as a classifier due to its ability to generalize the data analysis independently of how small or noisy the dataset is.

A neural network can learn very intricate features from the input data, including how the data structure is composed. One

example of such a model is a deep belief network (DBN), which is a nonlinear classification model composed of an unsupervised generative model and a supervised discriminative model. The model can be made more efficient by adding sparsity to the network, preventing it from over-fitting [49]. Some neural network models can also learn how to model temporal relationships of state variables, making the classification algorithm dynamic [50].

Besides classification, a regression approach can also be applied for MF detection. A sequential discounted autoregressive (SDAR) model of order N sequentially represents each data point in a time series as a combination of N previous point, with a discount factor for older points. When using statistical features, such a model can be trained to predict changes in EEG data with timing precision close to 150 ms [58]. Regression models can also be used for accounting for EEG differences among multiple subjects. Online weighted adaptation regularization for regression (OwARR) achieves this by online fusing data from old and new subjects, constantly adapting the classification model. Source domain selection (SDS) is applied to reduce the number of previous subjects needed to estimate new regression for real-time applications [56].

Table I presents the discussed time domain methods, comparing their main characteristics.

B. Frequency Domain Methods

1) Preprocessing: On the frequency domain, the preprocessing phase receives special attention from researchers. Besides the more common digital filters like bandpass and notch filters, specialized algorithms are frequently used for artifact rejection. Slow artifact rejection are especially relevant on frequency domain analysis, since they have a big impact in the spectral powers of δ, θ, α, and β subbands.

ICA is the most popular choice due to its simplicity and efficacy in rejecting EOG and movement artifacts [59]–[66]. It decomposes the original signal into independent components that can be inspected in order to reject the ones related to artifacts.

Other approaches similar to ICA present in the surveyed literature are PCA [67], [68] and denoising wavelets [69]–[71]. Both approaches consist of decomposing the original signal in a new set of components (principal components and subband intervals, respectively) and rejecting the components closely related to EOG, ECG, and movement artifacts. The Fisher bilateral test was applied as a thresholding method for artifact rejection [72]. In this case, the EEG data is compared to a 60-s long reference signal known to be artifact free.

2) Feature Extraction: By far the most common approach to feature extraction on the frequency domain is spectral analysis through FFT and PSD. Researchers are not only interested in the spectral powers for δ, θ, α, and β subbands, but also in the relation and ratios among these frequency bands, since they carry important information about MF state changes. Additionally, the ratios between spectral powers of different frequency bands help the classification model to account for cross-subject variations in the input data [60], [61], [72], [73].

In the literature, statistics are also applied together with spectral analysis for feature extraction. They can be fed to a classifier directly [59] or filtered by some kind of feature selection algorithm [66]. Nonparametric feature extraction was also applied due to the advantages of parametric models, such as better performance for nonnormal distributed data [64], [65].

3) Classification: Since most of the feature extraction methods on frequency domain rely on discrete spectral features, there is a large number of linear classifiers in the surveyed literature. The most common linear classifiers are SVM with linear kernel functions [68], [74], [75] and simple thresholding [61], [72]. Kernel partial least squares (KPLS) decomposition was used on the EEG data to find and select a reduced set of orthogonal components with maximum covariance. It was coupled with discrete-output linear regression (DLR) classifier to define a linear hyperplane capable of separating the MF states in the appropriate classes [69].

Yet in the realm of linear methods, k-singular value decomposition (KSVD) was applied to generate an overcomplete dictionary of signals that can be used to, sparsely, represent the input signal as a linear combination of the learnt signals [70]. Also, Fisher's linear discriminant analysis (FLDA) was used to find a linear combination of features that characterizes different MF states. This set of features can be applied as a linear classifier [60].

SVM is the most common method used for MF classification in the frequency domain. Besides its use with linear kernels, it was also extensively applied with nonlinear kernels, for more robust classification performance [63], [64], [66], [67], [71], [76]. Some more specialized versions of the traditional SVM algorithm were also used. Temporal aggregation SVM was applied in order to make the SVM algorithm dynamic [73]. Adaptive bounded SVM (ABSVM) can be used for two reasons: the bounded part optimizes the SVM training procedure when more than two MF states are considered and the adaptive part optimizes the results for cross-section and cross-subject classification problems [77]. Support vector regression (SVR) is a variation of SVM that performs regression instead of classification and was used to create an MF predictor [78].

Neural network models were also found in the surveyed literature. Sparse DBN [59] and sparse deep auto encoders [62] were used to create generative models capable of accounting for, respectively, cross-subject and cross-session variability. Radial basis function neural network (RBFNN) is basically an MLP with exactly one hidden layer and uses radial basis functions as activation functions. It was applied to construct a nonlinear MF state classifier due to its training and classification performances [65]. A recurrent self-evolving fuzzy neural network (RSEFNN) was used to create a dynamic regression tool capable of accounting for cross-subject regression of MF states [79].

Table II presents the discussed frequency domain methods, comparing their main characteristics.

C. Time–Frequency Domain Methods

1) Preprocessing: Researchers typically focus on two approaches to the preprocessing phase in the time–frequency

domain: digital filtering and wavelets. As seen in the previous section, bandpass and notch filters are largely used as noise and artifact removal tools. Adaptive filters are also viable digital options for noise and artifact rejection [80]. Experts also use visual inspection to remove artifacts [81], [82]. This approach is especially effective when relying on data from other sensors, such as EMG and EOG [83].

Wavelet transforms will decompose a time series into a set of components with different frequency bands. They are effective for analyzing nonstationary signals, since they can represent trends, discontinuities, and patterns in the original signal very well [84]. Wavelet packet transform (WPT) is the simplest wavelet transform, disregarding boundary treatments in the original signal [85]. DWTs are most commonly used due to their more robust performance [86]–[88].

2) Feature Extraction: One of the reasons to conduct EEG analysis in the time–frequency domain is to work with nonlinear features for the MF state classification. Usually, authors opt for different entropy and complexity measures to capture the nonstationarity and nonlinearity of EEG signals. The separation of the original signal in its main subbands seems to improve the nonlinear features performance. Some of the applied measures include differential entropy (DEn) [89], AppEn, SampEn, Renyi entropy (RenEn), recurrence quantification analysis (RQA) [83], wavelet entropy (WEn) [86], and Kolmogorov complexity (KC) [90]. The use of sliding windows for the calculation of entropy measures can be applied for real-time MF detection [84].

When working on the time–frequency domain, most authors opt to use some type of wavelet transformation to convert the EEG time series to the time–frequency domain. When doing this, da Silveira *et al.* [87] used best *m*-term approximation to select the wavelet decomposition terms with the biggest influence in alpha and beta subbands. Kaur and Singh [92] opted for a different approach to make the EEG data domain transition. They applied the empirical mode decomposition method (EMD) to extract intrinsic mode functions from EEG signals.

Instead of focusing the MF detection method on only one domain, some authors try to expand their possibilities by transitioning between different domains during the EEG data analysis. Correa *et al.* [80] made use of statistical features in the time, frequency, and time–frequency domain to assess drivers drowsiness state. Lee *et al.* [91] extracted 51 statistical, frequency, and interval features from EEG and respiration signals, and performed feature selection using mutual information (MI).

Since time–frequency domain signals contain all frequency domain information available, the use of spectral analyzing on the decomposed signals by deriving power-based indices is still a viable option [85].

3) Classification: Most of the works surveyed on the time–frequency domain use some sort of entropy, complexity, or statistical measures as features. Therefore, the classification algorithms need to distinguish among different MF states based on a set of discrete measures. Three approaches to this task dominate thresholding, SVM, and neural networks.

Thresholding classifiers are the simplest. They are trained to find the limits for the features that define each MF state. They are mostly linear and were used on spectral features [85] and on the decomposed time series [87], [93]. As in the frequency domain, SVM is extensively applied as a clustering method, to group the input signals in different MF state classes based on nonlinear features. Only nonlinear kernels were used for SVM in the surveyed works [81], [82], [91]. SVR was also used together with a continuous conditional neural field and a continuous conditional random field to produce a dynamic estimator of the MF state [89].

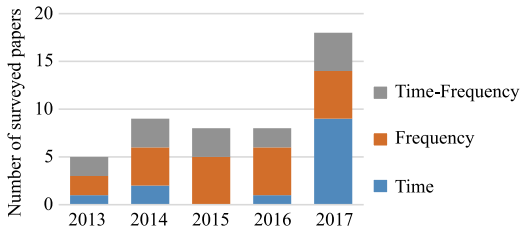


Fig. 4. Number of surveyed papers per year and domain.

One of the most simple neural network models is called single layer perceptron (SLP). Although SLP is not very useful for complex classification problems by itself, a variation of SLP called extreme learning machine (ELM) was successfully applied for MF state classification due to its ability to avoid local optima and fast training speed [83]. An MLP classifier was used in several works with good results [80], [84], [92] although more complex and robust neural network model was also used. Deep long short-term memory (LSTM) was used to construct a dynamic classifier to account for the fact that MF has a very important temporal factor, since it builds up over time [88]. Channelwise, a convolutional neural network (CNN) was used to automatically extract a complex feature from time-series data [90]. Such features are robust enough to provide effective cross-subject classification of the MF state.

Table III presents the discussed wavelet domain methods, comparing their main characteristics.

V. Discussion and Future Trends

The number of published works using EEG to assess MF has steadily increased in the recent years, as shown in Fig. 4. This increase reflects a change of paradigm in human–machine systems as machinery systems become increasingly reliable and consequently human operators account for a steady increase in accidents. This increase also shows that higher levels of automation in several industries, such as automotive [94] and maritime [95], have not made the topic less relevant. This is because in most cases automation does not remove the human element completely from the loop, but just reallocates it to a different role.

Fig. 4 also shows a big shift regarding the data domains used for analyzing the EEG data. The time domain, which was barely used in previous years, came to dominate in 2017. This most likely indicates the development of methods capable of dealing with time series and its particularities. including nonlinearity, high levels of noise, high dimensionality, and nonstationarity [45]). Some methods capable of dealing with these special

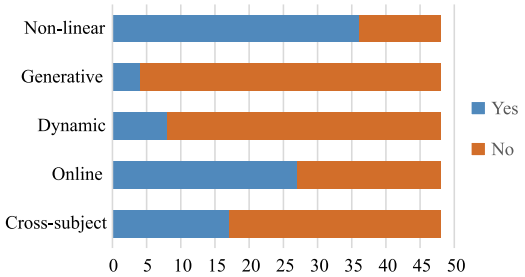


Fig. 5. Taxonomy distribution for the state of the art.

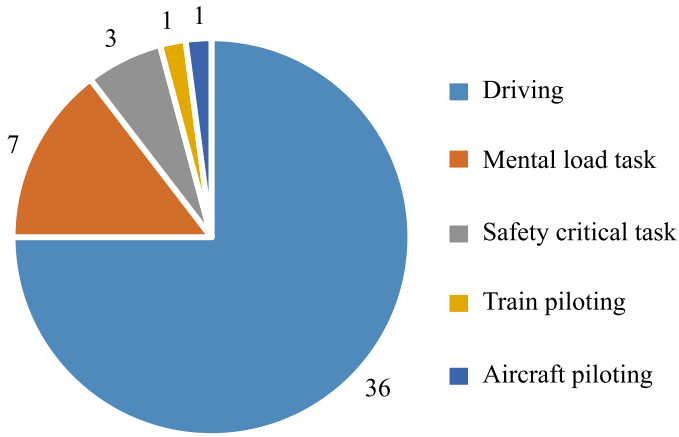


Fig. 6. Application areas distribution for the state of the art.

requirements from time series include deep learning methods, such as CNN, DBN, and auto encoders.

An overview of the taxonomic distribution can be seen in Fig. 5. Most of the taxonomic characteristics have an undeniable tendency across the field, which may indicate either a conformation about how the MF assessment methods should be structured or a possible turning point for further improvement. We can see that the vast majority of the papers applied nonlinear methods for handling EEG data. This is expected, since measuring linear features that can correctly represent time-series data can be very tricky, due to the intrinsic nature of this kind of signal. The greatest number of papers also use a discriminative approach, since most generative models are deep-learning based, and this kind of architecture has not been much explored in the field of MF assessment with EEG, although as noted there has been a shift toward deeper models. Regarding the dynamics, most papers use a static approach, since keeping track of temporal data to assess the MF state is demanding, especially for real-time applications. With the development of models such as LSTM, a dynamic approach is becoming more feasible, which is very relevant, since the MF state is a built-up process, where the tiredness accumulates over time, making this temporal dependency very relevant for an accurate analysis. The cross-subject aspect was evaluated by few researchers, although it is known that the EEG signals and MF state are subject-dependent, presenting variation among people. Future research should address this point. Regarding implementation, fortunately, almost 50% of the published work in the past five years presented methods capable of being implemented in real time, which is essential for real-life application. Usually this kind of online method uses little to no preprocessing, a simpler set of features, and a small amount of EEG channels in order to reduce the computational requirement and hardware footprint as much as possible.

The application areas where MF detection using EEG has been applied are driving, mental load tasks [51], [63], [68], [69], [82], [83], [88], safety-critical tasks [59], [62], [77], train piloting [71], and aircraft piloting [72]. Fig. 6 shows that driving tasks dominate as a case study for most of the published work. This reflects the automotive industry's efforts to provide systems to detect drowsiness in drivers, which have been implemented by several automobile manufacturers, including Audi, Volkswagen, Volvo, BMW, and Mercedes-Benz. Most of the systems available today use driving patterns, such as steering pattern, vehicle lane positioning, and acceleration and brake pattern [96] to identify fatigued drivers. Although some systems use driver's eyes and face monitoring to assess drowsiness [97], no commercial system the car industry has employed currently applies physiological signals to detect MF. Implementing a system that monitors physiological signals that can be used in real-life applications for detecting MF is the next step in accident avoidance in human–machine systems, and the automotive industry seems to be the most likely to achieve this first. Although other fields need such systems, minimal effort has been put into developing them, which provides an opportunity for researchers in other areas to develop new methods and tools for real-time MF detection.

A. *Feasibility of EEG-Based MF Detection Systems*

Several EEG-based MF detection methods can achieve classification accuracy over 90%. Although impressive, these numbers hide important limitations of current approaches that make the transition from research laboratories to real-life applications very difficult. These important points to consider are discussed below.

1) Cross-Subject and Cross-Session MF Detection: EEG signals are very sensitive electrical signals and can behave differently from person to person, so MF detection methods need to be robust enough to handle these variations. This robustness is also necessary to ensure a good generalization capability needed to handle subjects outside the training samples. On the surveyed literature, cross-subject and cross-session variability is most commonly approached in the frequency domain. The advantage of frequency domain methods is the use of spectral power ratios, which help to compensate for different ranges of theta, alpha, and beta subband activities in different individuals. Some entropy [47], regression [56], and neural network [90], [92] based methods were attempted on time and time–frequency domains, usually with similar accuracy ratios to frequency domain methods.

2) Computational Requirements: The process of MF assessment should be as time and energy efficient as possible. It should not rely on expensive computations that require a powerful computer to complete. Basically, any of the surveyed methods that can be considered online present a reasonable level of computational requirement for real-life implementation. Usually the time-consuming part of these algorithms is the training phase, but once the algorithm is trained, its application to new data is fast.

3) Portability: The hardware carrying the MF assessment algorithm should be as compact as possible. It should be portable, allowing the user to move around with no restrictions while providing long periods of battery autonomy. On the surveyed literature, portability is achieved using the Android platform [57], [75], [78], [91]. Computationally light MF assessment algorithms are implemented on mobile devices, receiving EEG data in real time via Bluetooth communication.

4) Intrusiveness: The MF assessment system should be as nonintrusive as possible so as not to interfere in any way with the performance of the user. If not properly designed, the system can cause distress to the operation. A main factor to improve is the number of electrodes used to acquire the EEG signal, balancing the tradeoff between number of electrodes and precision of MF detection. In the surveyed literature, we considered nonintrusive algorithms that rely on only EEG sensor and consider just few channels. In several studies, authors investigated the use of the simplest possible model, using only one EEG channel [47], [73], [74], [80], [85], [87], making the installation of electrodes as fast and simple as possible and reducing a lot of the computational requirements.

5) Number of MF States: MF develops as an accumulative process. Most published works use only two MF states to assess MF. The use of intermediate states between "no fatigue" and "fatigue" can help with the correct assessment of MF, since more gradual change between states can ensure better accuracy and greater response time for safety–critical systems. On the surveyed literature, few authors explored the use of intermediate MF states. We can identify the studies using three [59], [76], [89] or four [67], [77], [91] different MF states. Some methods capable of performing regression can present an almost continuous MF development output [78], [79].

6) Closed-Loop System: The assessment of MF is as important as what the system does with this information. There is a need to develop efficient ways to mitigate the effects of MF and to alert the user about the dangers in the operation during the MF state. The MF detection system needs to be capable of acting before any accident happens, in a preventive way. On the surveyed literature, several authors implemented closed-loop systems, but almost all of them used driving as a case study. In most cases, the fatigue threshold detection module sends a sonorous or visual warning feedback to alert the user about dangerous driving conditions [66], [67], [75], [78], [86], [91], [93]. A vehicle speed control model based on an MF assessment algorithm was successfully implemented in [70]. There was only one work that considered a closed-loop system in an application area other than driving, where the author implemented an MF warning feedback for train pilots using sonorous alarm and a massage chair [71].

VI. Conclusion

Following the continuous increase in quality of hardware and software present in all kinds of machine systems, the role of the human factor and human failure in accidents in human–machine systems has become more evident. One of the biggest causes of human failure lies in excessive MF, which can lead to drowsiness, lack of situational awareness, and slower response to external stimulus. When developing a system that can assess MF and warn the human operator about its critical condition, the use of EEG is recognized as the most reliable way to implement such a system.

This survey paper reviewed the current state of the art of the field of EEG-based MF detection systems. The fundamentals of data acquisition and interpretation were approached. The underlying structure of a typical EEG-based MF detection method was discussed and the most common preprocessing, feature extraction, and classification algorithms were presented and briefly discussed. The main goal of this paper is to provide an overview of the current trends in the area. To do this, we tried very hard to be inclusive of relevant papers published in the past five years. In this review, we discussed these papers' approaches, characteristics, and results based on their application domains, which we divided into time, frequency, and time–frequency domains.

In the final portion of the paper, we discussed the current state of the art using the presented taxonomy as a basis for discussion. The main points of the MF detection methods architectures were approached, and their current usage and future trends were briefly evaluated. There is a lot of opportunity to develop MF detection systems for applications other than driving, and now might be the right moment to put more emphasis into deep learning models that have been barely used to date, always keeping in mind the important role of online models in making real-world applications feasible.

Finally, from our survey study, we recommend the reader to take a closer look at Trejo *et al.* [69]. In this work, the authors implemented an MF detection model based on KPLS-DLR that meets most of the desired criteria for a feasible MF detection system in real life. Their model is dynamic, feasible for online implementation, and robust for cross-subject classification. The system is very minimally intrusive, using only two EEG channels and two EOG channels for artifact rejection. The mathematical algorithm implemented is simple but has outstanding performance (97% accuracy).

References

[1] Z. Yang, J. Wang, and K. Li, "Maritime safety analysis in retrospect," *Maritime Policy Manag.*, vol. 40, no. 3, pp. 261–277, 2013.

[2] K. Avers and W. B. Johnson, "A review of federal aviation administration fatigue research: Transitioning scientific results to the aviation industry," *Aviation Psychol. Appl. Human Factors*, vol. 1, no. 2, 2011, Art. no. 87.

[3] A. Williamson, D. A. Lombardi, S. Folkard, J. Stutts, T. K. Courtney, and J. L. Connor, "The link between fatigue and safety," *Accident Anal. Prevention*, vol. 43, no. 2, pp. 498–515, 2011.

[4] C.-H. Ting *et al.*, "Real-time adaptive automation system based on identification of operator functional state in simulated process control operations," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 2, pp. 251–262, Mar. 2010.

[5] D. San Kim and W. C. Yoon, "An accident causation model for the railway industry: Application of the model to 80 rail accident investigation reports from the UK," *Saf. Sci.*, vol. 60, pp. 57–68, 2013.

[6] P. Suraweera, G. I. Webb, I. Evans, and M. Wallace, "Learning crew scheduling constraints from historical schedules," *Transp. Res. Part C, Emerg. Technol.*, vol. 26, pp. 214–232, 2013.

[7] L. Giraudet, J.-P. Imbert, M. Bérenger, S. Tremblay, and M. Causse, "The neuroergonomic evaluation of human machine interface design in air traffic control using behavioral and EEG/ERP measures," *Behav. Brain Res.*, vol. 294, pp. 246–253, 2015.

[8] L. Reinerman-Jones, G. Matthews, and J. E. Mercado, "Detection tasks in nuclear power plant operation: Vigilance decrement and physiological workload monitoring," *Saf. Sci.*, vol. 88, pp. 97–107, 2016.

[9] C. Chauvin, S. Lardjane, G. Morel, J.-P. Clostermann, and B. Langard, "Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS," *Accident Anal. Prevention*, vol. 59, pp. 26–37, 2013.

[10] G. R. J. Hockey, *Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks*, vol. 355. Amsterdam, The Netherlands: IOS Press, 2003.

[11] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Adv. psychol.*, vol. 52, pp. 139–183, 1988.

[12] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *Int. J. Neurosci.*, vol. 52, no. 1–2, pp. 29–37, 1990.

[13] M. W. Johns, "A new method for measuring daytime sleepiness: The epworth sleepiness scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[14] T. Chalder *et al.*, "Development of a fatigue scale," *J. Psychosomatic Res.*, vol. 37, no. 2, pp. 147–153, 1993.

[15] P. M. Forsman, B. J. Vila, R. A. Short, C. G. Mott, and H. P. Van Dongen, "Efficient driver drowsiness detection at moderate levels of drowsiness," *Accident Anal. Prevention*, vol. 50, pp. 341–350, 2013.

[16] B.-C. Yin, X. Fan, and Y.-F. Sun, "Multiscale dynamic features based driver fatigue detection," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 03, pp. 575–589, 2009.

[17] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *Sensors*, vol. 12, no. 12, pp. 16 937–16 953, 2012.

[18] W. Zhu, H. Yang, Y. Jin, and B. Liu, "A method for recognizing fatigue driving based on Dempster-Shafer theory and fuzzy neural network," *Math. Problems Eng.*, vol. 2017, 2017, Art. no. 6191035.

[19] V. Menon, S. Rivera, C. White, G. Glover, and A. Reiss, "Dissociating prefrontal and parietal cortex activation during arithmetic processing," *Neuroimage*, vol. 12, no. 4, pp. 357–365, 2000.

[20] S.-W. Chuang, L.-W. Ko, Y.-P. Lin, R.-S. Huang, T.-P. Jung, and C.-T. Lin, "Co-modulatory spectral changes in independent brain processes are correlated with task performance," *Neuroimage*, vol. 62, no. 3, pp. 1469–1477, 2012.

[21] B. He, S. Gao, H. Yuan, and J. R. Wolpaw, "Brain–computer interfaces," in *Proc. Neural Eng.*, 2013, pp. 87–151.

[22] J. Liu, C. Zhang, and C. Zheng, "EEG-based estimation of mental fatigue by using KPCA–HMM and complexity parameters," *Biomed. Signal Process. Control*, vol. 5, no. 2, pp. 124–130, 2010.

[23] H. H. Jasper, "The ten twenty electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 10, pp. 371–375, 1958.

[24] G. H. Klem *et al.*, "The ten-twenty electrode system of the international federation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 52, no. 3, pp. 3–6, 1999.

[25] E. Suarez, M. Viegas, M. Adjouadi, and A. Barreto, "Relating induced changes in EEG signals to orientation of visual stimuli using the ESI-256 machine," *Biomed. Sci. Instrum.*, vol. 36, pp. 33–38, 2000.

[26] G. Chatrian, E. Lettich, and P. Nelson, "Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities," *Amer. J. EEG Technol.*, vol. 25, no. 2, pp. 83–92, 1985.

[27] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clin. Neurophysiol.*, vol. 112, no. 4, pp. 713–719, 2001.

[28] D. P. Subha, P. K. Joseph, R. Acharya, and C. M. Lim, "EEG signal analysis: A survey," *J. Med. Syst.*, vol. 34, no. 2, pp. 195–212, 2010.

[29] A. Craig, Y. Tran, N. Wijesuriya, and H. Nguyen, "Regional brain wave activity changes associated with fatigue," *Psychophysiology*, vol. 49, no. 4, pp. 574–582, 2012.

[30] A. Kandaswamy, V. Krishnaveni, S. Jayaraman, N. Malmurugan, and K. Ramadoss, "Removal of ocular artifacts from EEG—A survey," *IETE J. Res.*, vol. 51, no. 2, pp. 121–130, 2005.

[31] L. Bi, X.-A. Fan, and Y. Liu, "EEG-based brain-controlled mobile robots: A survey," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 2, pp. 161–176, Mar. 2013.

[32] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[33] H. Nolan, R. Whelan, and R. Reilly, "Faster: Fully automated statistical thresholding for EEG artifact rejection," *J. Neurosci. Methods*, vol. 192, no. 1, pp. 152–162, 2010.

[34] T.-P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T.-W. Lee, and T. J. Sejnowski, "Imaging brain dynamics using independent component analysis," *Proc. IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.

[35] P. Stoica *et al.*, *Spectral Analysis of Signals*, vol. 1. Englewood Cliffs, NJ, USA: Prentice-Hall, 2005.

[36] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang, "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Trans. Instrum. Meas.*, vol. 61, no. 6, pp. 1649–1657, Jun. 2012.

[37] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, and J. S. Suri, "Automated diagnosis of epileptic EEG using entropies," *Biomed. Signal Process. Control*, vol. 7, no. 4, pp. 401–408, 2012.

[38] L. Guo, D. Rivero, and A. Pazos, "Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks," *J. Neurosci. Methods*, vol. 193, no. 1, pp. 156–163, 2010.

[39] Z. Mu, J. Hu, and J. Min, "EEG-based person authentication using a fuzzy entropy-related approach with two electrodes," *Entropy*, vol. 18, no. 12, 2016, Art. no. 432.

[40] Y. Song, J. Crowcroft, and J. Zhang, "Automatic epileptic seizure detection in EEGs based on optimized sample entropy and extreme learning machine," *J. Neurosci. Methods*, vol. 210, no. 2, pp. 132–146, 2012.

[41] N. Kannathal, M. L. Choo, U. R. Acharya, and P. Sadasivan, "Entropies for detection of epilepsy in EEG," *Comput. Methods Programs Biomed.*, vol. 80, no. 3, pp. 187–194, 2005.

[42] A. Soria-Frisch, "A critical review on the usage of ensembles for BCI," in *Proc. Towards Practical Brain-Comput. Interfaces*, 2012, pp. 41–65.

[43] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The prisma statement," *Ann. Internal Med.*, vol. 151, no. 4, pp. 264–269, 2009.

[44] T. N. Alotaiby, S. A. Alshebeili, T. Alshawi, I. Ahmad, and F. E. A. El-Samie, "EEG seizure detection and prediction algorithms: A survey," *EURASIP J. Adv. Signal Process.*, vol. 2014, no. 1, 2014, Art. no. 183.

[45] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, 2014.

[46] J. Min, P. Wang, and J. Hu, "Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system," *PLoS One*, vol. 12, no. 12, 2017, Art. no. e0188756.

[47] J. Hu, "Comparison of different features and classifiers for driver fatigue detection based on a single EEG channel," *Comput. Math. Methods Med.*, vol. 2017, 2017, Art. no. 5109530.

[48] R. Chai *et al.*, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 715–724, May 2017.

[49] R. Chai *et al.*, "Improving EEG-based driver fatigue classification using sparse-deep belief networks," *Front. Neurosci.*, vol. 11, 2017.

[50] Q. He, W. Li, X. Fan, and Z. Fei, "Driver fatigue evaluation model with integration of multi-indicators based on dynamic Bayesian network," *IET Intell. Transp. Syst.*, vol. 9, no. 5, pp. 547–554, 2014.

[51] F. Razavipour, R. Boostani, S. Kouchaki, and S. Afrasiabi, "Comparative application of non-negative decomposition methods in classifying fatigue and non-fatigue states," *Arabian J. Sci. Eng.*, vol. 39, no. 10, pp. 7049–7058, 2014.

[52] Z. Mu, J. Hu, and J. Yin, "Driving fatigue detecting based on EEG signals of forehead area," *Int. J. Pattern Recognit. Arti. Intell.*, vol. 31, no. 05, 2017, Art. no. 1750011.

[53] Z. Mu, J. Hu, and J. Min, "Driver fatigue detection system using electroencephalography signals based on combined entropy features," *Appl. Sci.*, vol. 7, no. 2, 2017, Art. no. 150.

[54] J. Hu and P. Wang, "Noise robustness analysis of performance for EEG-based driver fatigue detection using different entropy feature sets," *Entropy*, vol. 19, no. 8, 2017, Art. no. 385.

[55] J. Hu, "Automated detection of driver fatigue based on AdaBoost classifier with EEG signals," *Front. Comput. Neurosci.*, vol. 11, 2017, Art. no. 72.

[56] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OWARR)," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1522–1535, Dec. 2017.

[57] J. He, Y. Zhang, C. Zhang, M. Zhou, and Y. Han, "A noninvasive real-time solution for driving fatigue detection based on left prefrontal EEG and eye blink," in *Proc. Int. Conf. Brain Health Inform*, 2016, pp. 325–335.

[58] V. Lawhern, S. Kerick, and K. A. Robbins, "Detecting alpha spindle events in EEG time series using adaptive autoregressive models," *BMC Neurosci.*, vol. 14, no. 1, 2013, Art. no. 101.

[59] Z. Yin and J. Zhang, "Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights," *Neurocomputing*, 2017.

[60] T. Nguyen, S. Ahn, H. Jang, S. C. Jun, and J. G. Kim, "Utilization of a combined EEG/NIRS system to predict driver drowsiness," *Sci. Rep.*, vol. 7, 2017, Art. no. 43933.

[61] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, "Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and FNIRS data," *Front. Human Neurosci.*, vol. 10, 2016, Art. no. 219.

[62] Z. Yin and J. Zhang, "Cross-session classification of mental workload levels using EEG and an adaptive deep learning model," *Biomed. Signal Process. Control*, vol. 33, pp. 30–47, 2017.

[63] Y. Sun, J. Lim, J. Meng, K. Kwok, N. Thakor, and A. Bezerianos, "Discriminative analysis of brain functional connectivity patterns for mental fatigue classification," *Ann. Biomed. Eng.*, vol. 42, no. 10, pp. 2084–2094, 2014.

[64] C.-H. Chuang, C.-S. Huang, L.-W. Ko, and C.-T. Lin, "An EEG-based perceptual function integration network for application to drowsy driving," *Knowl.-Based Syst.*, vol. 80, pp. 143–152, 2015.

[65] C.-H. Chuang, L.-W. Ko, Y.-P. Lin, T.-P. Jung, and C.-T. Lin, "Independent component ensemble of EEG for brain–computer interface," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 2, pp. 230–238, Mar. 2014.

[66] S. Hu, G. Zheng, and B. Peters, "Driver fatigue detection from electroencephalogram spectrum after electrooculography artefact removal," *IET Intell. Transp. Syst.*, vol. 7, no. 1, pp. 105–113, 2013.

[67] L. Cao, J. Li, Y. Xu, H. Zhu, and C. Jiang, "A hybrid vigilance monitoring study for mental fatigue and its neural activities," *Cogn. Comput.*, vol. 8, no. 2, pp. 228–236, 2016.

[68] F. Laurent *et al.*, "Multimodal information improves the rapid detection of mental fatigue," *Biomed. Signal Process. Control*, vol. 8, no. 4, pp. 400–408, 2013.

[69] L. J. Trejo, K. Kubitz, R. Rosipal, R. L. Kochavi, and L. D. Montgomery, "EEG-based estimation and classification of mental fatigue," *Psychology*, vol. 6, no. 05, 2015, Art. no. 572.

[70] Z. Zhang *et al.*, "A vehicle active safety model: Vehicle speed control based on driver vigilance detection using wearable EEG and sparse representation," *Sensors*, vol. 16, no. 2, 2016, Art. no. 242.

[71] X. Zhang *et al.*, "Design of a fatigue detection system for high-speed trains based on driver vigilance using a wireless wearable EEG," *Sensors*, vol. 17, no. 3, 2017, Art. no. 486.

[72] F. Sauvet *et al.*, "In-flight automatic detection of vigilance states using a single EEG channel," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 12, pp. 2840–2847, Dec. 2014.

[73] F. Rohit, V. Kulathumani, R. Kavi, I. Elwarfalli, V. Kecojevic, and A. Nimbarte, "Real-time drowsiness detection using wearable, lightweight brain sensing headbands," *IET Intell. Transp. Syst.*, vol. 11, no. 5, pp. 255–263, 2017.

[74] S. N. Resalat and V. Saba, "A practical method for driver sleepiness detection by processing the EEG signals stimulated with external flickering light," *Signal, Image Video Process.*, vol. 9, no. 8, pp. 1751–1757, 2015.

[75] G. Li and W.-Y. Chung, "A context-aware EEG headset system for early detection of driver drowsiness," *Sensors*, vol. 15, no. 8, pp. 20 873–20 893, 2015.

[76] M. Guo, S. Li, L. Wang, M. Chai, F. Chen, and Y. Wei, "Research on the relationship between reaction ability and mental state for online assessment of driving fatigue," *Int. J. Environ. Res. Public Health*, vol. 13, no. 12, 2016, Art. no. 1174.

[77] J. Zhang, Z. Yin, and R. Wang, "Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 2, pp. 200–214, Apr. 2015.

[78] C.-T. Lin *et al.*, "Wireless and wearable EEG system for evaluating driver vigilance," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 2, pp. 165–176, Apr. 2014.

[79] Y.-T. Liu, Y.-Y. Lin, S.-L. Wu, C.-H. Chuang, and C.-T. Lin, "Brain dynamics in predicting driving fatigue using a recurrent self-evolving fuzzy neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 2, pp. 347–360, Feb. 2016.

[80] A. G. Correa, L. Orosco, and E. Laciar, "Automatic detection of drowsiness in EEG records based on multimodal analysis," *Med. Eng. Phys.*, vol. 36, no. 2, pp. 244–249, 2014.

[81] M. Awais, N. Badruddin, and M. Drieberg, "A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability," *Sensors*, vol. 17, no. 9, 2017, Art. no. 1991.

[82] Y. J. Xiong, R. Zhang, C. Zhang, and X. L. Yu, "A novel estimation method of fatigue using EEG based on KPCA-SVM and complexity parameters," in *Proc. Appl. Mech. Mat.*, vol. 373, 2013, pp. 965–969.

[83] L.-l. Chen, Y. Zhao, J. Zhang, and J.-Z. Zou, "Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 7344–7355, 2015.

[84] C. Zhang, H. Wang, and R. Fu, "Automated detection of driver fatigue based on entropy and complexity measures," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 168–177, Feb. 2014.

[85] T. da Silveira, A. d. J. Kozakevicius, and C. R. Rodrigues, "Automated drowsiness detection through wavelet packet analysis of a single EEG channel," *Expert Syst. Appl.*, vol. 55, pp. 559–565, 2016.

[86] H. Wang, C. Zhang, T. Shi, F. Wang, and S. Ma, "Real-time EEG-based detection of fatigue driving danger for accident prediction," *Int. J. Neural Syst.*, vol. 25, no. 02, 2015, Art. no. 1550002.

[87] T. da Silveira, A. d. J. Kozakevicius, and C. R. Rodrigues, "Drowsiness detection for single channel EEG by DWT best m-term approximation," *Res. Biomed. Eng.*, vol. 31, no. 2, pp. 107–115, 2015.

[88] R. G. Hefron, B. J. Borghetti, J. C. Christensen, and C. M. S. Kabban, "Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation," *Pattern Recognit. Lett.*, vol. 94, pp. 96–104, 2017.

[89] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *J. Neural Eng.*, vol. 14, no. 2, 2017, Art. no. 026017.

[90] M. Hajinoroozi, Z. Mao, T.-P. Jung, C.-T. Lin, and Y. Huang, "EEG-based prediction of driver's cognitive performance by deep convolutional neural network," *Signal Process., Image Commun.*, vol. 47, pp. 549–555, 2016.

[91] B.-G. Lee, B.-L. Lee, and W.-Y. Chung, "Mobile healthcare for automatic driving sleep-onset detection using wavelet-based EEG and respiration signals," *Sensors*, vol. 14, no. 10, pp. 17 915–17 936, 2014.

[92] R. Kaur and K. Singh, "Drowsiness detection based on EEG signal analysis using EMD and trained neural network," *Int. J. Sci. Res*, vol. 10, pp. 157–161, 2013.

[93] M. Pathak and A. Jayanthy, "Development of a real-time single channel brain–computer interface system for detection of drowsiness," *Biomed. Eng., Appl., Basis Commun.*, vol. 29, no. 03, 2017, Art. no. 1750019.

[94] M. Gerla, E.-K. Lee, G. Pau, and U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds," in *Proc. IEEE World Forum Internet Things*, 2014, pp. 241–246.

[95] S. Ahvenjärvi, "The human element and autonomous ships," *TransNav: Int. J. Marine Navigation Saf. Sea Transp.*, vol. 10, pp. 517–521, 2016.

[96] A. Čolić, O. Marques, and B. Furht, *Driver Drowsiness Detection: Systems and Solutions*. London, U.K.: Springer, 2014.

[97] D. J. Walger, T. P. Breckon, A. Gaszczak, and T. Popham, "A comparison of features for regression-based driver head pose estimation under varying illumination conditions," in *Proc. IEEE Int. Workshop Comput. Intell. Multimedia Understanding.*, 2014, pp. 1–5.

A Context-Supported Deep Learning Framework for Multimodal Brain Imaging Classification

Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong

***Abstract*—Over the past decade, "content-based" multimedia systems have realized success. By comparison, brain imaging and classification systems demand more efforts for improvement with respect to accuracy, generalization, and interpretation. The relationship between electroencephalogram (EEG) signals and corresponding multimedia content needs to be further explored. In this paper, we integrate implicit and explicit learning modalities into a context-supported deep learning framework. We propose an improved solution for the task of brain imaging classification via EEG signals. In our proposed framework, we introduce a consistency test by exploiting the context of brain images and establishing a mapping between visual-level features and cognitive-level features inferred based on EEG signals. In this way, a multimodal approach can be developed to deliver an improved solution for brain imaging and its classification based on explicit learning modalities and research from the image processing community. In addition, a number of fusion techniques are investigated in this work to optimize individual classification results. Extensive experiments have been carried out, and their results demonstrate the effectiveness of our proposed framework. In comparison with the existing state-of-the-art approaches, our proposed framework achieves superior performance in terms of not only the standard visual object classification criteria, but also the exploitation of transfer learning. For the convenience of research dissemination, we make the source code publicly available for downloading at GitHub (https://github.com/aneeg/dual-modal-learning).**

***Index Terms*—Deep learning, electroencephalogram (EEG), explicit learning modality, implicit learning modality, object classification.**

Manuscript received September 15, 2018; revised January 8, 2019; accepted February 24, 2019. Date of publication April 2, 2019; date of current version November 21, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61620106008, in part by the the Natural Science Foundation of Guangdong Province under Grant 2016A030310053, in part by the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant JCYJ20160226191842793, in part by the Shenzhen high-level overseas talents program, in part by the National Engineering Laboratory for Big Data System Computing Technology, and in part by the Inlife-Handnet Open Fund. This paper was recommended by Associate Editor J. Han. *(Jianmin Jiang and Ahmed Fares are co-first authors.) (Corresponding author: Sheng-Hua Zhong.)*

J. Jiang and S.-H. Zhong are with the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China (e-mail: jianmin.jiang@szu.edu.cn; csshzhong@szu.edu.cn).

A. Fares is with the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Department of Electrical Engineering, the Computer Engineering branch, Faculty of Engineering at Shoubra, Benha University, Cairo 2900, Egypt (e-mail: ahmed.fares@szu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2904615

I. Introduction

HUMAN brain analytics has long been researched across a number of communities, including neural science, brain science, psychology, etc. At present, the brain interface is primarily examined via two approaches, functional magnetic resonance imaging (fMRI) and electroencephalograms (EEGs), where an EEG is a recording of voltage fluctuations produced by ionic current flows in the neurons of the brain [1]. While reflecting the brain's spontaneous electrical activities, an EEG also has the potential to provide a subjective response based on individual experiences [2], [3]. As a noninvasive brain signaling technique, an EEG provides high spatiotemporal resolution data, presenting a vivid reflection of the dynamics of the brain [4], which makes it ideal for a variety of research fields, such as brain-computer interface (BCI) [5]–[7], affective state recognition [8], [9], and diagnosis of brain-related diseases, such as epilepsy [10], [11], Alzheimer's [12], [13], Parkinson's etc. [14].

For the past decades, an understanding of EEG data evoked by specific stimuli/objects has been the primary goal of BCI and other important EEG-related research fields. Hence, EEG-based object/event classification becomes a key component across all these communities. Further, a number of psychological and neuroscience studies have demonstrated that up to a dozen special object categories can be classified by the event-related potential (ERP) recorded through EEGs [15]–[17], such as human faces. With applications of machine learning, a range of models have been developed [18]–[20] to address the problem of classifying visual objects via EEGs. However, most stimuli/objects in these studies are designed with only a single object set inside a clean background because enormous ambiguity surrounds the interpretation of EEG data, and multichannel EEG data sequences are generally only available in small quantities. In addition, EEGs are high dimensional yet have low signal-to-noise ratios, and the differences among individual subjects incur considerable temporal and spatial variability [21]. One study that is similar to ours in terms of the categorization objective is reported in [22], in which Walther *et al.* proposed an approach to estimate the categories of natural scenes using fMRI for only six categories. Specifically, they used fMRI and distributed patterns to analyze what regions of the brain can classify natural scenes. In practice, fMRI showed great potential in the brain imaging classification process; however, its main disadvantage lies in the experimental costs. This limitation is overcome by lower-cost techniques, such as EEGs, which provide higher temporal-resolution data compared to fMRI, but are susceptible to the aforementioned problems represented by lower signal-to-noise

ratios and spatial resolutions, posing significant challenges for brain imaging classification. With the progress of machine learning, some important limitations of the traditional neural networks have been overcome [23]–[27]. Inspired by such advancements, content-based multimedia understanding has achieved remarkable success in object detection [28] and image scene classification [29]. In contrast to the success of content-based multimedia understanding over the recent years, research on EEGs is still limited, providing an enormous scope for considerable improvement in terms of information extraction and classification of EEGs, including accuracy, generalization, and interpretability.

At present, image classification based on physiological signal analysis (implicit analysis) and content-based multimedia analysis (explicit analysis) have been two independently active research areas, and the relationship between the massive amount of physiological signals and the corresponding multimedia content has been relatively unexplored [30]. Nevertheless, the information from these two sources is likely to be complementary. On the one hand, physiological signal data may help us better understand the process of image classification inside human brains, and thus, it can be helpful for designing a highly robust brain-inspired model to classify the visual objects under complex backgrounds and occlusions. On the other hand, the feature extraction methods proposed for multimedia content analysis could inspire us to discover new and unstudied physiological signal patterns for developing more powerful and more intelligent object classification algorithms. In other words, brain-based image classification could be significantly improved if we can simultaneously leverage both the implicit and explicit modalities in designing the classification algorithms. Our extensive literature survey indicates, however, that there is no existing work integrating them for EEG-based image classification, although multimedia content analysis has achieved impressive progress over the past decade.

In this paper, we propose a novel deep brain analytics framework together with a multimodal approach for EEG-based image classification by integrating implicit and explicit modalities. Specifically, a consistency test based on a mapping between image content features and EEG-based features is added to promote potential solutions, and better performances are achieved compared to the state-of-the-art methods for EEG-based object classifications. Further, our proposed deep framework also demonstrates a good generalization capability in object categorization over a number of publicly available and widely adopted benchmarking datasets. In comparison with the existing research efforts and the corresponding state-of-the-art methods, the novelty of our contributions can be highlighted as follows. 1) We introduce a new concept of integrated implicit learning and explicit learning modalities to provide an alternative solution for the problem of brain imaging classification. 2) We propose a new deep brain analytics framework to exploit not only the strength of integrated multiple modalities, but also the advantages of the added consistency test for recommending potential targets and the fusion of individual classification results. 3) We carry out extensive experiments, and the results demonstrate that our proposed deep framework achieves superior performances in comparison with the existing state-of-the-art approaches.

II. Related Work

In psychology, stimuli refer to objects or events that cause a sensory or behavioral response in an organism. Therefore, stimuli form the basis of perception and behavior for human brain analytics, which has been intensively researched across the areas of neural science, psychology, and neural computation. Researchers aspire to present, analyze, distinguish, and understand how the human brain receives, handles, and processes rich and varied information in the real world through EEG signals, among which information about visual content and emotions is the primary target for research and analysis. Therefore, multimedia data containing a large amount of visual content information and emotional information are considered to be extremely suitable stimuli material, which are widely used in the acquisition, analysis, and classification of EEG signals [31], [32]. The research on multimedia content computing and multimedia emotional computing based on EEG signals has attracted enormous attention across relevant research communities [19], [32]–[35].

Before the popularity of deep learning methods, the primary approaches for image classification were predominantly feature based, and the commonly used features mainly included time-frequency features extracted by signal analysis methods, such as the power spectral density [36], bandpower [37], independent component analysis (ICA) [38], and differential entropy [39].

With the extensive application and in-depth promotion of deep learning, an ever-increasing number of brain and neuroscience research teams are exploiting its strength in designing ambitious algorithms to achieve intelligent understanding and perceptional analysis of brain activities via EEGs or fMRI. In [16], deep belief networks and deep automatic encoders to resolve the ERP P300 and non-P300 signals were reported. In [40], Yin and Zhang proposed a single-channel EEG classification method with a deep belief network to evaluate mental workload and mental fatigue states. In [20], an SVM classifier was trained to classify visually evoked EEG data according to 12 different object categories. In [41], a frequential deep belief network (FDBN) for classification tasks in motor imagery and adaptive EEG analysis was proposed. In [42], Gogna *et al.* proposed deep learning methods to solve the problem of reconstruction and classification of EEG data. In [43], a four-layer convolutional neural network to detect interictal discharges from intracranial EEG data was described, and determination of the effects of convolutional neural networks on decoding and visualization of EEGs was attempted and reported [44]. In [45], Dong *et al.* used the rectified linear unit activation function and long short-term memory (LSTM) on time frequency domain features to classify sleep stages. In [46], Stober *et al.* used CNNs and an autoencoder to classify audio-evoked EEG recordings. In [47], a compact full convolutional network (EEGNet) was proposed and applied to four different brain-machine interface classification tasks. In [48], Spampinato *et al.* used long-term and short-term memory network learning to obtain an EEG data representation based on image stimuli and constructed a mapping relationship from natural image features

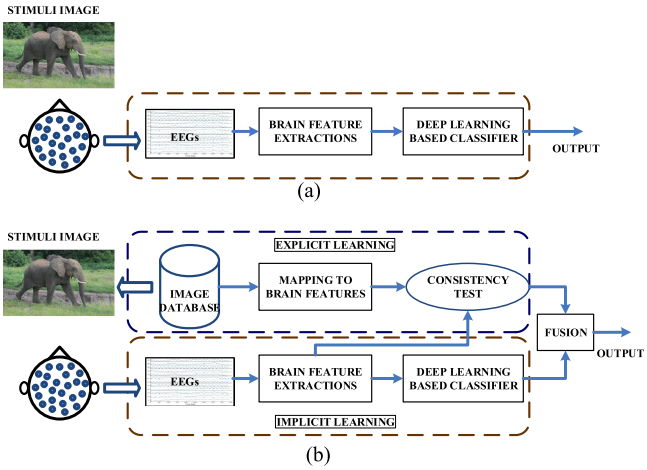


Fig. 1. (a) Illustration of the existing research on the single modality of implicit learning. (b) Illustration of our proposed framework with dual modalities, implicit learning, and explicit learning.

to EEG characterization. Finally, they used the new representation of EEG signals for classification of natural images. Compared with traditional methods, these deep learning based approaches have achieved outstanding results in realizing their respective research objectives. While these methods have demonstrated the capability of using brain signals and deep learning for classification purposes, none of them simultaneously integrated implicit and explicit modalities, and the state-of-the-art classification accuracy achieved to date by Spampinato *et al.* was 82.9% [48], leaving significant space for further research and improvement.

III. Proposed Multimodal Classification Algorithm

As shown in Fig. 1(a), the existing efforts on brain image classification are primarily limited to a single modality, so-called implicit learning, where EEG or fMRI signals are directly processed to extract brain features for learning and classification. Whether the learning process is designed as conventional or deep learning based, the essential strategy of a single modality remains unchanged. While such single modality strategies have achieved good progress across areas of both image processing and computer vision, the rich source of image content from which stimuli are selected for producing EEG or fMRI signals is basically ignored. To exploit the great successes, especially those achieved by deep learning based approaches toward intelligent image content analysis and classification, we introduce as a new strategy a second modality, so-called explicit learning, as shown in Fig. 1(b), which is added to target the rich source of images used for stimuli and to determine whether their analysis could provide further assistance in improving the classifications of EEGs.

Given m classes of images $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^{m}$, from which both training images and testing images are selected as stimuli to produce EEG signals, we apply deep learning based networks to extract brain feature vectors $\mathbf{B}_i \in \mathbb{R}^{d_b}$, where d_b stands for the dimension of the brain features. To create the second modality and exploit the rich source of the image database for improved brain signal classification, we propose to examine the m classes of images inside the original image database and determine if any of their content description and analysis in the so-called explicit learning can boost the implicit learning-based classification. To minimize the computing cost and the algorithm complexity, we select a number of representative images out of each class to characterize all the images inside the corresponding class, i.e., $\mathbf{g}_i, i \in [1, m]$. As the widely researched approach of clustering proves to be powerful in characterizing images, we apply a pixel-based clustering method to cluster all the images within each individual class such that the centroid of each cluster is taken as the most representative image for its corresponding class. In this way, the K-means clustering takes each class of images $\mathbf{g}_i$, as input and produces the most representative images per class, $\mathbf{G^r} = \{\mathbf{g}_i^{\mathbf{r}}\}_{i=1}^{m}$, as output, where $\left\{\mathbf{r}_1^i, \mathbf{r}_2^i, \ldots, \mathbf{r}_{n_i}^i\right\} \in \mathbf{g}_i^{\mathbf{r}}$ declares that each representative image class $\mathbf{g}_i^{\mathbf{r}}$ actually contains n_i representative images. In our algorithm design, we simply use the number of centroids as the number of representative images since the clustering is applied to all images of each individual class. While $\mathbf{G}$ denotes the original image set, with $\mathbf{g}_i$ being the ith class of images, $\mathbf{G^r}$ denotes the extracted representative image database, with $\mathbf{g}_i^{\mathbf{r}}$ being the ith class of representative images.

As seen in Fig. 1, the implicit learning modality essentially relies on deep learning of brain features to determine which class the input brain feature is associated with. To this end, the brain features of all the training images can be regarded as providing a certain level of ground truth for describing all the different classes. To allow a certain level of flexibility and tolerance for those images that could be selected as the test image, we cluster the entire database rather than the training images exclusively to produce the representative class for the consistency test and hence propose to add a second modality, i.e., the so-called explicit learning, by directly mapping all the representative images into the brain feature space and hence construct an explicit brain feature for each of them. In other words, we do not actually extract the brain feature from the EEG signals of those representative images; rather, we derive the brain feature directly from the mapping process because not all of the features have EEGs available. We then carry out a similarity-based consistency test to determine which representative brain feature provides the closest match for the brain feature of the stimuli image, and thus, the corresponding class can be selected as the recommended classification output to assist the classification from the implicit learning modality, the details of which are described as follows.

Let $\mathbf{B^t} = \left\{\mathbf{b}_1^{\mathbf{t}}, \mathbf{b}_2^{\mathbf{t}}, \ldots, \mathbf{b}_{d_b}^{\mathbf{t}}\right\} \in \mathbb{R}^{d_b}$ be the brain feature vector of the test image and $\mathbf{B}_{ij}^{\mathbf{r}} = \left\{\mathbf{B}_{i1}^{\mathbf{r}}, \mathbf{B}_{i2}^{\mathbf{r}}, \ldots, \mathbf{B}_{in_i}^{\mathbf{r}}\right\}$ be the brain feature set for the ith class of representative images; we calculate the distance between $\mathbf{B^t}$ and $\mathbf{B}_{ij}^{\mathbf{R}}$ as follows:

$$d\left(\mathbf{B^t}, \mathbf{B}_{ij}^{\mathbf{R}}\right) = \frac{1}{d_b} \sum_{k=1}^{d_b} \left(\mathbf{b}_k^{\mathbf{t}} - \mathbf{b}_{ij}^k\right)^2 \tag{1}$$

where $\mathbf{b}_{ij}^k$ is the kth element of the jth brain feature vector $\mathbf{B}_{ij}^{\mathbf{r}}$ inside the ith class of representative images.

The index of the representative image, for which the mapped brain feature of the input test image is the closest, can be derived via the following:

$$(i', j') = \underset{i\in[1,m], j\in[1,n_i]}{\operatorname{argmin}} d\left(\mathbf{B^t}, \mathbf{B}^{\mathbf{r}}_{ij}\right). \tag{2}$$

To ensure that such recommended class candidates have the best possible opportunity to include the true classification result, we allow a certain level of tolerance by applying (2) not only to the first minimum value, but also to the second and third minimum values. As the ith class contains n_i representative images, it is likely that a number of images across different classes are within the inclusion of the minimum match. If this is the case, all the brain features are integrated and averaged as a new brain feature that is then sent to the implicit learning model for reclassification, as if the brain features of the selected representative images were extracted from the EEG signals. Given that concatenation fusion is widely used in the machine learning community, examples of which include the inception model of GoogLeNet [49], applications in multimodal deep learning [50], etc., we propose to concatenate the input brain feature $\mathbf{B^t}$ with the brain features $\mathbf{B^r}$ from representative images to complete the reclassification. Such classified results are referred to as $\mathbf{C}_2$, and the direct classification of EEGs via implicit learning is referred to as $\mathbf{C}_1$. On the other hand, if all the index values derived by (2) are within a single class, this class candidate will be directly used as the recommended classification result, which is referred to as $\mathbf{C}_3$. Details of such a recommendation test are summarized as follows:

$$\gamma_c = \begin{cases} \mathbf{C}_3 & \text{if } (i', j') \in \mathbf{B}^{\mathbf{r}}_{i'} \forall_{i', j'} \\ \mathbf{C}_2 = \varphi\left(\dfrac{\alpha}{\eta} \displaystyle\sum_{k=1}^{\eta} \left\{\mathbf{B}^{\mathbf{r}}_{i'j'}\right\}_k + \beta \mathbf{B^t}\right) & \text{else} \end{cases} \tag{3}$$

where γ_c stands for the recommended class, η is the total number of brain features that achieve the minimum distance for $\mathbf{B^t}$ via (2), and α and β are the two weighting coefficients balancing the contributions of $\mathbf{B^r}$ and $\mathbf{B^t}$, which are determined via empirical study and training. Finally, $\varphi(\cdot)$ stands for the deep learning based classification obtained via the implicit learning modality.

To make the final decision for the classification output and integrate all individual classification results, $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3\}$, we add a simple fusion as follows:

$$\mathbf{C} = \begin{cases} \mathbf{C}_1 & \text{if } \mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}_3 \\ C_3 & \text{else if } P(\mathbf{C}_2) < T \bigcap P(\mathbf{C}_1) < T \\ \bar{C} & = \underset{\mathbf{C}\in[\mathbf{C}_1, \mathbf{C}_2]}{\operatorname{argmax}} \left(\delta P(\mathbf{C}_1), P(\mathbf{C}_2)\right) \text{ else} \end{cases} \tag{4}$$

where T is a threshold, which is determined empirically during the training process, $P(\mathbf{C})$ is the probability that $\mathbf{C}$ is the correct classification, and δ is an adjustment coefficient designed to constrain or boost $P(\mathbf{C}_1)$.

Essentially, (4) is designed to integrate all the classification results produced by the multimodal information fusion (two modalities), in which the first choice is straightforward and the second choice states that if neither $\mathbf{C}_1$ nor $\mathbf{C}_2$ has sufficient probability to justify its output, we would directly adopt the recommended classification as the output. Finally, the third choice states that the output is the one corresponding to the maximum probability among $P(\mathbf{C}_1)$ and $P(\mathbf{C}_2)$.

IV. Deep Learning Based Multimodal Framework

A. System Overview

Fig. 2 shows an overview of our proposed deep framework. As seen, the purpose of the real image is two-fold: As the input for the explicit learning and as the stimuli for evoking brain signals. The integrated EEG-based brain image classification consists of four stages, i.e., feature encoding, EEG regression, consistency test, and information fusion.

In the modality of implicit learning, the information is first extracted from raw EEG signals to construct brain cognitive features via an LSTM network, and these features are then fed into a number of later stages inside the framework, including the EEG regression stage, the consistency test stage, and the information fusion stage.

In the modality of explicit learning, on the other hand, the most representative images of each class are obtained via the clustering technique, and information is extracted from the image to construct visual features via the CNN, which is referred to as the feature encoding stage. To introduce the explicit learning into our framework, we propose to apply a KNN regression process and map the encoded visual feature into an EEG description, paving the way for the consistency test and hence producing recommendations for potential classification candidates $(\mathbf{C}_3)$ as described in (4).

The consistency test plays a gap-bridging role between the two modalities of implicit learning and explicit learning, where the brain features from EEGs and the content features from images are comparatively tested to estimate their consistency and fulfil the integration of the two modalities. Following that, a fusion stage is added to optimize the collective considerations of these individual classification results and hence deliver the best possible final classification performances. Under this circumstance, our fusion design is critical to ensure that both false positives and false negatives can be significantly reduced.

B. Feature Encoding

The feature encoding stage aims at extracting the brain cognitive feature representation $\mathbf{B}$ and the visual feature $\mathbf{V}$ from raw EEG signals and the input images, respectively.

In preparing the modality of implicit learning, the raw EEG signals are processed through a recurrent module, in which an LSTM is used as an encoder, and the temporal sequence is projected into a feature space $\mathbf{B} \in \mathbb{R}^{d_b \times n}$. Specifically, the LSTM-based encoder network consists of a single LSTM layer and the regular nonlinear output layer. It takes EEG brain signals as input and produces the brain cognitive features representation $\mathbf{B}$ as output.

To prepare the modality of explicit learning for context support and complementary classification, we send all representative images to a CNN-based GoogLeNet [49] encoder to

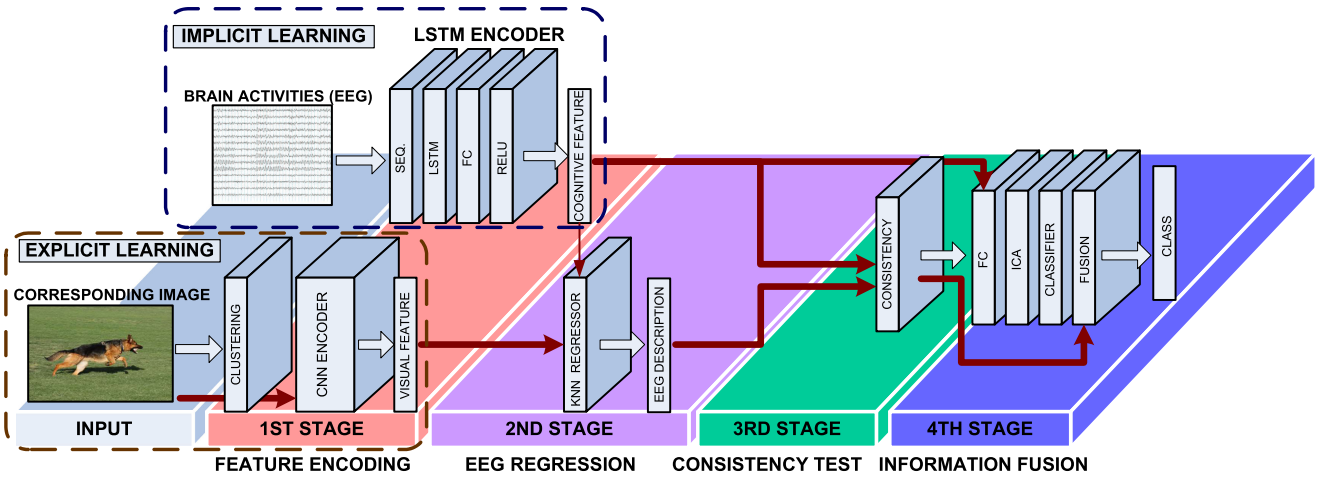


Fig. 2. Structural illustration of our proposed deep framework.

produce a set of visual feature vectors aimed toward mapping the visual content into EEG-compatible brain features. The overall structure and the relationship among all individual elements are shown in Fig. 2.

C. EEG Regression

To complete the mapping from the visual content representation to the brain feature description and ensure that the two different modalities, implicit learning and explicit learning, are compatible for integration, we add an EEG regression stage to project the visual features of the most representative images per class onto the brain cognitive features $\mathbf{B}$ via the visual features $\mathbf{V}$, producing the EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$.

The KNN regressor layer has three inputs and one output. The three inputs include the visual feature of the images $\mathbf{V}$, the visual feature of the representative images $\mathbf{V}^{\mathbf{r}}$, and the brain cognitive feature $\mathbf{B}$. The output includes the regressed EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$. The KNN regressor layer compares the visual features of the representative images to those of the images, and the K-nearest images to the representative images are retrieved. After that, the mean of the brain cognitive features associated with the K-returned images is calculated and considered to be the EEG description or characterization of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$.

D. Consistency Test

As it is known that EEGs are often degraded by noise interference, leading to the possibility that their classifications could be less reliable, we propose a consistency test to overcome this problem and help reduce the nonreliability. Given that stimuli images presented to human subjects for extracting EEG sequences are bounded in the 40 classes inside ImageNet-EEG [48], we extract n_i most representative images from each class via clustering techniques, and use these representative images to perform a consistency test on the brain activity analysis result, i.e., the classified output from the implicit learning modality, and to determine which representative images it is consistent with. As a result, the corresponding class can be taken as an alternative classification result, and we expect that the alternative classification should be the same as that from the implicit learning and classification. For those results that differ from each other, we apply a further fusion stage to finalize the classification output.

Essentially, the aim of the third stage is to produce the EEG representation of the consistent set based on the representative images. On the one hand, the first input of the consistency test is the brain cognitive feature vector, which is derived from EEG monitoring of the brain activities $\mathbf{B}^{\mathbf{t}}$. On the other hand, the second input is the EEG description of the most representative images $\mathbf{B}_{ij}^{\mathbf{r}}$, which is derived from the EEG regression stage. In this way, it is guaranteed that both $\mathbf{B}^{\mathbf{t}}$ and $\mathbf{B}_{ij}^{\mathbf{r}}$ are consistency testable. In addition, the KNN similarity measure is utilized to check the similarity between the inputs ($\mathbf{B}^{\mathbf{t}}$ and $\mathbf{B}_{ij}^{\mathbf{r}}$) and produce the K-nearest EEG representations of each image in the test set based on the EEG description of the representative images. After that, the mean of the EEG representation associated with the K representative images is calculated and considered as the EEG description of the consistent set. Consequently, either the classified results are verified by the consistency test or alternative classifications are produced for further fusion and integration.

E. Information Fusion

The information fusion stage consists of three main parts, including ICA, the classifier layer, and the fusion layer. This stage plays a constructive role in improving the classification accuracy for the proposed deep framework.

ICA is placed before the layer of classifiers as a feature selection module, which takes the EEG features B from the LSTM network as input and returns the independent statistical features as output. We implement the reconstruction ICA objective function based on the work reported in [51]. After ICA, two classifiers have been investigated, including the SoftMax classifier and the multiclass support vector machine (SVM).

To optimize the individual classification results derived by the consistency test and by direct classification of EEG signals, we add a multimodal information fusion layer, attempting to exploit both their individual strengths and their complementary advantages. Regarding the specific multimodal information fusion algorithm design, our fusion problem contends with two classification results corresponding to two modalities. On the one hand, we have direct classification results from the first stage, and on the other, we have the consistency test results, which could contradict each other in principle. Therefore, a question arises: What classification result should we adopt as the final one in order to maximize the classification accuracy?

Two main approaches have been investigated in our work, including decision-based multimodal fusion and feature-based multimodal fusion. In the decision-based multimodal fusion, we test the probability-based approach, and in the feature-based multimodal fusion, we further test three techniques, concatenation fusion, SUM, and MAX fusion.

For the probability-based fusion, we use the output probability of the classifier layer from both modalities, where the first modality output is $\mathbf{C}_1$ and the second modality outputs are $\mathbf{C}_2$ and $\mathbf{C}_3$, to determine the final classification result according to (4). Following the strategy given in (4), the second modality outputs work as a rectifier to improve the final classification accuracy.

V. Experiments

To evaluate our proposed deep framework and the introduced concept of context-supported multimodal learning, we conduct three phases of experiments. In the first phase, we try to evaluate the EEG-based object classification performance for our proposed deep framework on the publicly available dataset ImageNet-EEG [48]. In the second phase, we measure the generalization capability of the proposed framework on a subset of the visual classification dataset Caltech-101 [52]. By generalization capability, we mean how our proposed deep framework performs if it is applied to classify those objects or images that have not been seen before. In the third phase, the proposed deep framework is tested under a transfer learning setup. To benchmark our approach with multimodalities, we first compare the proposed framework with the existing state-of-the-art methods to verify its effectiveness. Then, we conduct additional evaluations to explore the performance of the proposed framework in more detail.

A. Experimental Settings and Training Details

Our experiments are conducted on two datasets, including the EEG-based classification dataset ImageNet-EEG and a subset of the visual-based classification dataset Caltech-101. ImageNet-EEG is a publicly available EEG dataset for brain imaging classification proposed by Spampinato *et al.* [48]. Caltech-101 includes 17 classes that coincidentally have the same names as those in ImageNet-EEG. Hence, those images are selected to construct the subset utilized to evaluate the generalization capability of our deep framework. For benchmarking purposes, the proposed deep framework is compared with the EEG-based object classification method [48], which is the latest research work published in 2017 on the same dataset. We also perform comparisons with several of the latest deep learning models for visual-based object classification, including AlexNet [23], VGGNet-16 [53], VGGNet-19 [53], GoogLeNet [49], and ResNet-101 [24].

In the modality of implicit learning, the iteration limit is set to 200 and the batch size is set to 440 for the parameters of the LSTM encoder in the first stage of the proposed deep framework. In the modality of explicit learning, K is set to 3 for the parameters of the pixel-based clustering and the feature-based clustering in the first stage of the proposed deep framework. In the third stage of the proposed deep framework, concerning the parameters for the consistency test, the number of nearest neighbours K is set to 3. In the fourth stage of the proposed deep framework, the number of extracted features from ICA is set to 70, and the iteration limit is set to 400. Our method is implemented on the Tesla P100 GPU.

As KNN and clustering employ an unsupervised learning mode and the CNN is pretrained on ImageNet for visual feature extraction, all three modules, the KNN regressor, the clustering, and the CNN, do not need to be trained. The LSTM encoder is trained on the EEG data with their labels. For the KNN regressor, which attempts to map the features into a user-specific EEG space, both features at the input are visual features, one for representative images and the other for training images. For the consistency test, both inputs for KNN are cognitive features from the representative images and training images.

TABLE I
Classification Performance Comparison Between Our Proposed Deep Framework and the State-of-the-Art Method [48]

Models	Accuracy
Proposed deep framework	94.1%
RNN-based model [48]	82.9%

B. EEG-Based Object Classification

In the first phase of experiments, we try to validate the effectiveness of our deep framework for EEG-based object classification. Our experiments are tested on the standard EEG dataset ImageNet-EEG. As ImageNet-EEG is collected using a 128-channel cap with active, low-impedance electrodes (actiCAP 128Ch), it includes the EEG signals of six subjects produced by asking them to look at visual stimuli, which are images selected from a subset of ImageNet [54], containing 40 classes with 50 images in each class. During the experiment, each image was shown on the computer screen for 500 ms.

Table I summarizes the experimental results in terms of the classification accuracies for both our proposed deep framework and the existing state-of-the-art method reported in [48]. As seen, while the precision rate achieved by our proposed deep framework is 94.1%, the existing state-of-the-art comparison is 82.9%.

To quantify the contribution of each stage designed in our proposed deep framework, we further carried out experiments to explore the effectiveness of different configurations of the

TABLE II
COMPARATIVE ASSESSMENT OF THE PROPOSED FRAMEWORK UNDER DIFFERENT CONFIGURATIONS

			Framework											
	Configurations		1	2	3	4	5	6	7	8	9	10	11	12
1ST stage:	Feature encoding	Pixel-based	✓		✓		✓		✓	✓	✓	✓	✓	✓
		Feature-based		✓		✓		✓						
4TH stage:	Classifier	SoftMax	✓	✓										
		SVM			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Feature selection	ICA					✓	✓		✓		✓		✓
	Decision-based multi-modal fusion	Probability-based	✓	✓	✓	✓	✓	✓						
	Feature-based multi-modal fusion	Concatenation							✓	✓				
		SUM									✓	✓		
		MAX											✓	✓
Accuracy	%		89.5	89.7	92.5	92.3	94.1	94.1	87.8	89.5	87	92.2	86.5	92.2

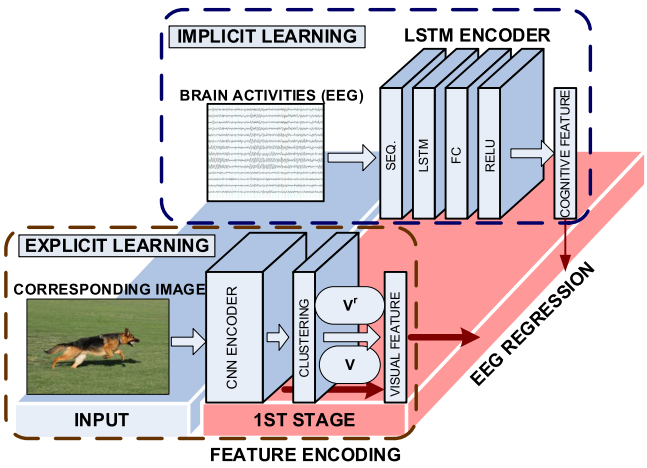


Fig. 3. Illustration of the alternative feature-based clustering.

individual stages. For the clustering stage, an alternative consideration is feature based, in which all the deep features of images inside each class can be clustered instead of their pixels, and then, the centroids are taken as the most representative deep features for their corresponding class (see Fig. 3). For the fusion stage, individual elements considered include 1) with or without ICA; 2) selection of different classifiers, including SoftMax and SVM; and 3) choice of different fusion methods, including probability based, corresponding to the pixel-based clustering, and feature based, corresponding to the feature-based clustering. Under the feature-based fusion, we could directly concatenate features as described in (3) or add every individual element of the features together as the fusion method (SUM). Finally, we could also select the most influential feature via $\mathrm{MAX}\{\mathbf{B^t}, \mathbf{B^r}\}$ (MAX).

Table II reports the experimental results in terms of the classification precision rates for all the configurations, from which we can observe and draw a number of conclusions, as described below.

First, the feature-based clustering in the first stage is better than the pixel-based clustering method, although the improvement is limited. This occurs if we select SoftMax as the classifier in the fourth stage of the proposed framework (see Fig. 2). These results are demonstrated in Table II by configurations 1 and 2.

Second, the performance of the pixel-based clustering is similar to (or even better than) the performance of the feature-based clustering if we select SVM as the classifier in the fourth stage. These results are demonstrated in Table II by configurations 3 to 6. These results illustrate why we select the pixel-based clustering method in the first stage when using SVM as the classifier in the fourth stage.

Third, we find that the SVM classifier is always better than the SoftMax classifier in the fourth stage. These results are demonstrated by configurations 1 to 4 in Table II. While the best performance of the SoftMax classifier is 89.7% (configuration 4), the best performance of the SVM classifier is 92.5% (configuration 5).

Fourth, we find that the performance of ICA plus SVM is always better than using SVM alone in the fourth stage. If we use ICA to reduce the feature dimension, the performance is always better, as demonstrated by configurations 3–6 in Table II. While the best performance of the SVM implementation is 92.5% (configuration 3), the best performance of employing ICA plus SVM is 94.1% (configurations 5 and 6).

Fifth, in the feature-based fusion method, SUM and MAX are better than concatenation fusion in the fourth stage. This is demonstrated by configurations 7–12 in Table II. While the best performance of the concatenation fusion method is 89.5% (configuration 8), the best performance of the SUM/MAX fusion method is 92.2% (configuration 10). The reason that the SUM/MAX fusion method outperforms the concatenation-based fusion is mainly due to the nature of the inputs to the fusion function. As the inputs are consistent features, the model-free fusion, SUM/MAX, is more suitable, while the concatenation fusion requires universal approximation to estimate the model parameters.

Sixth, the probability-based fusion method is better than the feature-based fusion. This is demonstrated by configurations 1–12 in Table II. While the best performance of the feature-based fusion method is 92.2% (configurations 10–12), the best performance of the probability-based fusion method is 94.1% (configurations 5 and 6).

Finally, we can conclude that the addition of the explicit learning modality does help the implicit learning modality achieve better performances for EEG-based object classification. Without the explicit learning modality, the best performance of implicit learning alone is only 90.5% when SVM is used as

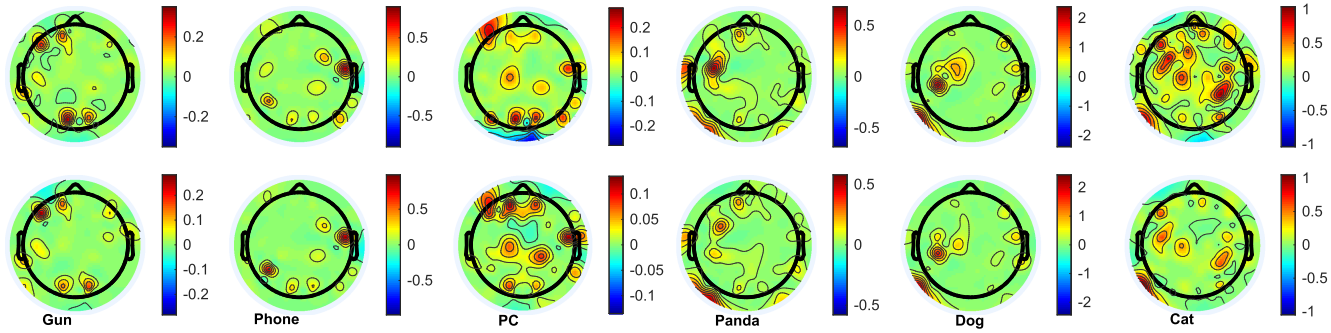


Fig. 4. Scalp distribution of the average energy for all participants and sessions for six categories, including "gun," "phone," "desktop PC," "panda," "dog," and "cat."

the classifier, and the best performance of the implicit learning alone is 82.9% when SoftMax is used as the classifier (due to space limitations, we do not list these results in Table II). With the addition of the explicit learning modality, however, the best performance of these settings is 94.1% (configuration 5) and 89.7% (configuration 2).

One novelty of our proposed framework is to integrate explicit and implicit learning modalities with a similarity-based consistency test on the representative images for every category. While the representative images are selected via pixel-level clustering, whether these representative images will truly trigger particular responses at the brain level for their corresponding classes remains questionable. To answer this question, we provide an average energy distribution of six categories for all participants and all sessions (first row) and the average energy distribution of the most representative images (second row) in Fig. 4. As seen, the neural activations of the representative images are close to that of all images in the same category, and an interesting observation is that the activations of three objects ("gun," "phone," and "desktop PC") are different from that of three animals ("panda," "dog," and "cat"). As three adorable animals, the activations of "panda," "dog," and "cat" share some level of similarity, as seen in Fig. 4, especially in the temporal area, indicating a strong sensitivity to visual perceptions, such as animal faces. In the ImageNet-EEG dataset, which includes 40 classes, most of the categories are not related to human emotions. From the scalp distribution of the average energy for all participants across all sessions, however, it is obvious that there exist higher responses at prefrontal areas, and these EEG data could be used for emotion classification.

C. Generalization Test for the Proposed Deep Framework

To test our proposed deep framework for its generalization capability in classifying brain images that are not previously seen by the framework via EEGs, we carry out the second phase of experiments on another widely used dataset, Caltech-101 [52]. Caltech-101 has 17 classes that are named the same as those in ImageNet-EEG, which creates an opportunity for us to carry out the generalization test by using the corresponding EEG signal sequences provided in [48]. For the convenience of result analysis and comparative studies, we construct a subset with all 17 classes from Caltech-101 to implement the second experiment, and the 17 classes include airplanes, bass, butterfly, camera, car with side view, cellphone, chair, cup, Dalmatian (dog), electric guitar, elephant, grand piano, lotus, panda, pizza, revolver, watch, and wildcat. The total number of images is 2059, and the number of images for each class is 121 on average.

TABLE III
COMPARATIVE GENERALIZATION TEST BETWEEN OUR PROPOSAL AND THE EXISTING STATE-OF-THE-ART METHOD [48]

Models	Accuracy
Proposed deep framework	80%
CNN-based model [48]	77%

Specifically, images from the 17 classes are taken as the input for GoogLeNet, and the output of the last fully connected layer is used as the extracted visual features. To maintain the necessary compatibility between the extracted visual features (explicit learning) and the brain cognitive features (implicit learning) for a smooth integration of the two modalities, we project all these visual features onto the brain feature space via the learned KNN regression module as shown in Fig. 2, and this is implemented without performing training on any image in Caltech-101.

The experimental results are summarized in Table III, which lists the classification performances for both the proposed deep framework and the existing benchmark [48]. As seen, our proposed framework outperforms the benchmark by 3%, indicating the following: 1) our proposed framework has a better generalization capability in classifying visual objects not previously seen, and 2) human visual capabilities can be learned and exchanged via machine learning.

For the convenience of further analysis and comparative investigation, Fig. 5 presents the confusion matrix of each category for Caltech-101. The rows represent the 17 classes from Caltech-101, and the first 17 columns represent the corresponding classes in ImageNet-EEG. The last column is used to represent the rest of the 40 classes from ImageNet-EEG. As we directly use the learned models trained via ImageNet-EEG to predict the images in Caltech-101, we find that some images are incorrectly classified into classes that do not exist in Caltech-101. In the confusion

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Others
1	73	7	0	0	10	0	0	0	0	0	0	0	1	0	0	0	0	9
2	3	85	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	6
3	3	0	31	0	5	0	0	0	0	0	0	0	1	0	1	1	8	50
4	6	0	2	91	0	0	0	0	0	0	0	0	0	0	0	0	0	2
5	8	3	0	0	66	0	0	0	0	0	0	0	0	0	0	0	0	24
6	0	0	0	0	3	84	0	0	0	0	0	0	0	0	0	0	0	14
7	0	0	0	0	0	0	56	0	0	19	0	0	3	0	0	0	0	22
8	2	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	4
9	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	5
10	0	0	0	0	0	0	0	0	0	81	0	0	1	0	0	0	0	17
11	3	0	0	0	5	0	0	0	0	0	65	0	0	0	4	0	0	23
12	2	0	0	0	0	0	2	0	0	2	0	61	21	0	0	0	0	13
13	1	0	0	0	0	0	0	0	0	0	0	0	94	0	0	0	0	5
14	0	0	0	0	0	0	8	0	0	4	0	0	0	58	0	0	0	30
15	0	0	0	0	1	0	0	0	0	1	0	0	0	0	94	0	0	4
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	0	8
17	2	0	2	0	5	0	0	0	0	0	0	0	0	0	0	0	80	12

Fig. 5. Confusion matrix, where the rows represent the 17 classes from Caltech-101, and the first 17 columns represent the corresponding classes in ImageNet-EEG. The column "Others" represents the rest of the 40 classes from ImageNet-EEG.

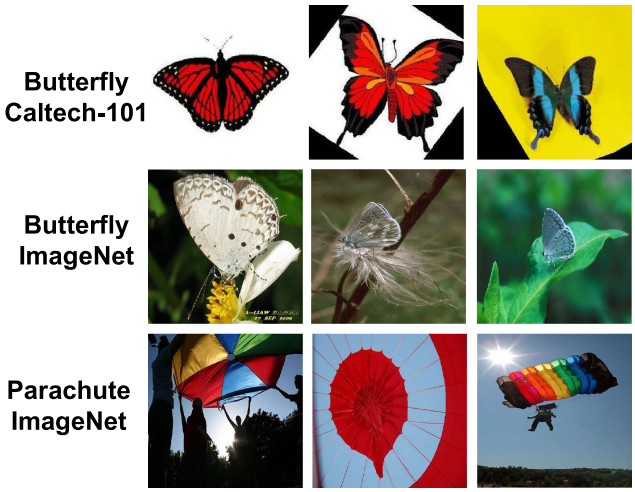


Fig. 6. Sample images from Caltech-101 and ImageNet.

matrix, these samples contribute to the confusion values labelled as "Others."

As seen in Fig. 5, the classification accuracy of the "butterfly" category is worse than others, only 31%. Additional examination even indicates that most of the images from this category are wrongly classified as "parachute" by our framework, which prompted our further investigation into the results. Additionally, Fig. 6 shows several sample images from the class "butterfly" in both Caltech-101 and ImageNet-EEG. As seen, the visual content from the "butterfly" images of Caltech-101 is obviously very different from that from the "butterfly" images of ImageNet-EEG. In other words, although the two classes are named the same, their images do not have any similarity in terms of visual objects. In contrast, the visual content from the "butterfly" images of Caltech-101 is very similar to that from the "parachute" images of ImageNet-EEG.

D. Transfer Learning via the Proposed Deep Framework

To improve the classification performances of our proposed deep framework on Caltech-101, we add a training process by selecting images from the 17 classes of Caltech-101. We do not want to change the EEG sequences inside ImageNet-EEG [48]; however, the power of transfer learning can be exploited to augment the EEG-based classification with help from the training process via images from Caltech-101.

As shown in Fig. 2, the essential integration of the two different modalities of explicit learning and implicit learning is supported because the compatibility between the visual features directly extracted from images and the brain cognitive features extracted from the EEGs is preserved. To this end, we establish an indirect mapping from the visual level to the brain cognitive level by transfer learning. In the first round of indirect mapping, we aim to obtain the brain cognitive level representation of each image in the training set of the 17 classes in Caltech-101. Specifically, the visual features extracted from the Caltech-101 images are further processed by KNN-based regression and then represent the brain cognitive features. By comparing the extracted visual features to those of the training images in ImageNet-EEG, the nearest neighboring images from ImageNet-EEG can be retrieved, and the mean of the brain cognitive features associated with these returned images are taken as the brain cognitive-level representation of each image in the training set of Caltech-101. In the second round of indirect mapping, our goal is to obtain the brain cognitive-level representation of the test images in the 17

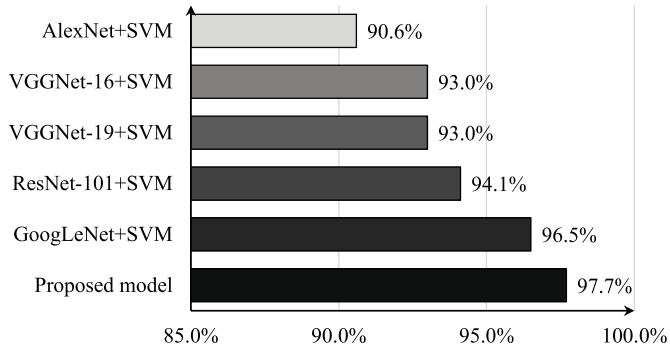


Fig. 7. Classification performance comparison between our proposed deep framework and the latest deep learning methods as the feature extractors, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101, where SVM is used as the classifier.

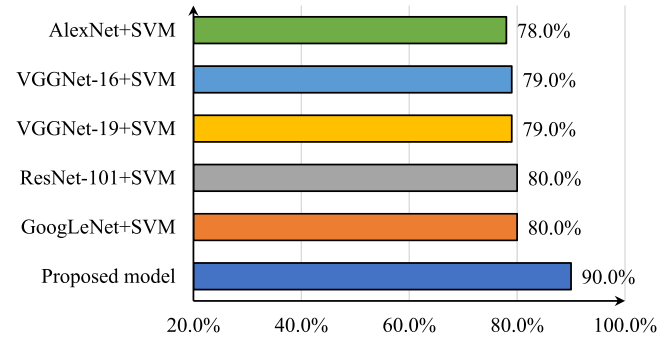


Fig. 8. Classification performance comparison between our proposed deep framework and the latest deep learning methods, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101, on the category of "butterfly".

classes of Caltech-101. The procedure is very similar to the first round, and the only difference is that the regression is carried out between the test set and the training set of Caltech-101 rather than between the training set of Caltech-101 and ImageNet-EEG. After the indirect mapping has been constructed, the brain cognitive-level representations of the training images in the 17-class Caltech-101 set are used to train the SVM classifier, and those of the test images are used to evaluate our proposed deep framework.

For the purpose of maintaining fair comparisons, AlexNet [23], VGGNet-16 [53], VGGNet-19 [53], GoogLeNet [49], and ResNet-101 [24], which are pretrained by ImageNet, are used as the feature extractors, and SVM is used as the classifier. The subset of the 17 classes in Caltech-101 is split into training, validation, and test sets, with percentages of 80%, 10%, and 10%, respectively. For all of those compared, images from the training set in the 17 classes of Caltech-101 are used to train SVM, the validation set is utilized to determine the parameters of SVM, and we evaluate the performance on the test set.

The experimental results are summarized in Fig. 7, from which we can see that the classification accuracy of our proposed framework is better than those of all the other approaches. Considering that our proposed deep framework primarily relies on classification of EEG sequences in ImageNet-EEG and the training with images from Caltech-101 is only exploited via the power of transfer learning, our proposed deep framework still remains competitive, particularly since the benchmark, GoogLeNet+SVM, directly exploits the training process with images from Caltech-101, rendering our proposal at a significant disadvantage.

For the convenience of further comparative analysis, we specifically focus on the category of "butterfly" and carry out a detailed investigation of the performance in the generalization test. As expected, the performance on "butterfly" is not good because the visual appearances of the images across the two datasets are significantly different and uncorrelated. As a result, a further analysis is conducted to study the performances after the training procedure. Fig. 8 demonstrates the classification accuracies on the category of "butterfly" achieved by our proposed deep framework and the latest deep learning methods, including AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101. As seen, the classification accuracy of our proposed deep framework is better than those of all the other methods. While the "butterfly" classification accuracy achieved by our deep framework is 90%, the "butterfly" classification accuracies achieved by AlexNet, VGGNet-16, VGGNet-19, GoogLeNet, and ResNet-101 are 78%, 79%, 79%, 80%, and 80%, respectively.

TABLE IV
PERFORMANCE EVALUATION OF THE PROPOSED DEEP FRAMEWORK FOR TWO DIFFERENT CONFIGURATIONS

	Framework	
Configurations	3	5
Accuracy	96.5 %	97.7 %

To assess the influence of different configurations, we further carry out experiments with a range of configurations to evaluate how the classification performances vary, and Table IV summarizes the top two results. As seen, configuration 5, which achieves the highest precision rate in the EEG-based object classification experiment, also achieves the highest precision rate and outperforms all the others.

VI. CONCLUSION

In this paper, by integrating implicit and explicit learning modalities, we propose a novel deep framework for EEG-based brain imaging classification. Our proposed framework provides an improved solution for the problem that, given an image used to stimulate brain activities, we should be able to identify which class the stimuli image comes from by analyzing the prompted EEG signals. As the visual cognitive capability of human brains is primarily researched via fMRI across the neural science and brain cognitive computing communities, significant challenges exist for processing EEG sequences, as they are exposed to noise and a high level of ambiguity exists. To address these challenges, we exploit the explicit learning widely researched across the areas of computer vision, digital media, and machine learning. To this end, we add a consistency test between the cognitive features extracted from EEG sequences and the visual features extracted from representative images to produce alternative and improved

solutions. Extensive experiments support that our proposed approach outperforms the existing state-of-the-art methods under various contexts and set-ups. The success achieved not only indicates that EEGs have significant potential for capturing brain activities for visual cognitive computing, but also opens up a new direction for explicit learning that, while widely researched in computer science, could also play significant roles in understanding and exploring the human brain.

A number of possibilities can be identified for further research, including applications of our deep framework for EEG-based brain activity understanding and interpretation, and testing the reliability and robustness of our framework toward recognition of visual objects and content inside human brains.

REFERENCES

[1] L. Yuan and J. Cao, "Patients' EEG data analysis via spectrogram image with a convolution neural network," in *Intelligent Decision Technologies*, I. Czarnowski, R. J. Howlett, and L. C. Jain, Eds. Cham, Switzerland: Springer, 2018, pp. 13–21.

[2] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image Vision Comput.*, vol. 31, no. 2, pp. 164–174, 2013.

[3] E. Kroupi, A. Yazdani, J.-M. Vesin, and T. Ebrahimi, "EEG correlates of pleasant and unpleasant odor perception," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1s, 2014, Art. no. 13.

[4] C. Guger, A. Schlogl, C. Neuper, D. Walterspacher, T. Strein, and G. Pfurtscheller, "Rapid prototyping of an EEG-based brain-computer interface (BCI)," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 9, no. 1, pp. 49–58, Mar. 2001.

[5] A. M. Green and J. F. Kalaska, "Learning to move machines with the mind," *Trends Neurosciences*, vol. 34, no. 2, pp. 61–75, 2011.

[6] D. Wu, "Online and offline domain adaptation for reducing BCI calibration effort," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 550–563, Aug. 2017.

[7] Y. Mishchenko, M. Kaya, E. Ozbay, and H. Yanar, "Developing a 3- to 6-state EEG-based brain-computer interface for a virtual robotic manipulator control," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 4, pp. 977–987, Apr. 2019.

[8] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[9] B. Hu, X. Li, S. Sun, and M. Ratcliffe, "Attention recognition in EEG-based affective learning research using CFS+KNN algorithm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 38–45, Jan. 2018.

[10] M. Fan and C. Chou, "Detecting abnormal pattern of epileptic seizures via temporal synchronization of EEG signals," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 601–608, Mar. 2019.

[11] D. Wang *et al.*, "Epileptic seizure detection in long-term EEG recordings by using wavelet-based directed transfer function," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2591–2599, Nov. 2018.

[12] Z. Song, B. Deng, J. Wang, and R. Wang, "Biomarkers for Alzheimer's disease defined by a novel brain functional network measure," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 41–49, Jan. 2019.

[13] D. Labate, F. L. Foresta, G. Morabito, I. Palamara, and F. C. Morabito, "Entropic measures of EEG complexity in Alzheimer's disease through a multivariate multiscale approach," *IEEE Sensors J.*, vol. 13, no. 9, pp. 3284–3292, Sep. 2013.

[14] X. Chen, X. Chen, R. K. Ward, and Z. J. Wang, "A joint multimodal group analysis framework for modeling corticomuscular activity," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1049–1059, Aug. 2013.

[15] K. Das, B. Giesbrecht, and M. P. Eckstein, "Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers," *Neuroimage*, vol. 51, no. 4, pp. 1425–1437, 2010.

[16] J. Kulasingham, V. Vibujithan, and A. De Silva, "Deep belief networks and stacked autoencoders for the p300 guilty knowledge test," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci.*, 2016, pp. 127–132.

[17] F. Li *et al.*, "Deep models for engagement assessment with scarce label information," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 598–605, Aug. 2017.

[18] J. Wang, E. Pohlmeyer, B. Hanna, Y.-G. Jiang, P. Sajda, and S.-F. Chang, "Brain state decoding for rapid image retrieval," in *Proc. 17th ACM Int. Conf. Multimedia.*, 2009, pp. 945–954.

[19] J. Moon, Y. Kwon, K. Kang, C. Bae, and W. C. Yoon, "Recognition of meaningful human actions for video annotation using EEG based user responses," in *Proc. Int. Conf. Multimedia Model.*, 2015, pp. 447–457.

[20] B. Kaneshiro, M. P. Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes, "A representational similarity analysis of the dynamics of object processing using single-trial EEG classification," *PLoS ONE*, vol. 10, no. 8, p. e0135697, 2015.

[21] S. Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 6175–6179.

[22] D. B. Walther, C. Eamon, F. F. Li, and D. M. Beck, "Natural scene categories revealed in distributed patterns of activity in the human brain," *J. Neuroscience Official J. Soc. Neuroscience*, vol. 29, no. 34, pp. 10573–10581, 2009.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 770–778.

[25] Z. Zhang, Y. Lu, L. Zheng, S. Li, Z. Yu, and Y. Li, "A new varying-parameter convergent-differential neural-network for solving time-varying convex QP problem constrained by linear-equality," *IEEE Trans. Autom. Control*, vol. 63, no. 12, pp. 4110–4125, Dec. 2018.

[26] Z. Zhang *et al.*, "A varying-parameter convergent-differential neural network for solving joint-angular-drift problems of redundant robot manipulators," *IEEE/ASME Trans. Mechatronics*, vol. 23, no. 2, pp. 679–689, Apr. 2018.

[27] Z. Zhang and L. Zheng, "A complex varying-parameter convergent-differential neural-network for solving online time-varying complex sylvester equation," *IEEE Trans. Cybern*, pp. 1–13, 2018.

[28] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[29] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[30] X. Li, X. Jia, G. Xun, and A. Zhang, "Improving eeg feature learning via synchronized facial video," in *Proc. IEEE Int. Conf. Big Data*, 2015, pp. 843–848.

[31] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[32] Y. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Oct. 2018.

[33] Y. Ding, X. Hu, Z. Xia, Y. Liu, and D. Zhang, "Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching," *IEEE Trans. Affect. Comput.*, pp. 1–12, 2018.

[34] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1809–1817.

[35] M. Bilalpur, S. M. Kia, T. S. Chua, and R. Subramanian, "Discovering gender differences in facial emotion recognition via implicit behavioral cues," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 119–124.

[36] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[37] S. Dähne, F. Bießmann, F. C. Meinecke, J. Mehnert, S. Fazli, and K.-R. Müller, "Integration of multivariate data streams with bandpower signals," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1001–1013, Aug. 2013.

[38] F. Cong *et al.*, "Linking brain responses to naturalistic music through analysis of ongoing EEG and stimulus features," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1060–1069, Aug. 2013.

[39] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng.*., 2013, pp. 81–84.

[40] Y. Zhong and Z. Jianhua, "Cross-subject classification of mental fatigue by neurophysiological signals and ensemble deep belief networks," in *Proc. 36th Chin. Control Conf.*, 2017, pp. 10 966–10 971.

[41] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted Boltzmann machines," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 566–576, Jun. 2017.

[42] A. Gogna, A. Majumdar, and R. Ward, "Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2196–2205, Sep. 2017.
[43] A. Antoniades *et al.*, "Detection of interictal discharges with convolutional neural networks using discrete ordered multichannel intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 12, pp. 2285–2294, Dec. 2017.
[44] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
[45] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.
[46] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep feature learning for EEG recordings," 2015, ArXiv:1511.04306, 2015.
[47] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional network for EEG-based brain-computer interfaces," 2016, arXiv:1611.08024.
[48] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6809–6817.
[49] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
[50] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[51] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Advances Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.
[52] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.
[54] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

Jianmin Jiang received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994.

From 1997 to 2001, he worked as a Full Professor of Computing with the University of Glamorgan, Wales, U.K. In 2002, he joined the University of Bradford, Bradford, U.K., as a Chair Professor of Digital Media and Director of the Digital Media & Systems Research Institute. He worked with the University of Surrey, Surrey, U.K., as a Full Professor during 2010–2014 and a Distinguished Chair Professor (1000-plan) with Tianjin University, Tianjin, China, during 2010–2013. He is currently a Distinguished Chair Professor and Director with the Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. He has authored approximately 400 refereed research papers. His research interests include image/video processing in the compressed domain, digital video coding, medical imaging, computer graphics, machine learning, and AI applications in digital media processing, retrieval, and analysis.

Dr. Jiang was a Chartered Engineer, Fellow of IEE, Fellow of RSA, member of EPSRC College in the U.K., and EU FP-6/7 evaluator.

Ahmed Fares received the Ph.D. degree from the Department of Computer Science and Engineering, Egypt-Japan University of Science and Technology (E-JUST), New Borg El Arab, Egypt, in 2015.

Currently, he is a Postdoctoral Researcher with the Research Institute for Future Media Computing at the College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China, and an Assistant Professor with the Department of Electrical Engineering and the Computer Engineering branch at the Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt. His research interests include brain science, cognitive science, computational modeling, theoretical computer science, and machine learning.

Sheng-Hua Zhong received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2013.

She worked as a Postdoctoral Research Associate with the Department of Psychological & Brain Sciences at The Johns Hopkins University, Baltimore, MD, USA, from 2013 to 2014. Currently, she is an Assistant Professor with the College of Computer Science & Software Engineering at Shenzhen University, Shenzhen, China. Her research interests include multimedia content analysis, cognitive science, psychological and brain science, and machine learning.

Exploring Short-Term Training Effects of Ecological Interfaces: A Case Study in Air Traffic Control

Clark Borst, Roeland M. Visser, Marinus M. van Paassen, *Senior Member, IEEE*, and Max Mulder, *Member, IEEE*

***Abstract*—In many work domains, the push toward higher levels of automation raises the concern of diminishing human expertise. Ecological interfaces could help operators in retaining and potentially even in acquiring expertise as they are hypothesized to lead to a deeper understanding of the work domain. This study explores the short-term impact of ecological interfaces on knowledge development and compares the results with an instruction-based training method. To monitor and compare students' progress, their decision-making strategies, identified from verbal comments recorded in "think-aloud" simulator sessions, are mapped onto the decision ladder. This method has been applied to an experiment ($N = 16$) aimed at training novices in conflict detection and resolution (CD&R) within a simplified air traffic control context. Results show that the overall CD&R performance in the final measurement sessions, featuring a transfer manipulation, was not significantly different between the "ecological" and "instructional" groups. In terms of cognitive behavior, however, students in the ecological group exhibited more laborious rule- and knowledge-based behaviors that sparked goal-oriented thoughts and corresponding control performances beyond the CD&R task. These findings indicate that ecological interfaces can change how people think and approach a control problem, even after removing the support. It is therefore reasonable to believe that ecological interfaces can play an important role in the early stages of deep knowledge development.**

***Index Terms*—Air traffic control (ATC), ecological interface design (EID), human–machine interface, training.**

I. Introduction

MANY work domains are moving toward higher levels of automation to meet more stringent safety, efficiency, and productivity demands. As articulated in Bainbridge's *Ironies of Automation* [1] and in recent work [2], a concern is that the cognitive expertise of human operators will diminish. Ironically, human expertise is critical for handling situations where automation support is unavailable (e.g., due to failures). Ecological interface design (EID) could help operators in retaining expertise as ecological displays provide a deeper insight into the physics and causal processes governing their work [3]–[6]. By serving as an "externalized mental model (...) that can support thought experiments and other planning activities" [4, p. 599], could ecological interfaces also contribute to building expertise by shaping the *internal* mental model? Previous longitudinal studies in process control have indeed reported skill and knowledge acquisition of students after being exposed to an ecological interface over a period of six months [7], [8]. Despite the encouraging findings, these studies also reported several limitations.

First, students did not receive any initial training, meaning that they had to engage in discovery learning while working with the interface. For some students, however, this incited "surface learning" and led to shallow knowledge as they did not actively reflect on the displayed information [8]. Second, the knowledge acquisition process was monitored by written "control recipes," in which students needed to write down a set of instructions on how they controlled the system. Although this gave insight into how students organized and chunked their knowledge, it cannot be ruled out that a hindsight bias may have confounded these recipes [8].

In this paper, a new empirical investigation is described that is aimed at overcoming the above-mentioned limitations while focusing on a different application domain: air traffic control (ATC). The scope will be training ATC novices, who are unbiased by previously developed strategies, in a conflict detection and resolution (CD&R) task in the horizontal plane. An ecological interface developed in a previous study, the solution space diagram (SSD) [9], [10], will be used for this purpose. The general approach will be similar to the study conducted by Christoffersen *et al.* [7], [8] in that the control performance and acquired knowledge of two participant groups are tested and compared after a transfer manipulation where the ecological support will be removed. Despite this similarity, there are three important differences.

First, both participant groups received the same initial training by a set of "best practice" instructions in CD&R [11]. This gives participants a head start in proper knowledge development and facilitates a more fair comparison between the two groups. Second, the gained knowledge and control strategies were monitored by mapping verbal comments, recorded during think-aloud simulation sessions, onto Rasmussen's decision ladder (DL) [12]. Thinking aloud while performing a task eliminates a potential hindsight bias. Third, the experiment took place over just two days, thereby aiming to explore short-term effects of EID on training.

The overall goal of the work described in this paper is twofold. First, related to the ATC work domain under investigation, to explore how the SSD contributes to picking up industry "best

Manuscript received March 19, 2018; revised October 15, 2018 and April 26, 2019; accepted May 5, 2019. Date of publication June 12, 2019; date of current version November 21, 2019. This paper was recommended by Associate Editor C. Marie. *(Corresponding author: Clark Borst.)*

The authors are with the Section Control and Simulation, Delft University of Technology, 2600 AA Delft, The Netherlands (e-mail: c.borst@tudelft.nl; roelmartijn@gmail.com; m.m.vanpaassen@tudelft.nl; m.mulder@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2919742

practices" and facilitates thinking and/or acting beyond those rules. Second, to provide new empirical insights into the merits and versatility of EID on knowledge development of system operators.

II. Background

A. Motivation for EID in ATC Training

In ATC, trainees are taught to expedite air traffic safely and as efficiently as possible, using a combination of learned strategies and procedures in response to recognized patterns and conflict geometries [13], [14]. Self-discovery of "what works" is typically how they learn the required skills. Trainees are also faced with unexpected disturbances that challenge earlier proven solutions, to discourage a "solve-all" strategy and to encourage knowledge-based problem solving rather than memorizing "tricks" [14].

Given the description of the ATC training process and its objectives, similarities can be discovered with the tenets of EID. First, similar to how ATC trainees are taught "robust" control strategies instead of fixed procedures, the EID framework was founded on the basic principle of providing support for unanticipated events for which no procedures exist. Ecological interfaces typically do so by portraying the range, or space, of possibilities (e.g., governed by the constraints of the work domain) instead of presenting single-optimized solutions that may fall short in situations that violate their specific assumptions [5].

Second, the air traffic controller needs to develop a mental model of the operational situation and how elements of the tactical traffic situation (e.g., aircraft positions, their flight directions, atmospheric conditions, etc.) relate to a higher level strategic "situation" that contains information about the current and future state of the airspace under control [14]. In EID, Rasmussen's abstraction hierarchy and/or the abstraction-decomposition space serve a similar purpose, by grouping domain-relevant constraints at different levels of abstraction, ranging from lower level states and whereabouts of objects to their relationships with higher level functional goals in the operational environment. A goal of EID is to portray this work domain structure on a display to serve as an "externalized mental model" of the system under control [4].

Third, controller expertise is influenced by a large number of perceptual factors. This is not surprising to consider that a controller needs to gain knowledge about the state of the airspace entirely from a plan view display (PVD), i.e., the electronic radar display. An ideal ATC training tool should thus support the trainee to become familiar with intricacies of the operational context by actively supporting this action–perception cycle. In this view, ecological interfaces typically aim to transform a cognitive task into a perceptual task [4], potentially enabling users to expedite situation recognition and formulate solutions to problems.

B. Supporting Solution Strategies for Workload Mitigation

An important trait of expert controllers is that they manage their own workload by applying solution strategies that minimize the required monitoring time [15]–[19]. Although ATC instructors do not teach particular solution strategies, literature indicates that expert controllers tend to converge to a range of "best practices" with workload-mitigating properties [15]–[19]. Also here, ecological interfaces are expected to support the development of such practices.

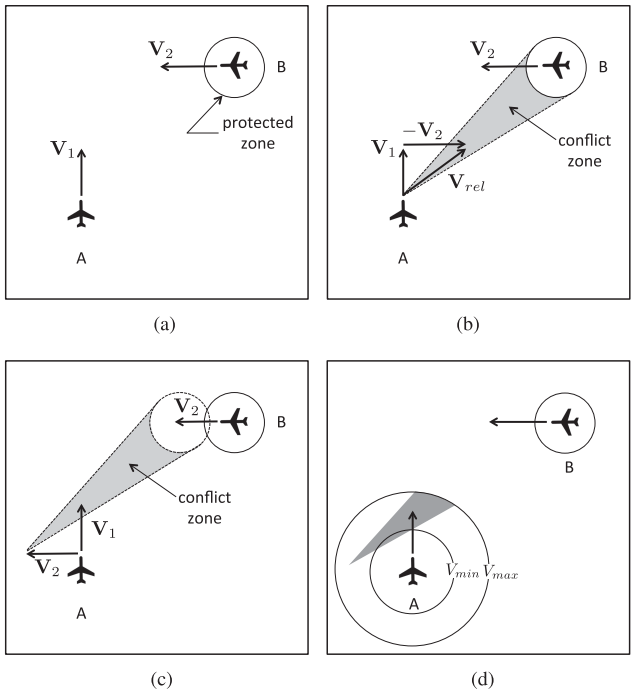


Fig. 1. SSD, showing the triangular velocity obstacle (i.e., conflict zone), formed by aircraft B within the speed envelope of the controlled aircraft A, and the absolute speed vectors of both aircraft (adapted from [9] and [11]). (a) Traffic geometry. (b) Conflict zone in relative space. (c) Conflict zone in absolute space. (d) Final SSD for aircraft A.

To illustrate, consider the SSD shown in Fig. 1, an ecological interface developed for ATC [9], [10]. In its most succinct form, the SSD portrays velocity obstacles (or, conflict zones) in speed and heading within the maneuvering envelope of the aircraft under control. The velocity obstacles constrain the maneuvering opportunities of the controlled aircraft in terms of potential loss of separation events. That is, if the velocity vector of the controlled aircraft lies within a triangular conflict zone, a loss of separation will occur in the near future. Vectoring the controlled aircraft outside such a conflict zone resolves the conflict. See [9] and [10] for more details on the design.

An example ATC "best practice" to resolve a crossing conflict in the horizontal plane, featuring two aircraft flying at different speeds, is to vector the slow aircraft behind the faster one. This is a typical "set-and-forget" strategy that minimizes monitoring time [19], [20]. Fig. 2(b) illustrates why this "best practice" is indeed a robust solution, one that requires less monitoring. That is, the available solution space on the right-hand side of aircraft A's maneuvering envelope is much richer than on the left-hand side. Additionally, placing the speed vector of aircraft A outside the velocity obstacle involves a small heading change to the right, making this a quick solution to resolve the conflict. Hence, the SSD has an explanatory value by making the best

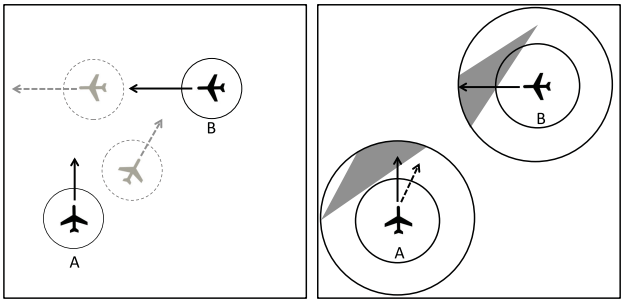


Fig. 2. SSD explains and re-enforces the ATC "best practice," potentially encouraging the development of control expertise. (a) ATC "best practice": put slow aircraft A behind faster aircraft B. (b) SSD: put slow aircraft A behind faster aircraft B.

practice visually salient. This could stimulate the development of control expertise, because trainees can literally "see" the "complete picture" governing a conflict and can thus actively think about and evaluate the best practice.

C. Decision Ladder Analysis

To support the development of control expertise, one could argue to simply provide ATC trainees with a range of best practices in the form of instructions. We hypothesize, however, the SSD to have certain advantages over instructions alone in the development of control expertise. To illustrate this, an analysis of information-processing steps and resulting knowledge gains for both instructions and the SSD (as illustrated in Fig. 2) has been carried out and mapped onto Rasmussen's DL [12].

The DL provides a qualitative model of human decision making in problem-solving activities and is defined by a sequence of knowledge states (circles) and information-processing actions (boxes) (see Fig. 3). In general, problem solving starts at the lower left corner when a worker is confronted with a certain "problem." Subsequent information processing would then enable the worker to gain a more detailed understanding of the problem at hand and thus reach higher levels in the DL. After that, several goal-oriented solution options are considered in an iterative cycle of knowledge-based behavior (KBB), followed by a selection of a specific solution and finally leading to the implementation of that solution. The left-hand side of the ladder thus represents problem analysis, whereas the right-hand side constitutes planning and executing solutions. Note that experienced workers rarely follow this sequence in a linear fashion. Based on earlier experiences, they can either skip steps (i.e., knowledge leaps) or make (rule based) shortcuts between the left- and right-hand sides of the ladder. A shortcut that directly connects the "activation" and "execute" boxes of the DL represents skill-based behavior (SBB), which features unconscious automated sensorimotor responses to stimuli. This behavior is commonly associated with highly experienced experts.

Decision support tools, but also instructions, could support novices to reach higher knowledge states as well as make certain shortcuts salient. For example, in Fig. 3(a), it can be seen that best practice instructions primarily encourage rule-based

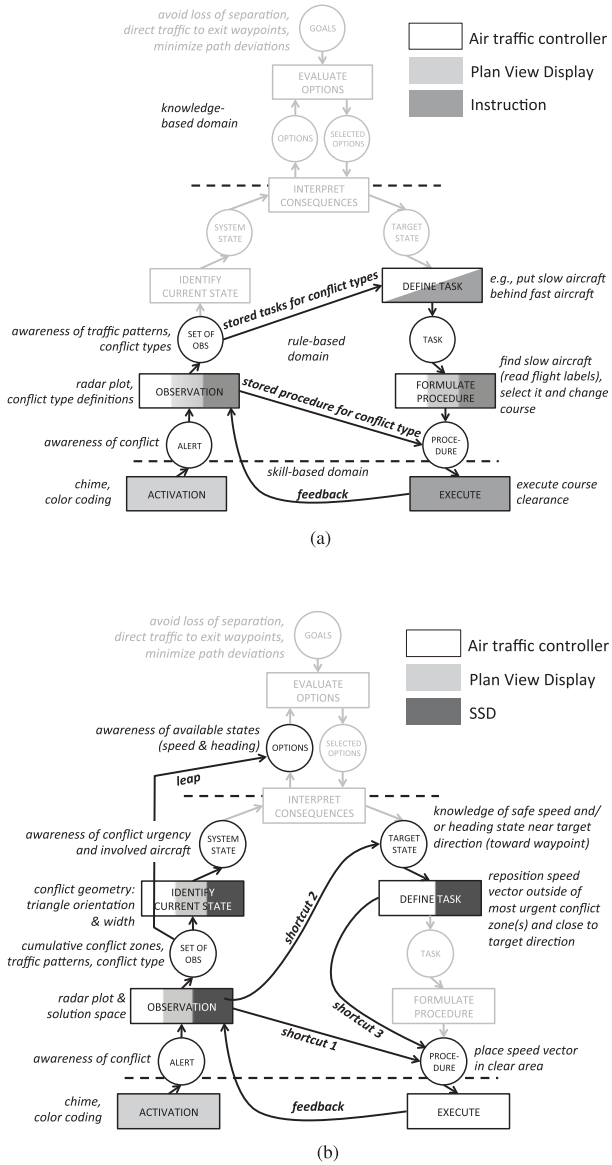


Fig. 3. Decision-making behavior supported by "best practice" instructions versus the SSD in conjunction with the PVD. (a) Instruction-based decision making. (b) SSD-based decision making (without instructions).

shortcuts that directly connect learned conflict geometries to task execution. A virtue of rules is that learning them by heart would facilitate swift decision making, but not necessarily provide support in analyzing the situation. In contrast, the SSD provides more support for the left-hand side of the ladder, corresponding to knowledge-based analysis of a traffic situation [see Fig. 3(b)]. It can support novices to reach higher knowledge states in the ladder, toward and into the region of KBB, whereas instructions would let novices remain more in the rule-based domain.

Decision aides that make rule-based shortcuts salient could result in novices showing expertlike behavior. However, this does not automatically mean novices will acquire the same level of expertise of experienced workers and they may not be able to properly handle situations where the support has been removed.

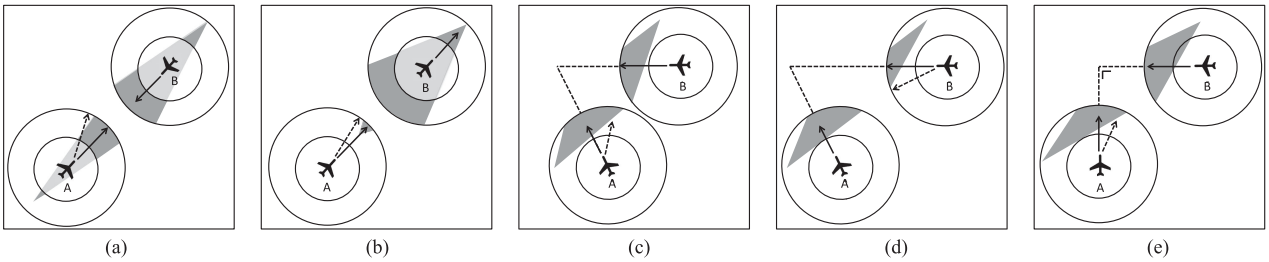


Fig. 4. Conflict types and their visualizations within the SSD. The length of the speed vectors indicate the aircraft speed magnitudes; the dashed speed vectors indicate the best practice solution to the conflict. The dashed line segments indicate the distance toward the crossing point of the aircraft pairs. (a) Head on (HON). (b) Overtake (OVR). (c) Crossing (CRO). (d) Crossing + distance bias (CRB). (e) Perpendicular (PER).

TABLE I
CONFLICT TYPES AND THEIR "BEST PRACTICES"

Conflict type	Heading difference [deg]	'Best practice' with aircraft speed difference	'Best practice' with equal aircraft speeds
Head on (HON)	170 - 180	Faster aircraft evades conflict	Either aircraft, depending on surrounding aircraft
Overtake (OVR)	0 - 10	Overtaking aircraft evades conflict	–
Crossing (CRO)	10 - 170	Slower aircraft evades conflict	Either aircraft, depending on surrounding aircraft
Crossing + bias (CRB)	10 - 170	Aircraft arriving later evades conflict	Aircraft arriving later evades conflict
Perpendicular (PER)	80 - 100	Slower aircraft evades conflict	Either aircraft, depending on surrounding aircraft

For example, in terms of task execution, the SSD provides several shortcuts from knowledge states toward the definition of tasks and procedures. It would thus depend on the individual user how the SSD will be used in the acquisition of deep knowledge about the traffic situation. That is, the SSD can solely be used as a rule-based tool [shortcuts 1 and 2 in Fig. 3(b)] to resolve a conflict, in which the shortcuts can be formulated as follows: "Direct the speed vector outside the conflict zone by a heading clearance." The risk of using the interface in such a fashion is that it cannot only lead to shallow knowledge and a dependence on the interface, as reported by Christoffersen *et al.* [8], but also lead to poor control performance (e.g., steer aircraft A in Fig. 2 in front of aircraft B). A person with a more analytical mindset would probably try to reflect on the feedback provided by the SSD (e.g., ask herself a question as "why is the conflict zone positioned and oriented in this fashion?"), reach higher regions in the ladder to gain deeper knowledge, and later use that knowledge to fall back to lower level rule-based shortcuts of better quality.

To facilitate proper decision making, mitigate potential large variability in SSD usage and encourage a deeper understanding of traffic situations; this study explored a hybrid approach where best practices are taught alongside the SSD. This would also ease measuring changes in achieved (higher level) knowledge states and help in analyzing to what extent that knowledge would persist after removing the SSD support.

III. EXPERIMENT DESIGN

A. Participants

TU Delft aerospace students were invited to participate voluntarily. After an intake questionnaire (probing their familiarity with ATC goals, displays, and practices) and a short skill test (several traffic stills for a conflict detection task), two balanced groups of eight participants (average age of 26 years; standard deviation of 1.9) were formed. All participants were familiar with the existence and looks of ATC radar displays and the overall ATC task (obtained from an introductory course in Avionics), but naive in terms of interpreting the radar display and CD&R best practices.

B. Instructions

All participants were given a mini lecture, in the form of a scripted PowerPoint slideshow, introducing ATC and the PVD and explaining the five best practices to paired aircraft conflicts (see Fig. 4 and Table I). These "best solutions" dictated in a specific conflict geometry of an aircraft pair what the best and most efficient action was to solve that conflict. The solutions to the conflicts were distilled from general rules of thumb that were adapted from research about controller strategies (e.g., [18] and [20]) and feedback from external experts on ATC training programs. These 'best solutions' provide a straightforward and quick fix for a conflicting pair of aircraft.

When practicing in the simulation environment, the task of the participant was to first guarantee safe separation of aircraft at all times by solving or preventing conflicts, and secondly to vector aircraft as efficiently as possible toward their respective exit waypoint. These goals and their priority closely resembled ATC practices.

Finally, participants were instructed to "think aloud" during *all* simulator sessions. Specifically, they were asked to mention the conflict type, the callsign of aircraft involved in the conflict, the aircraft they selected to resolve the conflict, and the type of solution. This allowed us to gain insight of their decision-making strategies and map those onto the DL after the experiment. It also allowed for classification of participants' decision-making behavior in terms of SBB, rule-based behavior (RBB), and KBB.

C. Independent Variables and Scenarios

The two independent variables were *training* (two levels), i.e., best practice instructions with or without the SSD and the *traffic*

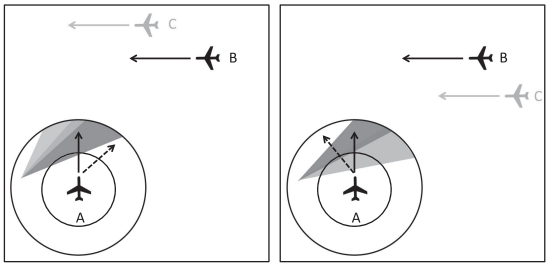


Fig. 5. Example of a three-aircraft scenario where the third aircraft (C) either amplifies, or requires deviation from, the best practice. (a) Amplifying best practice. (b) Deviating from best practice.

scenarios (five levels). Training with or without the SSD varied between participants and only applied to the training phase. After a transfer manipulation, both groups only had access to a baseline PVD to control traffic. Further, training participants in the five conflict types and their corresponding solutions (see Fig. 4 and Table I) featured aircraft pairs without any other traffic. Traffic scenarios were varied within participants, and this was realized by changing the order of appearance of the rehearsal exercises. Hence, a mixed design was used. During subsequent training exercises, also scenarios with three aircraft were encountered. Two-aircraft scenarios always had one "best solution," whereas in the three-aircraft scenarios, the third aircraft could either strengthen that "best solution" *or* cause the original "best solution" to create a conflict with the third aircraft. This required the controller to deviate from the learned best practice (see Fig. 5).

The traffic scenarios have been developed with the help of two external experts on ATC training (the Netherlands), with the specific focus on the horizontal plane and vectoring of aircraft by heading clearances. These restrictions (or simplifications) of the scenarios aimed to prevent confounds caused by the increased number of conflict solution possibilities. Scenarios were classified by the type of conflict (one of the five learning goals), whether the aircraft pair in conflict flew at different speeds, and the number of aircraft surrounding the conflicting pair. The geometrical orientations of the conflicts were rotated or mirrored in order to keep scenarios unrecognizable despite their similar geometries. All traffic scenarios had predefined solutions, such that near-unbiased performance comparisons could be made, and the same feedback could be given afterward to participants.

Three types of training elements were designed: still conflict scenarios, short dynamic scenarios (90 s), and long dynamic scenarios (900 s). In the dynamic scenarios, participants could interact with the aircraft and provide heading clearances. The short scenarios were variations of the five learning goals with each a single conflict between two or three aircraft. The longer scenarios were a compilation of at least seven consecutive conflicts, all with different geometries. In these scenarios, multiple-foil aircraft were present in the sector to distract the participant, increase complexity, and also to force participants to consider multiple solutions when a conflict situation emerged. A new conflict would present itself (turn amber) at least 120 s after a previous conflict.

D. Control Variables

All traffic was limited to the two-dimensional horizontal plane on flight level 290 and all aircraft were of the same type, featuring a speed envelope ranging from 150 to 290 kn and a fixed rate-one turning performance. The sector size and shape (squared 50×50 nmi area) and waypoint names were constant in all scenarios. To limit the control problem dimensions, conflicts were solved by giving heading clearances to aircraft (i.e., vectoring). No speed and/or altitude changes could be commanded.

E. Dependent Measures

The experiment collected dichotomous performance data during each scenario for three choices made: correct/incorrect conflict recognition, correct/incorrect choice of aircraft, and correct/incorrect choice of direction of the solution for the conflict. As the solutions to each conflict problem were predefined, simple yes/no answers were noted and cumulative error percentages could be calculated for each participant group. Also, the response times of these three decisions were recorded in seconds (using time-stamped audio and video recordings). In case the recognition of a scenario or the choice for a solution was altered, these would be recorded separately as well. The response time after an initial action was then noted such that in this way the "penalty" time of the first incorrect choice was included. Other control performance measures included the number of heading clearances (before and after solving the conflict), how often a loss of separation occurred, and the total additional flown track miles.

To analyze differences in behaviors between the two participant groups, audio and video recordings (a video capture of the computer screen) of the measurement sessions were manually transcribed. To assist the transcription, each information processing step was linked to specific behavioral markers and events observed and measured in the simulation sessions (see Table II). From the DL analysis described in Section II-C and the experiment setup, four most likely DL traversals were identified and labeled as variants of RBB and KBB (see Fig. 6).

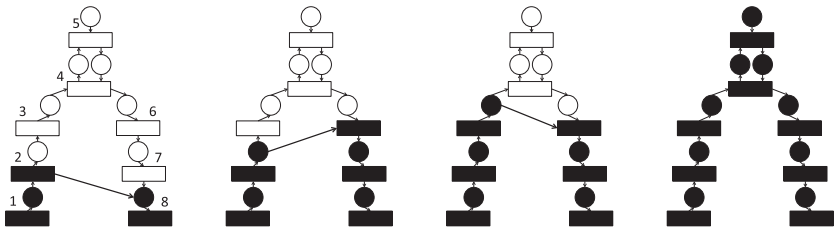
In Fig. 6, the first RBB type represents the "fastest" shortcut in which the observation of the traffic scenario immediately leads to an action, irrespective of the correctness of that action. RBB+ involves a more careful identification of the specific conflict type, followed by recalling the procedure involved to act upon the conflict. RBB++ also entails the evaluation of the conflict urgency and the identification of all aircraft involved in (solving) that conflict. Finally, KBB involves "thought experiments" where first (multiple) solutions are evaluated in terms of the higher order goals (i.e., adhere to target state, avoid new conflicts, and minimize path deviations).

F. Apparatus

Aircraft were simulated by linear kinematic equations and described by their position coordinates, velocities, and heading angles. Simulations ran on a desktop computer with a

TABLE II
IDENTIFICATION OF INFORMATION PROCESSING STEPS

Step in the DL	Behavioral markers: recognizing these steps during the simulation	Source of the marker
1. Activation	Spots a conflict, attention is drawn towards new conflict	Start of the conflict
2. Observe	Spots all (multiple) aircraft involved	Voice + Video + Cursor
3. Identify	Identifies the type of conflict	Voice
4. Interpret	Considers multiple options as solutions	Voice + Cursor
5. Evaluate	Considers safety of operation and space around aircraft	Voice
6. Define Task	Selects aircraft for the solution	Voice + Video + Mouse click
7. Form Procedure	Selects the direction of the solution	Voice + Video + Mouse click
8. Execute	Executes the solution	Video + Keyboard enter



(a) RBB (b) RBB+ (c) RBB++ (d) KBB

Fig. 6. Decision ladder sequences used to analyze control behavior of participants, ranging from fast RBB toward more laborious and slower KBB. The numbers in ladder (a) correspond to the information processing steps in Table II. (a) RBB. (b) RBB+. (c) RBB++. (d) KBB.

TABLE III
EXPERIMENT TRAINING PROCEDURE

Day 1 (morning)		Training Element	Duration
1.	Briefing	Mini lecture (scripted slideshow)	30 min
2.	Training	24 still scenes	25 min
3.	Training	Practice vectoring aircraft	5 min
4.	Training	8 Short dynamic scenes	25 min
5.	Training	2 Long dynamic scenes	30 min
		– full-day break –	
Day 2 (afternoon)			
6.	Training	8 Still scenarios (recap, with SSD)	10 min
7.	Training	8 Short dynamic scenes (no SSD)	25 min
8.	Training	2 Long dynamic scenes (no SSD)	30 min
9.	Measurement	5 Short dynamic scenes (no SSD)	15 min
10.	Measurement	1 Long dynamic scenes (no SSD)	15 min
11.	Debriefing	Retrospective questionnaire	20 min

30-in HD display with a resolution of 2560 × 1600 pixels and a refresh rate of 60 Hz. Interaction with aircraft was done by direct manipulation using a computer mouse and keyboard. No voice communication was required to command heading changes to aircraft, as this would interfere with the "think-aloud" task of the participants.

G. Procedure and Data Analysis

For each participant, the experiment was divided into two half-day sessions (including breaks), with exactly one day in between and balanced to different times during the day (morning or afternoon) (see Table III). On the first half-day, participants received the 30-min mini lecture. For the ecological group, an additional slide was presented that explained the SSD and the five best practices were shown in conjunction with the SSD.

The exercises following the briefing first featured 24 still scenarios with just 2 aircraft that were either in conflict or not. For the SSD group, the SSDs for both aircraft were shown. To prepare for the dynamic scenes in which participants could interact with aircraft and resolve conflicts, a short 5-min session was dedicated to vectoring an aircraft using the mouse cursor device and the ENTER key. Day 1 was concluded with eight short dynamic scenes, each lasting 90 s—first four scenes contained two aircraft and last four scenes contained three aircraft—and two long scenarios. Here, the short dynamic scenes with three aircraft all reinforced the learned best practice. Note, the ecological group always had access to the SSD in addition to the baseline PVD during all still and dynamic exercises, whereas the instructional group only had access to a baseline PVD.

Day 2 started with a short recap exercise featuring still scenarios with just two aircraft. After that, the transfer manipulation was done for the SSD group in which the remaining exercises (eight short dynamic scenes and two long dynamic scenes, same as in Day 1) featured a baseline PVD. The final measurement session consisted of five short dynamic scenes, all with three aircraft and five conflict types that either reinforced the best practice or required deviation (see Section IV).

On analyzing the results, it was decided to omit the data analysis from the long scenarios, because the large variability in the evolution of traffic situations generally makes the results very difficult to compare between participants and groups. A total of 2 h of video and audio recordings was analyzed. First, all video and audio material were anonymized by removing any reference to a particular participant group, allowing for a double blind analysis. After that, all material was transcribed. The transcription was carried out by two evaluators (both familiar with the DL) to ensure unbiased results as much as possible. Given the relatively low sample size for each experimental condition, conservative nonparametric tests were used to compare the control performance between the two participant groups. Kruskal–Wallis and Friedman tests were applied to analyze between- and within-group effects, respectively.

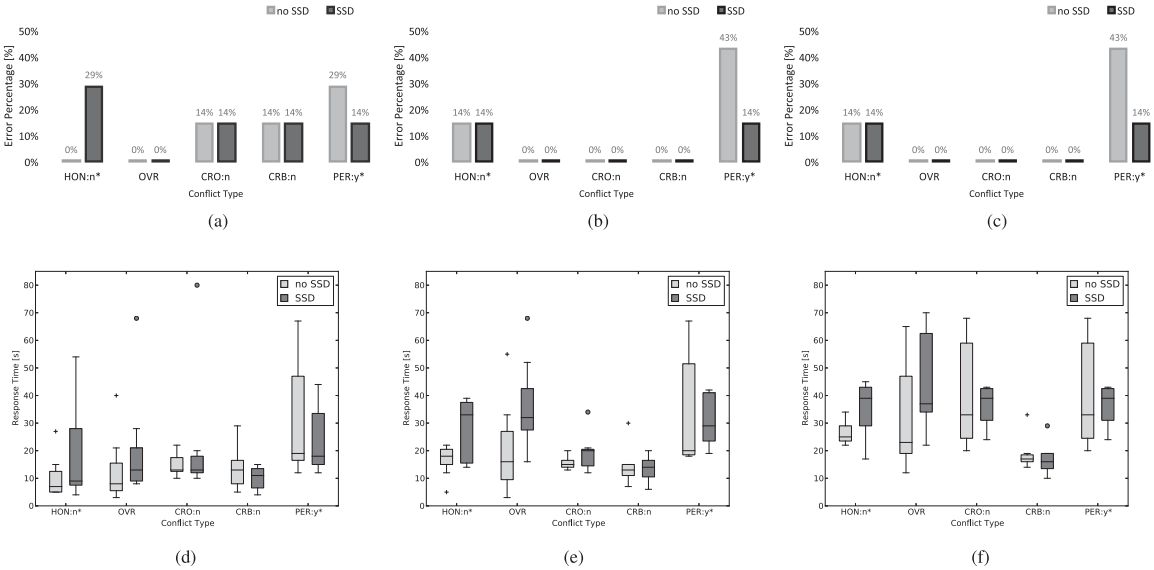


Fig. 7. Cumulative error percentages and response times of conflict type recognition, aircraft choice and solution response. The * symbol indicates deviation from the best practice and "y" or "n" designates the presence of a speed difference 'yes' or 'no.' (a) Incorrect conflict recognition. (b) Incorrect aircraft choice. (c) Incorrect solution choice. (d) Correct conflict recognition response time. (e) Correct aircraft choice response time. (f) Correct solution choice response time.

H. Hypotheses

First, it was hypothesized that training with the SSD would incite more higher level cognitive behavior (i.e., RBB++ and KBB), because it encourages participants to attend to the complete traffic situation instead of just the conflict pair and the specific rules. In other words, training without the SSD would encourage participants to only "find the right rule" to solve a conflict, whereas training with the SSD was expected to encourage participants in gaining insight into the situation, evaluate the learned rules, and pick the best one. Second, the SSD group was expected to make less mistakes in conflict-type recognition, choosing the correct aircraft to solve the conflict and implementing the correct solution, albeit at the cost of higher response times. Third, in "novel" scenarios that required deviations from the best practice due to the presence of a third aircraft, the SSD group was expected to develop "new rules" and better handle those situations than the "instructional" group.

IV. Results

In the short scenario sessions, no loss of separation events occurred as all participants managed to keep aircraft separated more than 6 nmi. Two participants were removed from the analysis, because they caused significant outliers and showed deviating behavior. Fortunately, they belonged to a different group, yielding two balanced groups of each seven participants.

A. Conflict Detection and Resolution

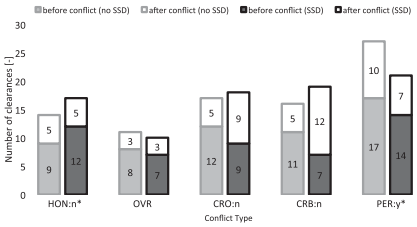
The control performance results, shown in Fig. 7, reveal trends in support of the second and third hypotheses, with overall larger response times for the SSD group and observed improvements for this group in the "novel" traffic scenarios (i.e., HON and PER). However, Kruskal–Wallis tests did not find any significant difference between the two participant groups on all aspects of the CD&R control performance.

Friedman tests did find significant differences for the within-group manipulations, i.e., the conflict types. A significant effect of conflict type on the recognition response time was found ($\chi^2(4) = 19.476$, $p < 0.01$). Pairwise comparisons (with Bonferoni correction) revealed that the PER scenario had significant longer response times than the HON scenario, despite that both were "novel" scenarios. Also a significant increase in aircraft choice response time was found for the SSD group in conflict-type recognition ($\chi^2(4) = 15.928$, $p = 0.03$). Pairwise comparisons revealed that CRB was significantly different from OVR and PER. The solution response time showed similar results, with a significant effect for conflict type ($\chi^2(4) = 29.789$, $p < 0.01$). Here, pairwise comparisons showed a significant difference between CRO and CRB, OVR and CRB, and PER and CRB.

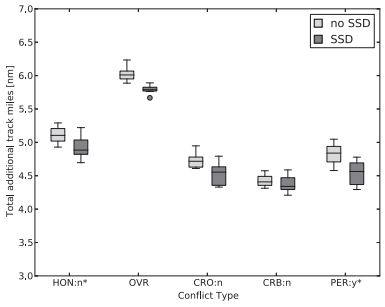
B. Control Efficiency

Fig. 8 shows the total number of heading clearances and the total additional track miles flown by aircraft. The heading clearances were split by clearances before (i.e., commanding an evasive maneuver) and after (i.e., commanding aircraft back toward their target waypoint) the conflict. The additional track miles were determined by the flown distances relative to the shortest paths toward the exit points.

In general, the two participant groups did not perform significantly different in terms of number of heading clearances. However, there was a significant group effect on the total additional track miles across the majority of conflict types (HON: $H(1) = 3.922$, $p = 0.048$; OVR: $H(1) = 9.016$, $p < 0.01$; CRO: $H(1) = 4.445$, $p = 0.035$; PER: $H(1) = 4.446$, $p = 0.035$). There was also a significant difference in track miles



(a)



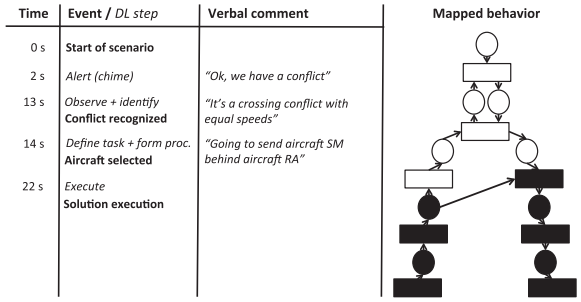
(b)

Fig. 8. Total number of heading clearances and additional track miles. The * symbol indicates deviation from the best practice and "y" or "n" designates the presence of a speed difference "yes" or "no." (a) Heading clearances. (b) Additional track miles.

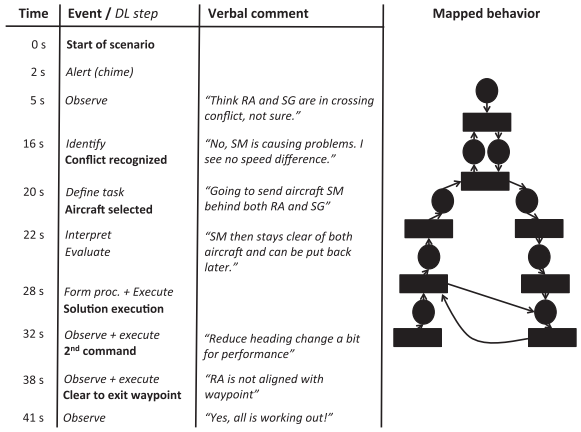
between conflict types ($\chi^2(4) = 48.514, p < 0.01$), which is not surprising given the different traffic geometries. The reduction in additional track miles [see Fig. 8(b)] for the SSD group indicates that the heading clearances were of better quality by adhering more closely to the aircraft shortest routes toward the exit waypoints. Interestingly, the CRB conflict type was the only one that resulted in similar additional track miles for the two groups. In this scenario, several participants in the SSD group had a tendency to put aircraft on their designated course too early after solving the conflict, occasionally requiring corrective actions to stay clear of the conflict that was initially solved. The instructional group showed less corrective heading adjustments, especially after the conflict was solved. They either waited longer before clearing aircraft to their exit waypoints or not do this at all. This resulted in slightly less heading clearances after the conflict, but more additional track miles.

C. Decision-Making Behavior

In Fig. 9, two transcripts from two different participants are shown for the CRO scenario along with their mapped behavior. As can be seen in this figure, it was common to find more than one behavioral type within one trial per participant. For example, Fig. 9(b) indicates that the participant first engaged in KBB by evaluating a possible solution, but later, when executing a second control action, a rule-based shortcut was made representing RBB from Fig. 6(a). Instead of counting this as both KBB and RBB, we decided to only count the first identified behavior (i.e., KBB) as we believe this to be most telling and meaningful of how a participant started to approach a (new) scenario. Additionally, it was also common to find iterations within one behavioral type,



(a)



(b)

Fig. 9. Example transcripts and their matching behavior types for the crossing (CRO) scenario. (a) Best matching behavior type: RBB+. (b) Best matching behavior type: KBB.

e.g., several attempts to detect the correct conflict type (RBB+), which were counted separately.

The resulting distribution in decision-making behavior, ranging from fast RBB toward slower KBB (with and without observed iterations), is provided in Fig. 10. From this figure, it can be observed that, overall, the SSD group reached higher in the DL than the instructional group, as hypothesized. Participants who trained with the SSD evaluated potential solutions against secondary ATC goals (e.g., adhering to target waypoints and avoiding new conflicts) rather than just on solving the conflict. This can also be observed from the transcripts provided in Fig. 9, where Fig. 9(b) is from a participant in the SSD group. The reduced additional track miles for the SSD group also supports this observation.

Participants in the SSD group consistently showed more RBB++ and KBB counts than participants in the instructional group, irrespective of conflict type. This was not only the case in the "easier" scenarios (e.g., see the CRO transcripts in Fig. 9), but also in the novel PER scenario as shown in Fig. 10(b). In this scenario, the SSD group did make less mistakes compared to the instructional group. Note that none of the participants showed the lowest level of RBB behavior. This type of behavior is commonly reserved to highly experienced controllers, and the short duration of our experiment apparently did not allow students to reach that level of expertise.

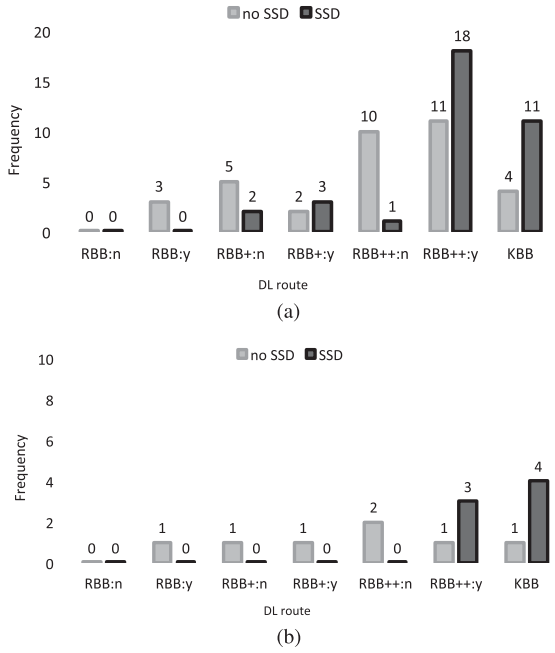


Fig. 10. Histogram of the observed decision-making behavior, where "y" or "n" indicates the presence of iterations within the behavior type. (a) Overall decision-making behavior. (b) Decision-making behavior in PER: y*.

V. Discussion

The gist of quantitative and qualitative results suggests that the SSD changed how participants thought and approached their control problem, even after removing the ecological support. Although no significant differences were observed on the primary task performance (i.e., CD&R), the control efficiency and decision-making results indicate that the SSD group was also more attentive to secondary ATC goals (i.e., adhering to target waypoints and minimizing path deviations) and better handle the "novel" PER scenario. Similar functional and goal-oriented behavior and knowledge organization was found by Christoffersen *et al.* [8].

From these results, however, the long-term impact and benefits for employing ecological interfaces in (ATC) training are difficult to predict. For example, specific to ATC experts is that they try to mitigate their own cognitive workload, but the SSD group revealed longer response times and more KBB. Although these results may be explained by the increased attentiveness to secondary ATC goals, it cannot be ruled out that also some confusion, caused by potential "compatibility issues" [21], [22] arisen after the transfer manipulation, contributed to the increased cognitive load. Only a longitudinal study would be able to shed light on how far this effect would persist after a prolonged period of time.

The conflict-type effects for the SSD group are also interesting, suggesting the SSD is context sensitive (regarding learning). In Fig. 4(a) and (b), the SSD patterns for HON and OVR conflicts make the best practice solution indeed less salient than for the crossing conflict types, given the orientation of the conflict zones. However, these results are largely consistent with an earlier study in ATC conflict detection, which showed that conflicts with "large convergence angles" and "short conflict times" (i.e., HON) and "small angles" and "long times" (i.e., OVR) are more difficult to detect than conflicts with "small angles" and "short conflict times" (i.e., CRO) [23]. In that sense, the SSD did not trivialize an intrinsically difficult problem, which is a common experience with ecological displays [5].

Our study also had limitations that need to be mentioned and addressed in future studies. First, transcribing the audio and video material was a laborious process. Although this process was carried out by two evaluators in a double-blind fashion, the results depended on how well participants articulated their decision-making behavior and how this was interpreted. In addition, the verbal comments did not always follow a chronological order relative to the sequence of steps in the DL. No reliability test was undertaken by a separate analyst, which could be identified as a limitation in our exploratory study. For future work, developing a set of standardized verbal protocols and matching the steps in the decision ladder could make this process more streamlined.

Second, we did not take into account specific instructional design methods in teaching students to comprehend and take full advantage of the SSD. After a brief explanation of the SSD and the best practices, students needed to rely on some form of discovery learning. It cannot be ruled out that personal differences in learning style affected our results. Other studies have indeed shown that people who are "holists" tend to pick up information from an ecological interface more easily than "serialists" [24]. For future studies, it is recommended to include an instructional design method (e.g., 'visual scaffolding' [25]) that complements ecological interfaces to better guide the information pickup and learning process.

To conclude, the findings in our study have shed light on broader implications for EID in terms of training requirements and its potential role in addressing expertise degradation. As mentioned by Christoffersen *et al.* [8], an ecological interface only contributes to *proper* knowledge acquisition when operators actively reflect on its visual feedback. Our experiment was specifically geared toward training, thereby creating a learning environment that encouraged evaluation and explorative thought experiments next to the industry "gold standards." This appeared to have effect, implying that ecological displays would *always* require training before users can take advantage of the visual feedback in a way that can build and/or retain expertise.

VI. Conclusion

We investigated the short-term effects of training a group of novices (in air traffic CD&R) with an ecological interface and compared the control performance and decision-making behavior with a group that only received instructions. An experiment was conducted wherein two groups underwent a two-day training program featuring a transfer manipulation in the final measurement scenarios on the second day. Results show that the primary task performance between the "ecological" group and the "instructional" group was not significantly different. Interestingly, students in the ecological group exhibited more laborious RBB and KBB that sparked goal-oriented thoughts and corresponding control performances beyond the primary task. These findings indicate that ecological interfaces can change

how people think and approach a control problem, even after removing the support. It is therefore reasonable to believe that ecological interfaces can play an important role in the development of deeper knowledge.

REFERENCES

[1] L. Bainbridge, "Ironies of Automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

[2] B. Strauch, "Ironies of automation: Still unresolved after all these years," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 5, pp. 419–433, Oct. 2018.

[3] K. J. Vicente and J. Rasmussen, "The ecology of human-machine systems II: Mediating direct-perception in complex work domains," *Ecological Psychol.*, vol. 2, no. 3, pp. 207–249, 1990.

[4] K. J. Vicente and J. Rasmussen, "Ecological interface design: Theoretical foundations," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 4, pp. 589–606, Jul./Aug. 1992.

[5] C. Borst, J. M. Flach, and J. Ellerbroek, "Beyond ecological interface design: Lessons from concerns and misconceptions," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 2, pp. 164–175, Apr. 2015.

[6] R. C. McIlroy and N. A. Stanton, "Ecological interface design two decades on: Whatever happened to the SRK taxonomy?" *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 2, pp. 145–163, Apr. 2015.

[7] K. Christoffersen, C. N. Hunter, and K. J. Vicente, "A longitudinal study of the effects of ecological interface design on skill acquisition," *Human Factors*, vol. 38, no. 3, pp. 523–541, Sep. 1996.

[8] K. Christoffersen, C. N. Hunter, and K. J. Vicente, "A longitudinal study of the effects of ecological interface design on deep knowledge," *Int. J. Human-Comput. Stud.*, vol. 48, no. 6, pp. 729–762, 1998.

[9] C. Borst, V. A. Bijsterbosch, M. M. van Paassen, and M. Mulder, "Ecological interface design: Supporting fault diagnosis of automated advice in a supervisory air traffic control task," *Cognition, Technol. Work*, vol. 19, no. 4, pp. 545–560, Nov. 2017.

[10] S. B. J. van Dam, M. Mulder, and M. M. van Paassen, "Ecological interface design of a tactical airborne separation assistance tool," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 6, pp. 1221–1233, Nov. 2008.

[11] C. Borst, R. M. Visser, M. M. van Paassen, and M. Mulder, "Ecological approach to train air traffic control novices in conflict detection and resolution," in *Proc. Sixth SESAR Innov. Days*, 2016, no. Nov., pp. 1–9.

[12] J. Rasmussen, *Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering*. Amsterdam, The Netherlands: North Holland, 1986.

[13] Federal Aviation Administration, "Review and evaluation of air traffic controller training at the FAA Academy," U.S. Dept. Transp., Federal Aviation Admin., Washington, DC, USA, Tech. Rep. AV-2008-055, 2013.

[14] M. Schuver-van Blanken, "Clarifying cognitive complexity and controller strategies in disturbed inbound peak ATC operations," in *Advances in Aviation Psychology*, M. A. Vidulich, P. S. Tsang, and J. M. Flach, Eds. Farnham, U.K.: Ashgate, 2014, ch. 6, pp. 85–102.

[15] J. D'Arcy and P. Della Rocco, "Air traffic control specialist decision making and strategic planning—A field survey," U.S. Dept. Transp., Federal Aviation Admin., Tech. Rep. DOT/FAA/CT-TN01/05, 2001.

[16] S. Loft, S. Bolland, M. S. Humphreys, and A. Neal, "A theory and model of conflict detection in air traffic control: Incorporating environmental constraints," *J. Exp. Psychol.: Appl.*, vol. 15, no. 2, pp. 106–124, Jun. 2009.

[17] E. M. Rantanen and A. Nunes, "Hierarchical conflict detection in air traffic control hierarchical conflict detection in air traffic control," *Int. J. Aviation Psychol.*, vol. 15, no. 4, pp. 339–362, 2005.

[18] E. M. Rantanen and C. D. Wickens, "Conflict resolution maneuvers in air traffic control: Investigation of operational data," *Int. J. Aviation Psychol.*, vol. 22, no. 3, pp. 266–281, Jul. 2012.

[19] S. Fothergill and A. Neal, "Conflict-Resolution heuristics for en route air traffic management," in *Proc. 57th Annu. Meeting Human Factors Ergonom. Soc.*, 2013, vol. 57, no. 1, pp. 71–75.

[20] B. Kirwan and M. Flynn, "Investigating air traffic controller conflict resolution strategies," EUROCONTROL, Brussels, Belgium, Rep. ASA.01.CORA.2.DEL04-B.RS, Mar. 2002.

[21] N. Lau, G. A. Jamieson, G. Skraaning, and C. M. Burns, "Ecological interface design in the nuclear domain: An empirical evaluation of ecological displays for the secondary subsystems of a boiling water reactor plant simulator," *IEEE Trans. Nucl. Sci.*, vol. 55, no. 6, pp. 3597–3610, Dec. 2008.

[22] Y. Li, X. Wang, and C. M. Burns, "Improved monitoring performance of financial trading algorithms using a graphical display," in *Proc. Annu. Meeting Human Factors Ergonom. Soc.*, 2018, pp. 187–191.

[23] R. W. Remington, J. C. Johnston, E. Ruthruff, M. Gold, and M. Romera, "Visual Search in complex displays: Factors affecting conflict detection by air traffic controllers," *Human Factors*, vol. 42, no. 3, pp. 349–366, Sep. 2000.

[24] G. L. Torenvliet, G. A. Jamieson, and K. J. Vicente, "Making the most of ecological interface design: The role of individual differences," *Appl. Ergonom.*, vol. 31, no. 4, pp. 395–408, 2000.

[25] J. J. G. van Merriënboer, P. A. Kirschner, and L. Kester, "Taking the load off a learner's mind: Instructional design for complex learning," *Educ. Psychol.*, vol. 38, no. 1, pp. 5–13, 2003.

Clark Borst received the M.Sc. (*cum laude*) and Ph.D. degrees in aerospace engineering from Delft University of Technology, Delft, The Netherlands, in 2004 and 2009, respectively.

He is currently an Assistant Professor in aerospace human–machine systems with Delft University of Technology. His work and research interests lie in developing human-centered aviation automation through the application and empirical evaluation of cognitive systems engineering and ecological interface design principles.

Roeland M. Visser received the M.Sc. degree in aerospace engineering from Delft University of Technology, Delft, The Netherlands, 2016, for his work in ecological interface design and training for air traffic control.

He is currently an Airport Consultant with Netherlands Airport Consultants, The Hague, The Netherlands, on strategic and tactical studies that include operational analyses, air traffic forecasting, flight schedule predictions, organization structuring, and operational readiness for airports around the world.

Marinus (René) M. van Paassen (M'08–SM'15) received the M.Sc. and Ph.D. degrees in aerospace engineering from Delft University of Technology (TU Delft), Delft, The Netherlands, in 1988 and 1994, respectively, for his studies on the role of the neuromuscular system of the pilot's arm in manual control.

He is currently an Associate Professor with the section Control and Simulation, Aerospace Engineering, TU Delft, working on human–machine interaction and aircraft simulation. His work on human–machine interaction ranges from studies of perceptual processes and manual control to complex cognitive systems.

Dr. van Paassen is currently an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS.

Max Mulder (M'14) received the M.Sc. and Ph.D. (*cum laude*) degrees in aerospace engineering from Delft University of Technology (TU Delft), Delft, The Netherlands, in 1992 and 1999, respectively, for his work on the cybernetics of tunnel-in-the-sky displays. He is currently a Full Professor and the Head of the section Control and Simulation, Faculty of Aerospace Engineering, TU Delft. His research interests include cybernetics and its use in modeling human perception and performance and cognitive systems engineering and its application in the design of "ecological" human–machine interfaces.

Prof. Mulder is currently an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS.

Drilling Into Dashboards: Responding to Computer Recommendation in Fraud Analysis

Natan Morar, Chris Baber, Faye McCabe, Sandra D. Starke, Inna Skarbovsky, Alexander Artikis, and Ivo Correai

***Abstract*—Credit card fraud analysis is almost entirely automated. However, there may be occasions when a human analyst is required to intervene. In this paper, we consider situations in which a transaction triggers an automated alert but not sufficiently to allow automated response. On such occasions, automated analysis makes a recommendation as to the fraud pattern that has been identified and the human analyst decides, if this recommendation is correct and what action to take. In order to support the analyst, a "dashboard" can be used to display information that is relevant to the fraud pattern. Thus, a computer could analyze transaction data, define this as a known fraud pattern, and then present the analyst with a dashboard to illustrate how the transaction might fit the pattern. We explore the efficiency with which people respond to the computer's recommendation, and whether computer confidence has an impact on this response. We define efficiency in terms of information search: the user will either have the relevant information on screen or will need to drill-into a dashboard (i.e., open additional windows for information). The results show that participants adapt their decision making to the confidence of the automated support, and, although they drill-down even when not required, are efficient in terms of time spent looking at relevant information.**

***Index Terms*—Automation transparency, credit card fraud, decision support systems, human-automation interaction.**

I. Introduction

THE analysis of credit card fraud has become highly automated in recent years [5], [8], [9], [13], [30]. However, there remain situations in which a human analyst might be required to contact a cardholder to check a transaction or to review the decisions made by automated systems. For example, in the EU-funded SPEEDD project, online inductive logic programming (using Online Learning of Event Definitions (OLED) [14]) and online complex event processing (using IBM's PROTON)[1] is applied to credit card transaction data[2] in order to learn relational patterns in fraudulent activity [1]. In this project, the set of fraud patterns is as follows:

1) increasing amounts of money spent on a single card over a sequence of transactions (increasing amount—IA);
2) using a card to make a very large transaction, compared to the average transaction in the region (large amount—LA);
3) a very high number of transactions in a very short time period (flash attack—FA);
4) transactions in geographically remote places in a short time period (transactions in faraway places—TF).

In credit card fraud investigation, automation analyzes and flags suspicious transactions. Ideally, the automated system would process a transaction and make a decision in milliseconds. When automation confidence (or, certainty) is very high, transactions are labeled as fraud and cards are automatically blocked. When there is no indication of fraud, the transaction is allowed. However, there are situations where there is some indication of fraud but not enough evidence to take an automated action, e.g., there might be a new form of fraud, or the data do not quite meet the threshold required [13]. In such cases, the analyst might review the computer's recommendation and then decide whether to call the cardholder to obtain additional information, or permit, or block the transaction. Visualizations of information relevant to the transaction can aid the analyst. Such visualizations can be designed on an analogy with the "dashboard" of an automobile: *"When properly designed… dashboards support a level of awareness… that could never be stitched together from traditional reports"* [11, p. 5]. A dashboard offers a high-level perspective on the situation, with the opportunity to drill-down to lower levels of detail, such that interaction follows discrete stages, e.g., *"Overview first, zoom and filter, then details on demand"* [31] or *"Analyze first, show the important, zoom/filter, analyze further, details on demand"* [15].

Some fraud patterns require contextual information for proper diagnosis, while others require the investigation of transaction-specific "detailed" information. For fraud analysis, contextual information could include a transaction suspiciousness score, which could be a probability between 0 and 1 based on the bespoke algorithm of a particular organization, together with the

Manuscript received August 6, 2018; revised May 13, 2019; accepted June 11, 2019. Date of publication July 29, 2019; date of current version November 21, 2019. This work was supported by the European Union FP7 project SPEEDD under Grant 619435. This paper was recommended by Associate Editor K. M. Feigh. *(Corresponding author: Chris Baber.)*

N. Morar, C. Baber, F. McCabe, and S. D. Starke are with the University of Birmingham, Birmingham B15 2TT, U.K. (e-mail: nsm120@student.bham.ac.uk; c.baber@bham.ac.uk; fxm493@student.bham.ac.uk; s.d.starke@bham.ac.uk).

I. Skarbovsky is with the IBM Research Haifa, Haifa 3498825, Israel (e-mail: inna@il.ibm.com).

A. Artikis is with the NCSR Demokristos, Athens 15310, Greece (e-mail: a.artikis@gmail.com).

A. Artikis is with the Feedzai, Lisbon 1050-045, Portugal (e-mail: ivo.correai@feedzai.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2925619

[1] https://github.com/ishkin/Proton/

[2] The dataset used for this paper can be publicly accessed from the project website http://speedd-project.eu/data

type of transaction made, e.g., cardholder present (in a store) or cardholder not present (online), time of day, geographical region of transaction, customer profile, and spending behavior [7], [22], [29], [30]. It is not clear whether it is more important to show the context in which a suspected fraudulent transaction has occurred (i.e., overview) or show the specific flagged transaction (i.e., details), as different fraud types are diagnosed using different types of information [8], [9], [29]. Moreover, it is not known how different modes of interaction affect the analyst's information search. People tend to optimize information search, depending on the task they are required to perform and on information accessing costs, so that relevant information gain per unit time is maximized [6]. From the concept of "information foraging" [25] we assume that people filter out irrelevant information sources and select relevant ones based on previous searches. By analogy, the specific information sources used by the automation could be shown to the human analysts and would have (because they are on the display) a low access cost. Other information could be accessed through pop-up windows, which would have higher access costs. The four fraud types outlined above would place different weight on the relevance of available information. In this study, we investigate whether presenting participants with information relating to these specific fraud types, first, decreases their decision time and, second, changes their information search behavior (in terms of drill-down activity), and third, whether the confidence of the automation influences these measures.

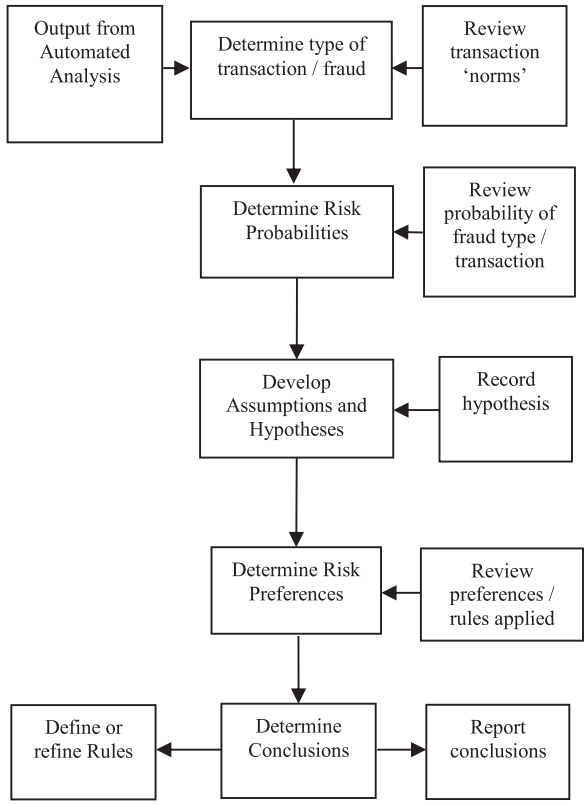


Fig. 1. Possible decision process for (human) credit card fraud analysis used in this paper.

II. Credit Card Fraud Analysis

The credit card industry is understandably very protective of the approaches used in the analysis of credit card fraud. While we have benefitted from discussions with a number of fraud analysts operating in the U.K., Europe, and the USA, the following description presents a high-level account of decision making in which analysts engage. This does not represent analysis conducted by any individual organization, but a general description of how analysis is approached. The description of this process underpins our experimental design.

Fig. 1 shows a decision process for credit card fraud analysis. In terms of the system output, a transaction suspiciousness or risk score will be based on probabilities that are defined in terms of an organization's risk models. The risk model will be tailored for specific types of client, region, transaction, etc., but could include such measures as number of transactions for an account in a given time period, value of a transaction, number of cash withdrawals at automated teller machines, etc. The risk models inform the design and operation of algorithms used by the automated system.

Credit card fraud can involve several types of analysts, from those involved in the definition and running of machine learning algorithms on big datasets to those who respond to alerts from the fraud detection system. In cases involving response to alerts, call-center analysts will triage these according to risk level; alerts with the highest level of risk are worked on first (i.e., "priority mode"). From our interviews, we believe that credit card organizations can employ between 250 and 1500 fraud analysts, and the total number of cases to be processed by an analyst is around 200 per day. In terms of the baseline decision time, a typical decision by a call handler might take around 1.8 min (assuming 200 cases to investigate in a 7.5 h working day, with 1.5 h of breaks). Given that the handling of a case involves blocking or allowing the transaction (and completion of forms for audit purpose) or speaking to the cardholder prior to making the decision (and so involve a telephone call as well as form filling), then one might anticipate the decision on the fraud (based on the information provided) to be performed relatively quickly. Consequently, having a system that supports quick but accurate decision making could be advantageous. Thus, a well-designed dashboard, together with a reliable recommender system, ought to minimize time spent on each call.

If a transaction meets the criteria to which the algorithms apply, then it would be automatically blocked. However, if only some of the criteria were met or if there was some uncertainty concerning the criteria, the transaction might be presented to a human analyst. In this paper, we are interested in the role of call-center analysts, who perform customer verification on suspicious transactions. In general, the decision to contact a customer would be made if the automated system was able to match some but not all of its criteria. In this instance, the call-center analyst would be presented with some of the transaction details together with the automated output (in the form of a score). In Fig. 1, this activity would involve the top three boxes. The call-center analyst would interpret information from the dashboard in order to define the situation; if there was insufficient information, then the analyst would contact the cardholder. In the telephone call, the call-center analyst would follow a clearly defined script in order to establish whether the

person at the other end of the phone is the genuine cardholder and whether the cardholder made the purchase or whether it was made fraudulently. The call-center analyst would then confirm that the transaction was acceptable or mark it as suspicious. This could involve reimbursing the cardholder or could trigger further investigation. For this role, analysts have access to transaction and customer details, both past and present.

III. Humans and Automation

A. Recommender Systems and Transparency

Recommender systems provide suggestions for a specific action or solutions to a problem [28]. Automated reasoning on the data computes an answer and displays it to the user, e.g., in the form of a recommendation for an action to be taken, or in the form of a detected event. In the context of fraud handling, this can relate to the detection of aspects of transactions which are related to definitions of fraud (see above). "Transparency" is a defining factor of a "good" recommender system [34], i.e., the extent to which the computational process behind the recommendation is visible and clear to the human. It has been shown that increasing transparency of recommender systems, that is, making explanations available to the user along with recommendations, improves decision performance [26], [33], [34].

In this paper, we display recommendations in terms of a specific fraud that has been identified, with the computer's confidence in its recommendation, and the most relevant data.

B. Dealing With Automation Confidence

If the automation has high confidence in its recommendation, then the role of the human could be to confirm this. A well-supported observation is that human-automation system performance with "perfect" (i.e., high reliability) automation tends to be superior to that without automation, and performance with "imperfect" (i.e., low reliability) automation tends to be inferior to that without automation [16], [20]. What makes the findings of these studies counterintuitive is that participants, in the high-reliability conditions, will follow the advice of automation even when it is wrong. This could be interpreted as automation bias, where humans are complacent and not checking the automation output against other information sources [23], [24], i.e., the human merely accepts the automation's output and does not contribute to the decision making. Conversely, if the automation has low confidence, then the human might ignore useful output. Such automation bias has been interpreted in terms of operator characteristics, e.g., individual differences [10], experience of automation failing [4], decision accountability [33], or the design of the automation, e.g., level of automation [18], or use of information sources [21], [32].

From a review of automation reliability, Wickens and Dixon [36] conclude that there is a "cross-over" point at 70% reliability, below which unreliable automation was worse than no automation (in terms of human-automation system performance). However, operators might use unreliable automation in order to free up cognitive resources, which would explain reliance on unreliable automation under high workload [36], or operators might use the recommendation as a cue to apply their own heuristics [19].

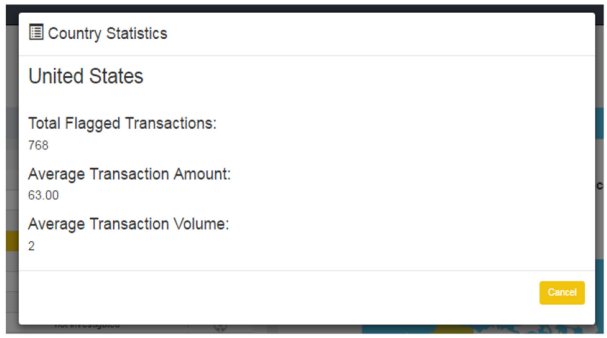


Fig. 2. Country pop-up.

It has been suggested that acceptance of (and, by implication, trust in) automated decision support can be considered in terms of "conformance" between human and automation problem-solving style [35]. In other words, if the automation appears to be addressing the problem in a manner that is similar to the one used by the human, then this could lead to higher level of acceptance by the human. For this paper, conformance involves the decision to block or allow a transaction on a credit card and the use of available information.

IV. Designing Dashboards for Fraud Analysis

In this experiment, dashboards are designed according to visualizations that we had observed to be common in fraud analysis; the aim was not to produce a single organization's display but to produce designs that had sufficient family resemblance across the domain. In addition, the dashboards needed to present the confidence that the computer applies to its classification of the fraud pattern, and to present the most relevant information in support of that classification.

The experiment also considered the need to drill down for information. By "drill down," we mean whether the participants need to call pop-up windows (see Figs. 2 and 3) to obtain further information. Some fraud types require contextual information, such as information about the normal card usage in the region where it occurred. When a transaction is outside the bounds of what is considered "normal," this is a potential indication of fraud. For other fraud types, specific transaction information is needed for diagnosis. For example, it is sufficient to know that one card has been physically used in two distant countries in a very short amount of time to suspect that the card may have been cloned.

In this experiment, we contrast an "overview" design (which summarizes transactions occurring in a particular country) with a "detailed" design (which provides a view of transactions on a given card). In each dashboard, it was possible to identify *some* fraud patterns solely on the information presented to the user, but there was the option to "drill down" to find further information—which meant that all of the fraud patterns used in this experiment could be identified using both dashboard

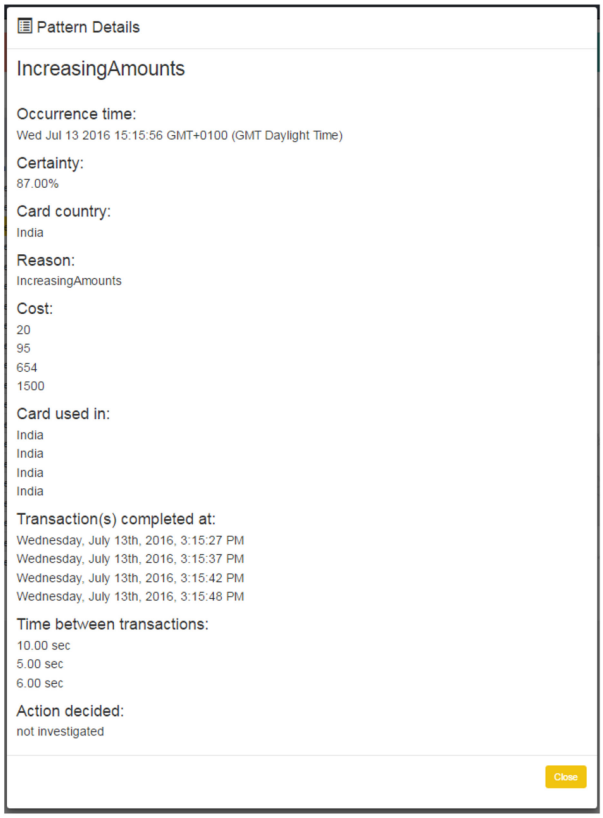


Fig. 3. List pop-up.

designs. Depending on the fraud pattern, one might assume that a given dashboard design would be more suitable, and, as the use of pop-up windows is only relevant to some of the frauds, efficient performance could be defined in terms of their use. This is shown in Table I.

The overview dashboard (see Fig. 4) presents the user with a summary bar, on the top of the dashboard, which shows the total number of transactions investigated by the department, the number of transactions flagged by the automation, the average amount (of cash spent on a transaction), and the average volume (of transactions) for a selected country. To select a geographical region, the user clicks on the map on the right of Fig. 4.

The analyst also has a list of the patterns flagged by the automation in the form of an event list (on the left of Fig. 4). The event list shows the transaction number, the fraud pattern identified, the automation confidence level associated with this flag, and the current status of the pattern (not investigated, fraud, contact, allow). The analyst would take the next "not investigated" fraud from the event list and work on this.

At the bottom of each dashboard there are five buttons: three of them are for the possible decisions the user can make: allow a transaction, query the transaction with the cardholder, block the transaction. On both dashboards, pop-ups are called by clicking the "explain" buttons: the one on the map brings up information related to the selected country (see Fig. 2), and the other, under the "events list," brings up more data related to the selected pattern (see Fig. 3).

TABLE I
NEED TO DRILL-DOWN FOR FURTHER INFORMATION TO INVESTIGATE A GIVEN FRAUD PATTERN WITH A GIVEN DASHBOARD

Fraud Pattern	Information required	Overview Dashboard	Detailed Dashboard
IA: increasing amounts	a) number of transactions b) amounts trend	List pop-up required List pop-up required	
LA: large amount	a) transaction cost b) country average amount	List pop-up required	Country pop-up required
FA: flash attack	a) number of transactions b) country average volume	List pop-up required	Country pop-up required
TF: transactions in faraway places	a) countries where transaction were made b) time distance between transactions	List pop-up required	

As Table I indicates, when using the overview dashboard, analysts either have the information they require for diagnosis of fraud, or they have to search for it in the list pop-up.

In the detailed dashboard (see Fig. 5), participants can use the information available on screen or have to bring up the country pop-up window to get contextual information. The window on the left labeled patterns to investigate is the same as the event list window in the overview dashboard. At the top of the pattern view window (see Fig. 5), visualizations show the number of transactions and the interval between them (left) and transaction amounts (right). Below these, a map shows the geographical region in which the transaction(s) took place, and, at the bottom of the window, the date and time of flagging is shown along with the automation confidence level and reason for flagging. It should be noted that while Figs. 4 and 5 show different event lists, participants encountered all of the possible events during their trials.

V. EXPERIMENT

We are interested in whether the participants' time to make a decision is affected by automation confidence. We are also interested in how participants use the information available and whether their information-search strategy (i.e., how many pop-ups participants open and how long they choose to keep them open) changes with different confidence levels and/or fraud types. These provide measures of whether users recognize the need for extra information, and the effort put into seeking additional information for the transaction. Thus, the experiment was designed to explore the following two questions.

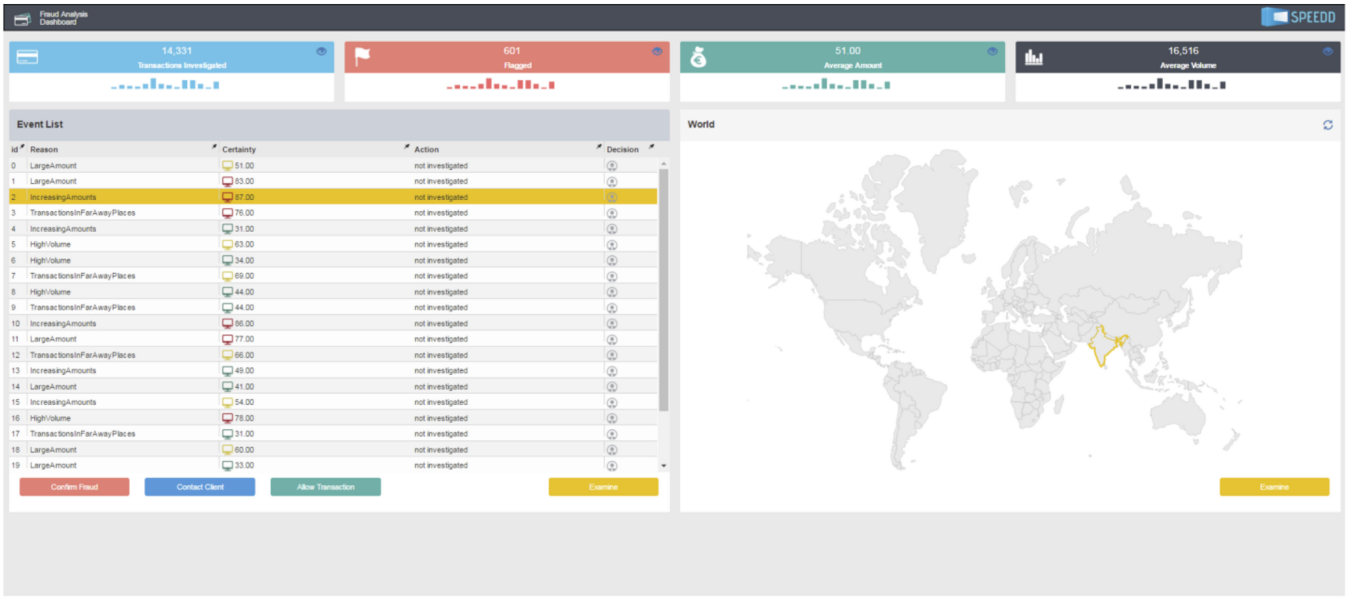


Fig. 4. Overview dashboard.

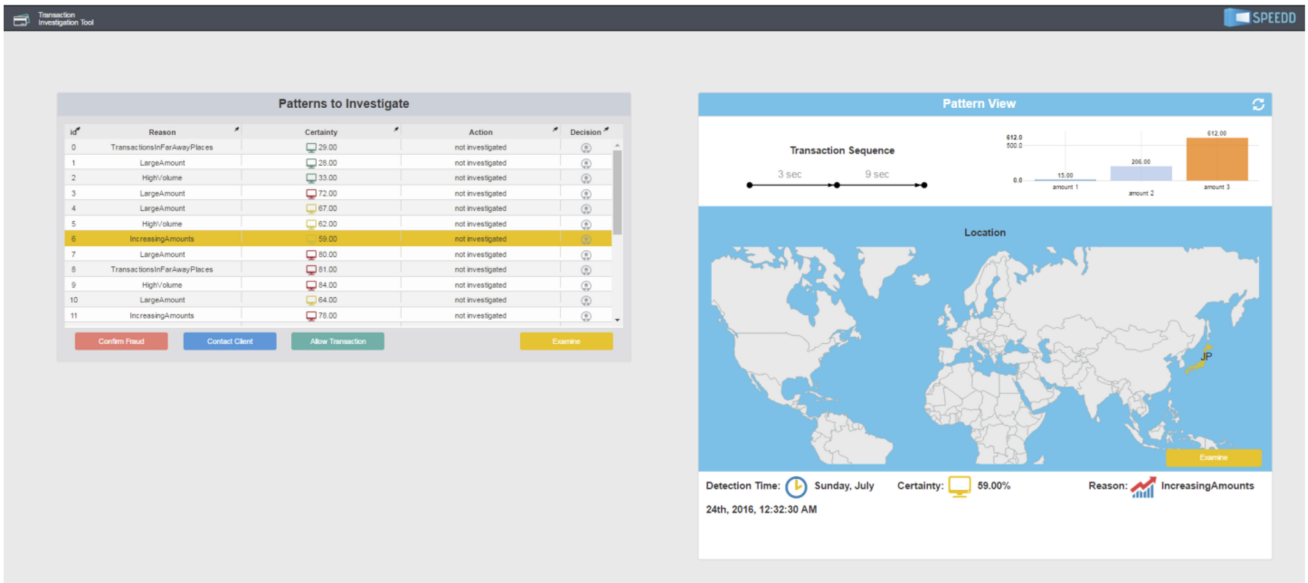


Fig. 5. Detailed dashboard.

1) Does user behavior vary with automation confidence?
2) Does information search-strategy differ when people are presented with information relevant to their decision activity?

We would consider optimal behavior to seek extra information only when it is required and to keep these pop-up windows opened until the needed information is extracted. We would also expect human decision making to reflect automation confidence levels, in line with previous work [2], [3], [19], [36], i.e., users will allow more transactions in the low confidence cases and flag more transactions as fraud in the high confidence cases, and make more decisions to contact customers in the medium confidence cases.

A. *Participants*

A total of 27 people took part in the experiment [17: male; 10: female; age range: 22–29]. None of the participants had experience of working in the credit card industry or financial sector. Given the observation that fraud schema are highly company-specific, we felt that it was inappropriate to recruit participants from a specific company because they would respond to the information according to their organizational policy. Having said this, the experiment has been repeated with a small number (four) of experienced credit fraud analysts, and performance is consistent with this study [1]. For our recruitment of participants, we assume that people educated to degree level and given training on

the fraud patterns to identify would provide a reasonable proxy with call-center agents, whose role is primarily to follow the script provided to them. Therefore, each participant was trained to criterion (see below) before the experiment began. In this way, we have a homogenous user group from which to explore the impact of dashboard designs on decision performance.

B. Procedure

The study was approved by the University of Birmingham Ethics Panel (Reference Number ERN_13-0997). All data were anonymized and participants provided informed consent.

Following a briefing on the task and training in using both dashboards, participants were given a demonstration of how fraud patterns could be recognized from the data presented in a dashboard and the other windows. Next, they were given four practice trials (to become accustomed to interacting with the dashboard) in order to familiarize themselves with the dashboard before beginning the trial. They were asked to process up to ten examples. Once participants were able to correctly process five examples consecutively the main experiment began. In order to replicate the script-based aspect of call-center analysts' work, participants were provided with an aide memoire, which defined the four fraud patterns.

Each participant investigated 48 patterns, 24 using each dashboard. The set of 24 patterns (for each dashboard) were randomized across participants in order to minimize order effects. The patterns were defined in terms of four fraud types—increasing amounts (IA); transactions in faraway places (TF); large amounts (LA); flash attack (FA)—and three levels of automation confidence—low ($\leq$51%); medium (51%–69%); high ($\geq$70%). The rationale for defining "high" confidence as >70% (rather than, say, 90% or 100%) was twofold. First, we assumed that if confidence is sufficiently high, then the decision will be made automatically, and thus only those decisions which fall below a threshold will be passed on to the human operator. Second, a reliability of 70% could be taken as a cutoff point, below which there is a drop in performance such that it is preferable to ignore unreliable automation [31]. Each fraud pattern was presented twice under each automation confidence level.

The independent variables for the experiment were: dashboard, fraud pattern, and automation confidence level. The dependent variables were: time to submit a decision, number of pops-up opened, and decision type. All statistical tests were performed using IBM SPSS v24, and tests of normality were applied to the data in order to select statistical tests [5].

C. Results

1) Average Time to Make Decisions: Tests of normality (using Shapiro–Wilk) indicated that the average time to make a decision did not follow a normal distribution in the low and medium-confidence conditions. Thus, nonparametric tests were applied to the time to make decision data. There was a significant difference between the high confidence and the medium-confidence conditions ($z = -2.407, p = 0.016$) (median high = 11.73 s, median med = 13.06 s) for time to make decision. No other differences were found (see Fig. 6).

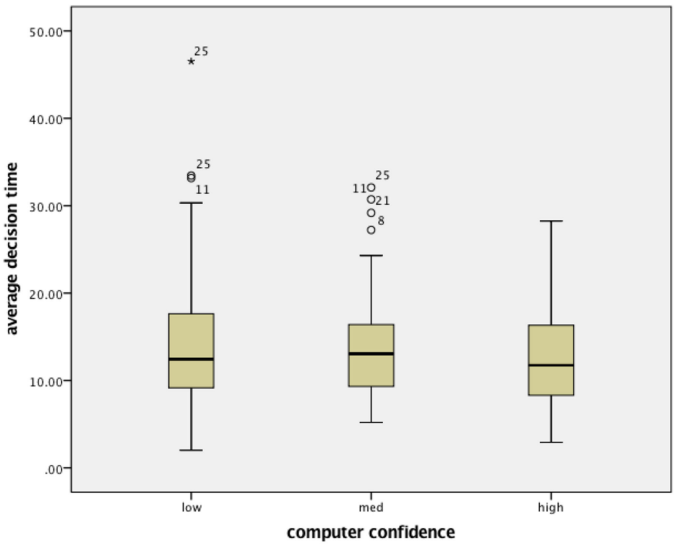


Fig. 6. Average time to make decisions, under different automation confidence levels with the two dashboards. Outliers are indicated as dots.

2) Type of Decision Made by User: The types of decision data [allow transaction, block transaction (i.e., fraud) or contact cardholder] were normally distributed except for fraud in the low-confidence condition and allow in the high-confidence condition. In cases, where normally distributed datasets were compared, parametric tests were applied otherwise nonparametric tests were applied. When automation confidence level was low, there were significantly more decisions to allow than block the transaction ($z = -3.69$, $p < 0.001$) (median allow low = 56%, median fraud low = 13%). There were also more decisions to contact the customer in the low condition than to flag the transaction as fraud (i.e., block it) ($z = -3.546$, $p < 0.001$) (median contact low = 31%, median fraud low = 13%).

In the medium computer confidence cases, the number of decisions to contact the customer was significantly higher than to allow transactions ($t(24) = 2.923$, $p = 0.007$) (mean contact med = 42.8%, mean allow med = 24.12%).

When automation confidence level was high, participants were significantly more likely to mark the transaction as fraud than to either allow ($z = -3.919$, $p < 0.001$) (median fraud high = 56.25%, median allow high = 0%) it or contact the customer ($t(24) = 3.131$, $p = 0.005$) (mean fraud high = 59.5%, mean contact high = 32.25%). Moreover, there were more decisions to contact the customer than to allow transactions ($z = -3.546$, $p < 0.001$) (median contact high = 31.25%, median allow high = 0%).

Furthermore, the proportion of allow decisions in the low-confidence condition was higher than in the medium-confidence condition ($t(24) = 5.85, p < 0.001$) (mean allow low = 52.36%, mean allow med = 24.12%), which was higher than in the high-confidence condition ($z = -3.472$, $p = 0.001$) (median allow med = 25%, median allow high = 0%). Fraud decisions are higher in high-confidence condition than the medium ($t(24)$

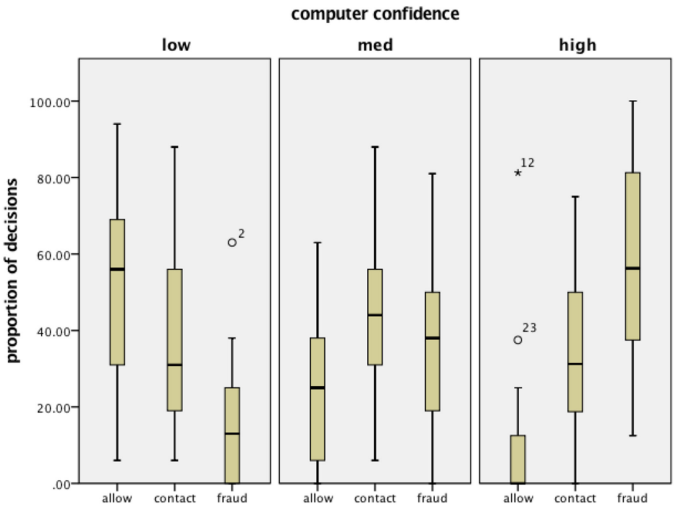


Fig. 7. Types of decision made by the user, for both dashboards, under different automation confidence levels.

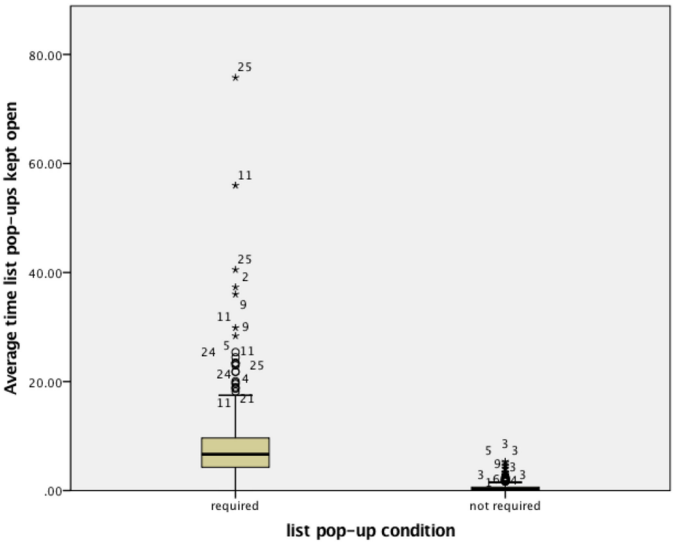


Fig. 8. Average time country pop-ups were kept open. Outliers are indicated as dots.

$= 4.76$, $p < 0.001$) (mean fraud high $= 59.5\%$, mean fraud med $= 35.32\%$) and from the medium to the low-confidence condition ($z = -3.415$, $p = 0.001$) (median fraud med $= 38\%$, median fraud low $= 13\%$). There were no differences in contact decisions between confidence levels. These can be seen in Fig. 7.

3) Pop-Ups Opened: Normality tests showed that data for the number of pop-ups opened were not normally distributed.

A Wilcoxon test showed significant difference in pop-ups opened between the required and not required conditions of the list pop-up ($z = -7.28$, $p < 0.001$). However, no differences were found for the country pop-up (Fig. 8). Participants opened fewer list pop-ups (Fig. 9) when not required at low ($z = -3.198$, $p < 0.001$), medium ($z = -3.642$, $p < 0.001$) and high ($z = -5.643$, $p < 0.001$) computer confidence (compared with required pop-ups). No other differences were found.

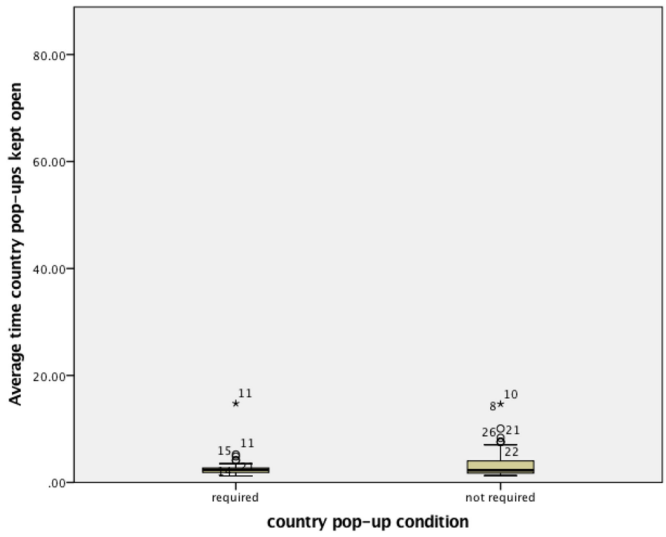


Fig. 9. Average time list pop-ups were kept open. Outliers are indicated as dots.

4) Time Pop-Ups Active: The Shapiro–Wilk test showed that data for the time pop-ups were kept active were not normally distributed. A Wilcoxon test showed significant differences between the required and not required conditions for list pop-up and country pop-up. Participants looked at the list pop-up for longer in the required condition compared to the not required condition ($z = -8.504$, $p < 0.001$) (median list required $=$ 6.26 s, median list not required $= 3.65$ s). However, participants looked at the country pop-up for longer when it was *not* required ($z = -2.306$, $p = 0.021$) (median country required $=$ median country not required $= 0$ s). The list pop-up was kept open for longer than the country pop-up both when required ($z = -10.687$, $p < 0.001$) (median list required $= 6.26$ s, median country required $= 0$ s) and when not required ($z = -12.608$, $p < 0.001$) (median list not required $= 3.65$ s, median country not required $= 0$ s).

VI. Discussion

The results are considered in terms of the questions posed in Section V.

1) Does user behavior vary with automation confidence level?

Participants responded faster to transactions associated with a high automation confidence level compared to the medium confidence transactions (although there was no significant difference between low confidence and either medium or high). Automation confidence level had a bearing on the type of decision that participants made: the greater the automation confidence, the more likely participants were to block the transaction, and the lower the automation confidence the more likely participants were to allow the transaction. The decision to contact the cardholder depended on the automation reliability. When reliability was low, participants were more likely to contact the cardholder than to declare a transaction fraudulent, and in the high-reliability condition, they were more likely to contact the cardholder than to allow the transaction. We suggest that

this points to two reasonable assumptions that the participants could be making: if the computer has made an error, it might not have useful information available to it, and speaking to the cardholder would involve a sort of unstructured query (which can be refined during the conversation, albeit on the basis of a script). In other words, when automation confidence is low, then it makes sense to seek information outside the computer before deciding to declare a transaction as fraudulent, but if the computer confidence is high (but you are not sure), then it makes sense to speak to the cardholder before allowing a transaction. In the medium-confidence conditions, then speaking to the cardholder was the most commonly selected option and this makes sense because it reserves judgment on the available options.

2) Does information search-strategy differ when people are presented with information relevant to their decision activity?

In all conditions, participants used at least one pop-up—even when this was not required. When using the "list" pop-up, participants tended to make more use of the pop-up windows when they were required, and to keep these open when they were required, than when they were not required. This was not so apparent for the country pop-ups, and participants kept the country pop-up open for longer when it was *not* required. This suggests that the relevance of the information in these pop-ups differed in terms of their content. This raises some interesting questions concerning the issue of "transparency." While the dashboards in this experiment had been designed to provide necessary information to support the recommendation for each specific fraud type, it is possible that participants did not regard this as sufficient—and so, would seek more information. This could relate to the suggestion that people seek more information to increase their confidence in a decision [12], [23], even when the additional information might not be useful. From this, the act of seeking additional information might not relate directly to the decision making activity so much as their emotional response to making a decision. It also implies that estimates of the value, or relevance, of information might differ from the actual value (in terms of a specific decision). It is apparent, in our experiment, that participants were able to make reasonable judgments about the value of information in the "list" pop-ups, but were less able to make such judgments for the map pop-up. For the list pop-ups, analysis shows that even though extra information is brought up, it is rapidly discounted when it is not required. When the information present in the list pop-up is required to make a decision, participants spend significantly longer looking at it than when it is not required. This effect does not occur in the case of the country pop-up, but this may be due to the fact that there was very little additional information in this window compared to the list pop-up and the users and so participants may have responded to them differently (see Figs. 2 and 3). It might be the case that the very fact that the country pop-up contained information that was not relevant could have evoked a longer decision time (to ensure that the information was either redundant or not relevant). Alternatively, it might be the case that what looks like a search for additional information is actually a means of bounding the information space (and determining what information to ignore). In support of this suggestion, it is worth noting that the fraud patterns we defined were not mutually exclusive (i.e., information can fit more than one pattern), and so the task of aligning information with a pattern is somewhat fuzzy. In response to this fuzziness, a reasonable response might be to rule out alternative explanations, and this could involve checking the other information that is available. Certainly, in our postexperiment debriefs, this strategy was mentioned by several participants. Thus, presenting information in an easily accessible dashboard is only part of the challenge of supporting decision making and people will respond to recommendation in terms of the transparency between their understanding of the task and the information available.

VII. Conclusion

We report the design of dashboards for analyzing four types of credit card fraud under different levels of computer confidence. The dashboards differed in the type of information presented to the user. Based on the information they presented and the type of fraud being investigated, efficient pop-up use was defined. Results show that participants would drill-down for information even when this is not required. However, at least in the case of the list pop-up, participants quickly discount the window when it was not required.

Relating these findings to the notion of conformance [35], it can be seen that participants adapted their decision to align with the confidence of the computer and that their use of available information was, to some extent, dependent on the type of fraud that was being flagged. In [35, p. 50], there was the suggestion that "conformance may only be relevant for expert users who hold consistent and well-developed decision-making strategies." Our findings suggest that training participants to criterion, providing them with an aide memoire to define the task, and providing clearly defined dashboards all help to ensure consistency in decision making, and this, in turn, relates to strategic conformance.

From this experiment, we propose that automation confidence interacts with user decision activity—not just in terms of decision outcome but also in terms of information search. However, it is not a simple matter of searching for more information when automation confidence is low. Such a strategy could be counterproductive for the user because it would lead to an increase in workload. Recently, we have reported a model of the optimal use of information sources by healthcare professionals that demonstrates that the strategies for information-search strategy and decision making emerge from the reliability of the information sources, the relative usefulness of these sources, and cost of accessing these sources [2]. From this model and the results reported in this paper, we propose that the ecology of the decision environment (in terms of the availability and access cost of information, and the confidence of support provided to the decision maker) creates a tradeoff space for the analyst. Consequently, rather than searching for more information that the computer might hold but not have used, the user would seek

information from sources outside the computer. The design of a dashboard could, therefore, not only focus on the behavior of the automation (in terms of indicating its recommendation and confidence in that recommendation) but also the availability of information that could be relevant to that recommendation.

As a final point, we note that there continues to be rather limited research into the design and use of dashboards (even though these are growing in popularity). Resnick proposed, in 2003, that there is a need to understand the ways in which dashboards (as high-level summaries of data) should allow users to understand variability in the source data, to ensure that users do not erroneously see patterns in data and should direct users to additional, relevant data [27]. In this paper, we contribute to these aims through the exploration of the ways in which information can relate to different decision tasks, and how user performance can be influenced by "source variability" in the form of reliability of automation recommendations. Further work could explore the decision processes that people make in response to dashboards, particularly when the dashboard seems to be designed to support rapid, intuitive decision making [11], [27].

REFERENCES

[1] A. Artikis *et al.*, "A prototype for credit card fraud management," in *Proc. 11th ACM Int. Conf. Distrib. Event-Based Syst.*, ACM, 2017, pp. 249–260.

[2] A. Atcharya, A. Howes, C. Baber, and T. Marshall, "Automation reliability and decision strategy: A sequential decision model for automation interaction," in *Proc. Annu. Meeting Human Factors Ergonom. Soc.*, HFES, 2018, pp. 144–148.

[3] C. Baber, N. S. Morar, and F. Mc.Cabe, "Ecological interface design, the proximity principle, and automation reliability in road traffic management," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 3, pp. 241–249, Jun. 2019.

[4] J. E. Bahner, A.-D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Human-Comput. Stud.*, vol. 66, pp. 688–699, 2008.

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decis. Support Syst.*, vol. 50, pp. 602–613, 2011.

[6] X. Chen, S. D. Starke, C. Baber, and A. Howes, "A cognitive model of how people make decisions through interaction with visual displays," in *Proc. 32nd Annu. ACM Conf. Human Factors Comput. Syst.*, ACM, 2017.

[7] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Mahwah, NJ, USA: Lawrence Erlbaum, 1988.

[8] W. N. Dilla and R. L. Rascke, "Data visualization for fraud detection: Practice implications and a call for future research," *Int. J. Accounting Inf. Syst.*, vol. 16, pp. 1–22, 2015.

[9] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," *Expert Syst. Appl.*, vol. 38, pp. 13057–13063, 2011.

[10] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human-Comput. Stud.*, vol. 58, pp. 697–718, 2003.

[11] S. Few, *Data Visualization Past, Present and Future*. Ontario, Canada: Cognos Innovation Center for Performance Management, 2007.

[12] C. C. Hall, L. Ariss, and A. Todorov, "The illusion of knowledge: When more information reduces accuracy and increases confidence," *Org. Behav. Human Decis. Process.*, vol. 103, pp. 277–290, 2007.

[13] D. J. Hand and G. Blunt, "Prospecting for gems in credit card data," *IMA J. Manage. Math.*, vol. 12, pp. 173–200, 2001

[14] N. Katzouris, A. Artikis, and G. Paliouras, "Online learning of event definitions," *Theory Pract. Logic Program.*, vol. 16, pp. 817–833, 2016.

[15] D. A. Keim, "Information visualization and visual data mining," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, pp. 1–8, 2002.

[16] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, pp. 153–184, 1994.

[17] K. J. Leonard, "Detecting credit card fraud using expert systems," *Comput. Ind. Eng.*, vol. 25, pp. 103–106, 1993.

[18] J. Meyer, L. Feinshreiber, and Y. Parmet, "Levels of automation in a simulated failure detection task," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2003, pp. 2101–2106.

[19] N. Morar and C. Baber, "Joint Human-Automation decision making in road traffic management," in *Proc. Annu. Meeting Human Factors Ergonom. Soc.*, HFES, 2017, pp. 385–389.

[20] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation bias: Decision making and performance in high-tech cockpits," *Int. J. Aviation Psychol.*, vol. 8, pp. 47–63, 1998.

[21] D. A. Norman, "The 'problem' with automation: Inappropriate feedback and interaction, not 'over-automation'," *Philos. Trans. Royal Soc. London B Biol. Sci.*, vol. 327, no. 1241, pp. 585–593, 1990.

[22] T. C. Ormerod, L. J. Ball, and N. J. Morley, "Informing the development of a fraud prevention toolset through a situated analysis of fraud investigation expertise," *Behav. Inf. Technol.*, vol. 31, no. 4, pp. 371–381, 2012

[23] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, pp. 381–410, 2010.

[24] R. Parasuraman and V. Riley, "Humans and Automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, pp. 230–253, 1997.

[25] P. Pirolli and S. Card, "Information foraging," *Psychol. Rev.*, vol. 106, no. 4, 1999, Art. no. 643.

[26] P. Pu and L. Chen, "Trust building with explanation interfaces, in *Proc. 11th Int. Conf. Intell. User Interfaces*, 2006, pp. 93–100.

[27] M. L. Resnick, "Building the executive dashboard," in *Proc. 47th Annu. Meeting Human Factors Ergonom. Soc.*, HFES, 2003, pp. 1639–1643.

[28] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. New York, NY, USA: Springer, 2011, pp. 1–35.

[29] J. T. S. Quah and M. Sriganesh, "Real-time credit card fraud detection using computational intelligence," *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1721–1732, 2008.

[30] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Syst. Appl.*, vol. 40, pp. 5916–5923, 2013.

[31] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Vis. Lang.*, 1996, pp. 336-343.

[32] I. L. Singh, R. Molloy, and R. Parasuraman, "Automation-induced monitoring inefficiency: Role of display location," *Int. J. Human-Comput. Stud.*, vol. 46, pp. 17–30, 1997.

[33] L. J. Skitka, K. Mosier, and M. D. Burdick, "Accountability and automation bias," *Int. J. Human-Comput. Stud.*, vol. 52, pp. 701–717, 2000.

[34] N. Tintarev and J. Masthoff, "Evaluating the effectiveness of explanations for recommender systems," *User Model. User-Adapted Interact.*, vol. 22, pp. 399–439, 2012.

[35] C. Westin, C. Borst, and B. Hilburn, "Strategic conformance: Overcoming acceptance issues of decision aiding automation?," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 1, pp. 41–52, Feb. 2016.

[36] C. D. Wickens and S. R. Dixon, "The benefits of imperfect diagnostic automation: A synthesis of the literature," *Theor. Issues Ergonom. Sci.*, vol. 8, pp. 201–212, 2007.

An Interface for Verification and Validation of Unmanned Systems Mission Planning: Communicating Mission Objectives and Constraints

Clayton D. Rothwell and Michael J. Patzek

***Abstract*—Anticipated advances in the use and capability of unmanned systems may increase the complexity in mission planning and human-machine teaming, creating challenges for mission performance, safety, and predictability. Verification and validation tools, such as model checking, have been applied to mission planning and is a promising approach but require expertise to use. This article describes the study of one such tool, with an interface that provides the human operator a way to communicate their high-level mission goals and objectives to the model checking software without having to learn temporal logics or syntax. We describe the development and refinement of a prototype, as well as an experiment testing if a verification and validation tool can improve mission planning compared to a baseline without a verification and validation tool. Results showed that a tool increased mission planning accuracy, while also increasing mission planning time. Also, subjective workload was not different between the two configurations (tool & baseline). These results indicated that verification and validation tools with a suitable interface can enable humans to communicate mission objectives and constraints to machines, improving performance, safety, and predictability. Results also suggested that additional research to improve this communication would further increase the benefits and impact of these kinds of tools.**

***Index Terms*—Human computer interaction, model checking, planning, user interfaces, unmanned vehicles.**

I. Introduction

EFFECTIVENESS, safety, and operator workload are key concerns of unmanned aerial vehicle (UAV) planning and execution. Current advanced technology development efforts, such as single operator control of multiple UAVs, human-machine teaming, increased control autonomy and decentralization, are likely to increase unexpected execution behaviors and many have studied interface techniques to address these concerns [1]–[3]. In addition, these technology advances will challenge the typical, manually intensive quality assurance and validation processes for ensuring effectiveness and safety of plans, which are time consuming, not exhaustive, and prone to error. Operators need automated verification tools that can improve the quality and speed of validation, but, similar to execution, require associated interface improvements to be adopted and useful. The particular issue with verification and validation tools centers around communication. It is difficult to convey the desired behaviors, goals, objectives, constraints, rules of engagement, etc., in such a way that a machine can reason over them [4].

The current study developed an interface to address these communication issues, the Specification Pattern Editor and Checker (SPEC), and tested its usability and impact on mission planning effectiveness within a tool for verification and validation, the Verifiable Task Assignment and Scheduling Controller (VTASC) [5]. VTASC is a collection of software components that features the SPEC interface to facilitate human-machine communication, a model checker for verification, and a UAV ground control station for mission planning. The SPEC interface supported communication by providing a set of tools to write specifications in English that are converted to the temporal logic representation for the model checker. Ideally, humans could communicate specifications and synthesis techniques could generate control automation and mission plans for all the UAVs that meet the specifications. Toward that end, the current work had human operators generate specifications using SPEC, as well as UAV mission plans, then VTASC checked if the plans met the mission objectives while complying with the rules of engagement and constraints.

The article is organized as follows: Section II discusses related work. Section III provides an overview of VTASC and the design of and interaction with SPEC. Section IV describes a usability test conducted on version 1 of SPEC. Section V describes version 2 of SPEC and the changes that were made based on the usability study's findings. Section VI presents an experiment comparing mission planning performance with and without a verification tool. With a verification and validation tool integrated into the control station, we expected that operators would perform complex mission planning for multiple UAVs more effectively than in baseline conditions without a verification tool.

II. Background

Verification and validation tools are a promising option that simultaneously address the challenges created by increasing mission and UAV control complexity and the need for

Manuscript received November 5, 2018; revised April 3, 2019 and July 30, 2019; accepted September 18, 2019. Date of publication November 11, 2019; date of current version November 21, 2019. This work was supported in part by Air Force Office of Scientific Research (13RQ03COR), and in part by Air Force Research Laboratory FA8650-14-D-6500/0002. This article was recommended by Associate Editor M. L. Bolton. *(Corresponding author: Clayton D. Rothwell.)*

C. D. Rothwell is with the Ohio State University, OH 43210 USA (e-mail: clayton.rothwell@osumc.edu).

M. J. Patzek is with the Air Force Research Laboratory, Wright-Patterson AFB, OH 45433 USA (e-mail: michael.patzek@us.af.mil).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2945618

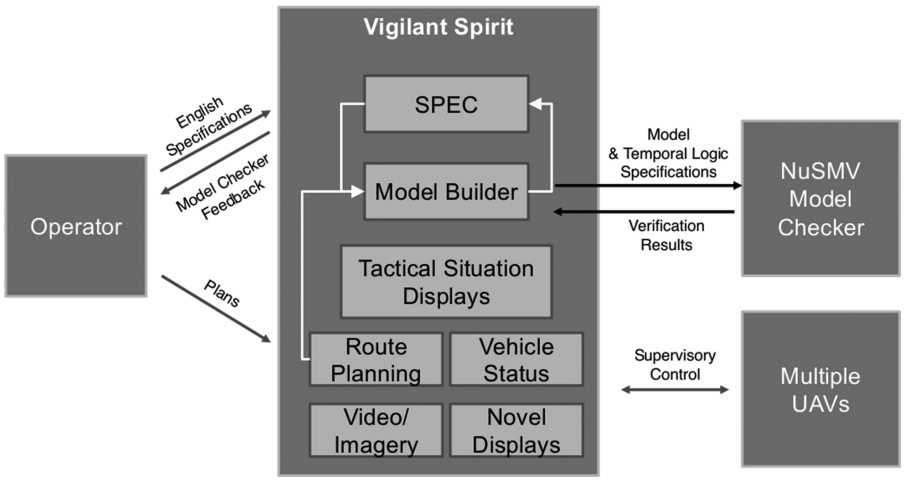


Fig. 1. Diagram of VTASC components that shows the connections between the operator (left), the UAV control station (middle), and the model checking software (top right). See text for more details.

human-machine communication. Researchers have applied formal methods—mathematically based languages, techniques, and tools used to design and verify the safe and reliable operation of systems—to robotic systems, including UAVs [6]–[8]. Formal methods include model checking, a technique in which a finite-state representation of a system is automatically checked against a set of desired specifications expressed in temporal logic. In the UAV domain, model checking can be used to verify that UAV assignments, routes, and on-board dynamic mission planning conform with mission specifications (e.g., objectives, constraints, and rules of engagement) during mission planning and during mission execution, thereby increasing the chances of mission success. However, the impact of model checking on mission planning accuracy and completion time has not been established.

Model checking of mission specifications provides a framework for communication between human and autonomous system members. The mission specifications would communicate the human's intent to the model checker that would check if the mission plans accomplish the intent. Communication is a vital element for successful collaboration/teaming that has been central to research on human teams, but is similarly important for collaboration between humans and machines [9]. However, formal methods and temporal logic specifications can be difficult to understand, use, or learn [4]. Therefore, an interface is needed that operators find easy to understand and can capture their intended meaning and expresses it in a formal methods framework. SPEC was designed to accomplish these goals, and features English structured language and everyday examples to improve pattern understanding.

III. Description of VTASC and SPEC

A diagram of the VTASC software is shown in Fig. 1. The human operator (left) provided English specifications to Vigilant Spirit Control Station (VSCS) (middle), a UAV control station that has been developed by the Air Force Research Laboratory (AFRL) for studying multi-UAV supervisory control and new interfaces and technologies [10], [11]. The English specifications were input to SPEC, which resided inside VSCS, and SPEC converted the specifications to temporal logic (SPEC is detailed in Section III-A). The operator also created UAV plans using VSCS's interface for waypoint-based route planning, in which each point had an associated speed and altitude. Other functions of VSCS such as displays of video/imagery were not used in this research.

A model builder component inside VSCS estimated UAV positions for each second of the mission using the planned paths and a simple flight model (i.e., instantaneous turns and linear climb/descent). The initial model builder is detailed in [8]. It assumed UAVs were constrained to fly over roads or point-to-point between nodes on the map. Here, UAV waypoints had no such restrictions. The final model was a deterministic finite-state model of the mission plans representing UAV position for each second of the mission and Booleans for UAVs present over any point of interest. This simple modeling approach facilitated the examination of the interface for mission planning, rather than attempting a valid representation of the uncertainties UAVs face in plan execution. The model builder combined this with temporal logic specifications from SPEC, and passed the combination to New Symbolic Model Verifier (NuSMV), a symbolic model checker [12], [13]. NuSMV was selected to accommodate both relative time and real-time specifications. NuSMV performed its model checking, and provided the results for each specification back to VSCS's model builder. The model builder fed the results to SPEC which displayed them to the operator.

A screen shot of the VTASC interface is shown in Fig. 2. This was displayed on two side-by-side monitors, with the left monitor containing the VSCS display that contained vehicle status information and route planning interfaces, and the right monitor containing SPEC. The following sections go into additional detail on SPEC's design and how SPEC provided a framework for communication between humans and machines.

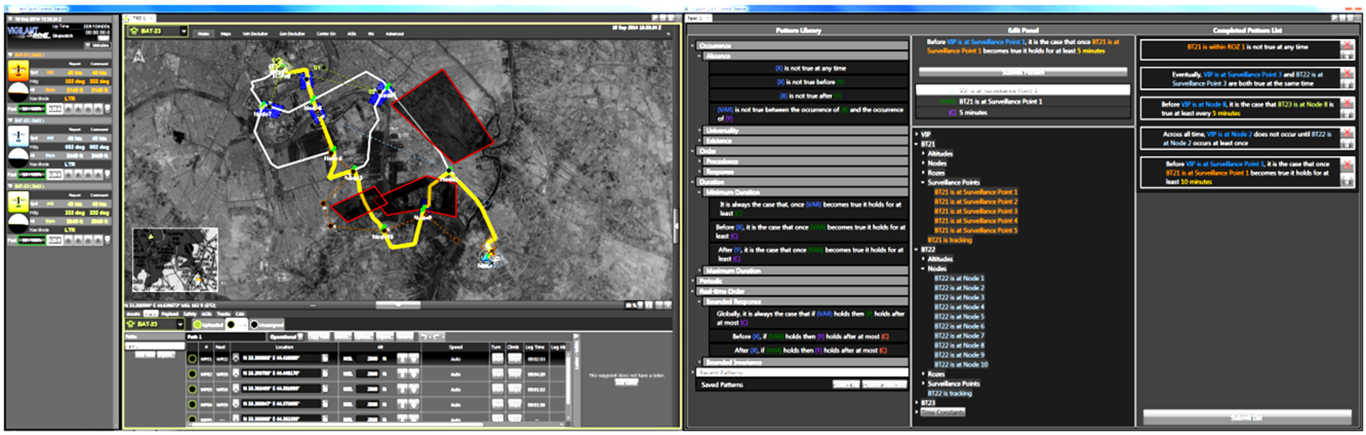


Fig. 2. Screenshot of VTASC which shows the side-by-side displays. The VSCS display was used for waypoint-based route planning (left) and the SPEC interface was used for writing specifications and viewing model checker feedback (right).

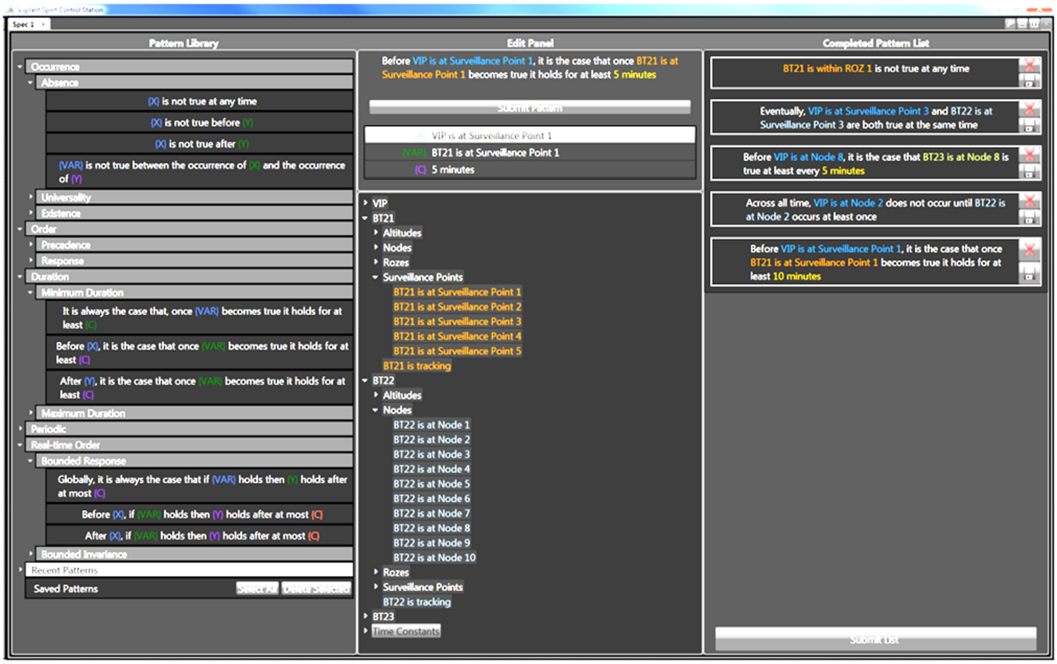


Fig. 3. Screenshot of SPEC, providing a detailed view of the Pattern Library (left), the Edit Panel (middle), and the Completed Pattern List (right).

A. SPEC Interface

The main functions of SPEC are:

1) allow human operators to write specifications in English form rather than learning temporal logic syntax;
2) initiate the model checking processes;
3) display the results of model checking.

Regarding specification writing, some prior work has used translation-based approaches [14], [15]. SPEC used a different approach, based on the notion of a specification pattern which has been described in [4] and used in [16]. A specification pattern captures a temporal relationship that expresses a commonly-desired system property with an expert-defined mapping between an English representation and the temporal logic syntax. Temporal relationships, such as *property P always happens at least once every t minutes*, are generic in nature such that *P* can be any property of a given system and *t* can be any one of the discrete time steps of that system. Patterns can be instantiated with variables specific to a particular system of interest and then used in model-checking. The majority of SPEC is devoted to providing operators with a library of specification patterns, and methods to fill in the generic patterns with variables of interest.

An expanded view of SPEC is shown in Fig. 3 to provide detail of the three panels. The leftmost panel is the *Pattern Library*, which is dedicated to displaying an organized collection of specification patterns that are available. The middle panel is the *Pattern Editor*, which is dedicated to filling in patterns with variables that are modeled in the finite-state model of the mission plans. The rightmost panel is the *Completed Pattern List*, which holds the patterns that have been written and that will be submitted to the model checker. This panel also displays feedback for each specification after the model checker has run.

1) Pattern Library: The Pattern Library is an organized and interactive tree view that contains all the patterns available for use (Fig. 3, left panel). Patterns are categorized as: occurrence, order, periodic, duration, and real-time order. These categories came from two extensive surveys of the specification literature and model checking applications [4], [17]. Occurrence patterns are relationships that capture if something should always occur, never occur, or occur sometime. Order patterns are relationships that capture if something should occur before or after something else. Periodic patterns are relationships that capture if something should repeatedly occur within some time window. Duration patterns are relationships that capture if something should occur for more or less than a certain length of time. Real-time order patterns are similar to order patterns, which use before and after, but add the ability to specify how much time before or after.

Most patterns came from the pattern literature but a few patterns were customized for this domain (e.g., to create the comparisons: equal to, greater than, less than). The English language representation of each pattern was taken from [17] then adapted to be easier to read and understand. Our process for adapting the English representation of patterns was: be as consistent as possible in the use of variables and words, use "happens" instead of "is true" or "occurs," and simplify understanding by rewording patterns to avoid negatives and double negatives wherever possible.

The patterns listed in the library (42 total) are configurable for expansion and updating, though patterns in the configuration file must have temporal logic and a corresponding English representation written by a human. The Pattern Library has two other categories that are dynamic: recent patterns and saved patterns. Recent patterns holds the ten most recently completed patterns. Saved patterns can be unlimited, are stored in a file from session to session and can be added to hold frequently used patterns from mission to mission.

2) Edit Panel: Upon selection of a pattern in the library, the pattern populated the top portion of the Edit Panel and replaced any existing contents (Fig. 3, middle panel). Immediately beneath that is a list of the variables contained in the pattern that need to be assigned to complete the specification. The first variable was selected by default, and the selected variable was changed by clicking on another variable's row. Variables in the patterns could be assigned a value from the interactive tree which lists what is in the finite-state model. Options included: the UAV's altitude, the UAV's location above a named intersection of roads, the UAV's location above possible bottlenecks for the convoy, and the UAV's location within restricted operating zones (ROZ)s (i.e., regions of airspace that are "no fly" areas for these UAVs). Unlike the patterns, the variables in the model cannot be configured. (Updates to the model require changes to the model builder software and possibly the flight model).

Once all variables were assigned, the completed specification was submitted to the Completed Pattern List. There were no checks that the completed specification was meaningful or appropriate, but there were some simple error messages that appeared when the pattern did not have all the variable fields assigned or when it was a duplicate.

3) Completed Pattern List: The Completed Patterns List holds the instantiated patterns, now considered mission specifications (Fig. 3, right panel). Each completed pattern can be edited, saved, or deleted. At the bottom of the panel is a button to submit the list of specifications to the model checker. Depending on the complexity of the check, the model checker could run for a number of seconds. Ellipsis symbols to the left of each specification told the user that the model checker was processing, and were replaced by a green checkmark or red prohibition symbol when the check was finished to indicate the positive or negative result, respectively. Each specification was checked independently in the NuSMV software, instead of concatenating them all together, which allowed the interface to provide individual feedback on each specification. Also, the specification checks were not run together but were broken into two groups: one for relative time specifications (e.g., LTL) and the other for real-time specifications (e.g, RTCTL). These checks were executed by separate instances of NuSMV that were run in parallel. In some instances, arrival of results from the two checks did not coincide.

B. Example Use Case

To perform a safety check that a UAV avoids a specific ROZ, first the user would browse the Pattern Library to find and select the "never occur" pattern. Then, using the Edit Panel, the user would browse to and select UAV-1 in ROZ-1, creating the pattern "UAV-1 in ROZ-1 never happens," and add this to the Completed Patterns List. The user would click to begin model checking and review the results.

IV. Usability Test

We conducted a usability test of the SPEC tool to receive preliminary feedback on the design and to identify areas for improvement prior to an experimental evaluation of SPEC. The test was within the context of a mission planning use case and also informed the scenario difficulty and duration for our eventual experiment. We used a combination of questionnaire and interview techniques. The test goals were: solicit feedback on the pattern concept, the method of creating specifications with the SPEC tool, and the model checker's output.

A. Method

Six participants were recruited and none had prior experience with SPEC or VTASC.

1) Materials: Testing was conducted in a quiet office setting using desktop computers with two side-by-side 24 monitors (Dell 2408WFPb; Round Rock, TX). VTASC was presented across the two monitors (Fig. 2) and participants used a keyboard and mouse. The questionnaire was adapted from the questionnaire for user interface satisfaction (QUIS) [18], a widely used usability measure. QUIS was adapted by adding extra questions about the sequencing of selecting, editing, and submitting patterns in SPEC (no questions were removed). For clarity, we refer

to it as QUIS-a. Participants were instructed to skip questions they thought were not applicable.

A semistructured interview was conducted after the participants finished the questionnaire, soliciting comments on these topics: the desired workflow for creating plans and specifications, if the English representation of patterns were easy to understand, design ideas for organizing the Pattern Library, design ideas for integrating the SPEC and the map, design ideas for the Pattern Editor, other types of specifications/patterns that they think would be useful.

The mission scenario was in the domain of convoy escort and had a combination of spatial and temporal constraints. The participant was tasked with planning routes for three UAVs that conformed with mission objectives. A convoy of vehicles containing a very important person (VIP) was beginning its route at Node 1 and end at Node 8. The primary convoy route and alternate contingency routes were displayed. Critical segments of the route were the intersections between the primary and alternate routes and the bridges. The intersections were critical because they were decision points for continuing on the primary route or switching to an alternate route. The bridges were critical because they were known points of vulnerability and high exposure. Bridges were called surveillance points (SPs) and had to be inspected from an altitude below 2400 ft mean sea level, due to the simulated sensor capabilities on these aircraft. The altitude space was banded in 200 ft bands, starting at 2000 ft and ending to at 3599 ft, and some UAVs had to be in separate bands during the whole mission. In addition, some current operations and known threats created ROZs.

Mission Objectives for the Usability test:

1) A UAV arrives at the destination (Node 8) before the VIP;
2) A UAV arrives at the first route decision point (Node 2) before the VIP;
3) A UAV visits the destination to take a picture every 5 min until the VIP arrives;
4) A UAV takes a picture of SP1 every 3 min until VIP goes through SP1;
5) A UAV takes a picture of SP3 every 3 min after VIP goes through SP1;
6) All UAVs avoid all ROZ;
7) UAV 2's altitude should always be different than UAV 3's altitude;
8) A UAV is located at SP2 to provide overwatch when VIP goes through.

2) Procedure: First, participants were introduced to the project and the goals of the test. Second, participants received training on VSCS, on how to create flight plans in VSCS, and on SPEC. The training used demonstrations and hands-on exercises with at least one pattern from each category in the Pattern Library to ensure participants were familiarized with the technology prior to the usability test. Training length varied with prior knowledge of VSCS, but on average was 1.25 h long. After training, the participants learned about the mission scenario they would perform for the usability test. The scenario was terminated by either: 1) correctly creating all eight specifications and building flight plans that met the specifications or 2) after 30 min elapsed, whichever came first. After the end of the scenario, participants filled out an usability questionnaire followed by a debriefing session that contained a semistructured interview. The purpose of the debriefing session was to capture free response comments from participants as well as clarify any written comments on the questionnaire (e.g., illegible handwriting, ambiguous statements, unfamiliar terms). On average, the whole procedure took 2.25 h to complete.

B. Outcomes

Mission planning performance and specification writing accuracy were characterized at a global level and were not analyzed in detail for this test. Of the six people that participated in the usability test, there was a wide range of progress on the task in the 30 min time limit. Two participants completed the mission before 30 min elapsed, two participants nearly completed the mission within 30 min, one participant was not quite halfway finished, and one did not attempt to use the model checker. The participant that did not use the model checker received the same training as the other participants and seemed to understand its purpose during training, but perhaps he would have benefited from additional training and familiarization time. This participant's data was included in the analysis because of his involvement with the model checker during training, and the findings did not change when his responses were removed from the analysis.

The questionnaire data were analyzed using the Kolmogorov–Smirnov nonparametric test that indicated when the distribution of responses is skewed with the majority of responses being above the midpoint of the scale. In general, ratings were higher than the midpoint; the overall mean was 6.53 (SD = 2.86). Fig. 4 shows mean ratings for each question with standard error bars. Despite the small sample size, the Kolmogorov–Smirnov test found responses for 5 QUIS-a items were significant: 7, 8, 11, 24, 29 (the light colored bars in Fig. 4). For item 7, characters on the computer screen were easy to read ($D(6) = 0.7$, $p < 0.01$). For item 8, highlighting on the screen simplified the task ($D(6) = 0.53$, $p < 0.05$). For item 11, the use of terms throughout the system was consistent ($D(6) = 0.53$, $p < 0.05$). For item 24, the system was reliable ($D(5) = 0.6$, $p < 0.05$; There were only 5 degrees of freedom (d.f.) because one participant did not respond to item 24). For item 29, the sequence of operations was clear ($D(6) = 0.53$, $p < 0.05$). Overall, these findings suggested participants had a positive impression of the user interface.

Comments volunteered by participants in response to QUIS-a items identified areas for improvement. On items 11, 17, and 20, which asked about confusion and difficultly of the system, participants remarked that the terms used were confusing at times and terms used in SPEC did not match the terms they used to think about specifications. Another topic of agreement was items 9a and 9b in which participants reported that the variable tree in the Edit Panel was difficult to navigate. This topic was further explored in the semistructured interviews, which are presented below.

The responses to interview questions were informative feedback on many areas but here we focus only on topics that

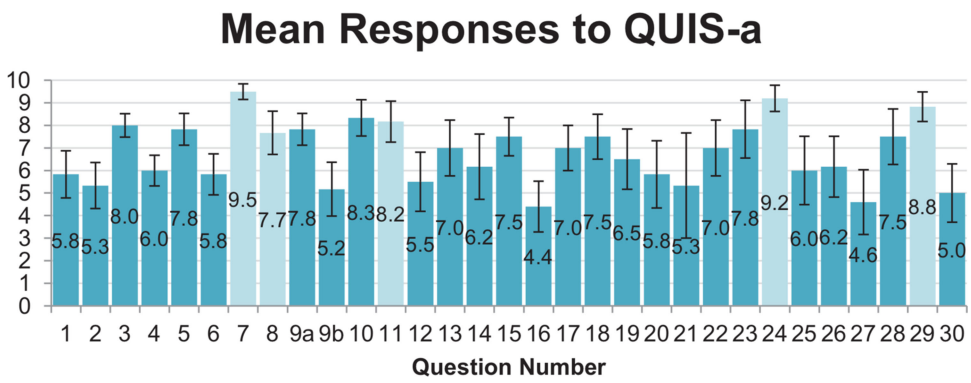


Fig. 4. Mean and standard error of QUIS-a responses from the usability study. Light-colored bars are significantly different from a uniform distribution of responses based on the Kolmogorov–Smirnov test (see text for explanation).

achieved the most consensus amongst participants, and which we were addressed when SPEC was refined. Most participants utilized an interleaved workflow for completing the mission planning and specification writing and checking (four of six participants) rather than a sequential workflow. Follow-up questions on workflow indicated that the interface is well-suited to interleaving and a few participants felt that the interface was not suited to building the patterns first. Participants were asked if it was clear how to fix issues that were flagged by the model checker. In general, participants felt that the model checker feedback was often ambiguous and more feedback about the cause of the failure or even a suggested solution to the failure would improve the interface. Participants mentioned that ambiguity often arose from troubleshooting the source of the problem: the UAV plans did not fulfill the mission or there was a mistake creating specification in SPEC (e.g., wrong pattern selected, wrong variables, etc.). One participant requested a visualization to show the violations flagged by the model checker on the map, as more informative feedback and to assist in the troubleshooting. When asked if they found the patterns in the Pattern Library easy to understand, most participants said that some patterns were easy to understand and other patterns were not easy to understand, but the examples each participant used did not agree. However, when asked to provide any wording suggestions to improve the patterns, most participants did not have any.

Participants were asked to provide design ideas, if they had them, for different components of SPEC. Interestingly, one participant did suggest typing the pattern and having it translated to temporal logic, which was a previously explored approach in this project (e.g., [15]). Many design ideas were offered to improve interaction with the Edit Panel. Consensus emerged that the variable tree was hard to navigate and too large when all the levels are expanded, but suggestions differed in how to improve it (e.g., closing all the lower tree levels or having a collapse all button). Three participants wanted alternative methods to enter variable information, either typing or from drop down (i.e., "combo") boxes. Three participants mentioned having the selected variable automatically advance to the next variable or by pressing the tab key on the keyboard, in addition to mouse click selection. Participants were asked about design ideas for integrating SPEC and the control station's map. Participants agreed that integration between SPEC and the map would be beneficial but were not sure how to accomplish it. One participant suggested a timeline display to go underneath the map and another suggested integrating pattern writing into the route planning interfaces of VSCS.

C. Discussion

Participant feedback was generally positive, showing the promise of these kinds of technologies and affirming the initial design work. The usability study also generated many ideas for further improvement of SPEC that might contribute to the impact of verification and validation tools. Specifically, participants desired more feedback from the model checker, found some pattern wordings difficult to understand, and found the Edit Panel's variable tree hard to navigate. These ideas became the focus for refinements to the design of SPEC.

The interviews revealed an unexpected troubleshooting aspect of using any verification tool. Participants desired more feedback from the model checker so they could determine more easily if the specification violation was due to a problem in their UAV plans or a problem in their specification creation. This troubleshooting process for verifying UAV mission plans largely parallels the analysis of model checking results in the original domain of model checking, hardware and software verification. Consider the UAV plans to be our instance of a system model and the specifications to be our instances of system properties. Baier and Katoen [19] discuss three possible causes of a specification violation. There may be a *modeling* error (the model is not a valid representation of the system), a *design* error (the system has a flaw which has been discovered), or a *property* error (the specification checked does not reflect the informal requirement). Participants of their own accord realized the possibilities for design errors (i.e., problems with their plans) and property errors (i.e., problems with their specifications). Diagnosing the cause of a negative result is not simple, and better feedback would most likely enhance the distinctions between design errors and property errors. (No participants encountered or suspected modeling errors, though these errors might be a real possibility for a fielded system and the task of

TABLE I
EXAMPLES OF PATTERN DESCRIPTIONS ADDED TO SPEC VERSION 2 TO INCREASE PATTERN UNDERSTANDING

Pattern	*Description*	*NuSMV Syntax*
Before the first occurrence of {Y}, {VAR} happens at least once	Before Bob leaves the grocery store, Bob checks his grocery list one or more times	G(!(Y)) | (!(Y) U (VAR & !(Y)))
Across all time, if {VAR1} happens then {VAR2} happens after at most {C}	Every time Bob gets a mortgage bill, he pays it within 30 days	AG(VAR1 −> ABF 1..C VAR2)
After the first occurrence of {X}, whenever {VAR} happens it holds for at least {C}	After Jill moved to the beach, whenever her parents visit her they stay for at least 2 weeks	AG((X) −> AG((!(VAR) & AX(VAR)) −> ABG 1..C VAR))

investigating and identifying these errors using SPEC is an area for future work.)

V. DESIGN REFINEMENTS

The usability test identified three areas of improvement in the SPEC design: some patterns were difficult to understand, the Edit Panel interaction, and model checker feedback. For a second version of SPEC, we addressed two of these areas: patterns were difficult to understand at times, and the interaction with the Edit Panel. These changes made to SPEC are detailed below. Model checking feedback, such as visualization of the counterexample or the flight plans, was not addressed as it is an active research area that would entail more formative design research than was within the scope of the current study.

A. Pattern Understanding

To increase understanding of patterns, two changes were made. First, the wording of many of the patterns was changed in an effort to make them more readable and less formalized while trying to avoid ambiguity in the temporal relationships they express. Second, a supplemental description of each pattern was developed with everyday circumstances, to try to further convey the logical relationship in a natural intuitive example (Table I).

B. Edit Panel Interaction

To improve the navigation of the Edit Panel, the majority of the interaction changes focused on the tree viewer's behavior. Many complaints focused on the length of the tree list due to the many variables in the NuSMV model and on the tree viewer's expand/collapse behavior. We added a "collapse all" button to provide a quick method to clean up the variable tree. In addition, we modified how actions at a higher (i.e., parent) level of the tree affect the lower (i.e., child) levels of the tree, so that any collapse action would also collapse all children of that level, rather than the default behavior where each level has to be manually collapsed.

The Edit Panel was criticized for the heavy mouse usage in selecting which variable to assign (at the top of the panel) and which variable value the user wants to use (in the variable tree). To address these issues, we had the selected variable (i.e., the variable that will be assigned when something is selected in the tree) automatically advance to the next variable in the pattern after a selection was made. Version 1 of SPEC required users to manually return to the list of variables within the pattern and change which variable was selected via mouse click. This one change eliminated a lot of mouse travel.

One change made to the Edit Panel was not motivated by the usability study but rather by a design goal of furthering the integration of the temporal patterns and the spatial representation on the control station's map. We added buttons to the variable tree that allowed the user to assign variables by selecting objects (e.g., Points, Areas, ROZ) from the VSCS tactical situation display (i.e., map).

VI. EXPERIMENTAL EVALUATION

An experiment tested if verification and validation tools can improve the accuracy of mission plans. The experiment was designed to measure the impacts of SPEC on mission planning accuracy, mission planning time, and mental workload. We hypothesized that using SPEC would increase mission planning accuracy but would also increase mission planning time, as it will likely take longer to do mission planning and checking than just mission planning alone. We hypothesized that using SPEC may also increase mental workload, from both writing specifications in SPEC and having to revise mission plans after receiving feedback from SPEC.

A. Method

The methods are summarized below, and more details can be obtained in [5].

1) Participants: Twelve individuals (four female) volunteered to participate in this experiment. None of these participants had prior experience with SPEC. Some participants had previous experience with prototype UAV ground control stations including VSCS.

2) Materials: The experiment relied on the same setting and questionnaire as the usability test. The VTASC displays were modified to manipulate the availability of SPEC. When SPEC was present, it was shown on the right monitor of VTASC. When SPEC was absent, the right monitor was blank. Two mission planning scenarios (referred to as A and B) were developed based on the scenario used in the usability study. They each had eight mission objectives that were similar in content and approximately equal in difficulty. Scenarios were counterbalanced with display conditions during the experiment to control for any unintended differences. Workload was measured using a computerized version of the NASA Task Load Index (NASA-TLX; [20]). The pairwise comparison aspect of the test was completed once, at the end of the experiment, and the weights applied to

all of that participant's responses. In addition, we performed a semistructured interview slightly modified from what was used in the usability study. It was shortened for the sake of time and one question was added to solicit feedback on the perceived benefit and value of SPEC.

3) Procedure: The experimental design was a 2×2 design with the within-subject factors display condition (SPEC absent or present) and scenario (A or B). Each participant performed two trials, one with SPEC and one without SPEC. The combination of display condition and scenario was counterbalanced across participants, making the interaction between-subjects. (The interaction was not a focus of this study and steps were taken to equate the two scenarios.) Also, the order of conditions was counterbalanced across participants to control for any order effects. Upon arriving to the lab, participants read informed consent documents and asked any questions. After providing their consent, they were assigned to one of four sequences depending on the counterbalanced combinations of display conditions and scenarios. The sequence of training differed depending on if SPEC was present or absent in the first trial. If SPEC was absent in the first trial, participants were trained on how to create mission plans using VSCS's route planning interfaces. Training consisted of: an explanation of the features and various methods of mission planning in VSCS, a demonstration of mission planning, and hands-on practice mission planning. Training was achievement-based and continued until participants demonstrated they could create a number of specific vehicle plans, which exercised many different mission planning features.

After completion of training, participants completed the first trial. All trials were self-paced and participants were instructed to work on the mission plan until they were satisfied they had met all the objectives. Accuracy was emphasized over speed. The completion time was recorded by the experimenter using a stop watch. After trial completion, participants filled out the NASA-TLX. Then prior to the second trial, participants were trained on how to use SPEC. SPEC training consisted of:

1) An explanation of the tool's features;
2) A discussion of the types of patterns available;
3) A demonstration of writing specifications with SPEC;
4) Hands-on practice using SPEC to create specifications that exercise many different aspects of the tool and types of patterns;
5) Practice using SPEC in a scenario—i.e., checking those specifications against mission plans, receiving feedback from SPEC, and addressing problems both in the mission plan and the specification.

After training, participants completed the second trial followed by the NASA-TLX questionnaire and pairwise comparison selections. If SPEC was present in the first trial, participants were trained on creating mission plans in VSCS followed by training on how to use SPEC. Then, they completed the first trial and NASA-TLX questionnaire followed by the second trial and NASA-TLX questionnaire with pairwise comparison. Across participants, total training time averaged 1.43 h.

After the completion of both trials (irrespective of order), participants completed the QUIS-a in regards to the SPEC tool. They were asked specifically to not consider VSCS in their responses, as that was not the focus of this research. Once QUIS-a was completed, the semistructured interview was conducted. The whole session lasted 3 h and 45 min on average.

TABLE II
SUMMARY OF EXPERIMENTAL RESULTS SHOWING MEAN AND *P*-VALUE FOR THE DISPLAY MANIPULATION AND THE TWO SCENARIOS (SEE TEXT FOR DETAILS AND WORKLOAD SUBSCALES)

Condition	*Accuracy*		*Time (s)*		*Overall Workload*	
	M	p	M	p	M	p
SPEC Present	0.92	$< .05$	2291	$< .001$	42	1.0
SPEC Absent	0.80		997		42	
Scenario A	0.78	.31	1820	$< .05$	47	$< .05$
Scenario B	0.86	-	1202		36	

4) Analyses: Quantitative analysis of the experimental manipulations was complemented by auxiliary analyses of the qualitative data. The quantitative analyses focused on mission planning accuracy, mission planning time, and workload. The auxiliary analyses addressed if participants were successful using SPEC and analyzed the ratings and comments collected in the questionnaires and interviews. To further investigate participants' performance with SPEC, we analyzed the accuracy of writing specifications in SPEC for each scenario mission objective. In addition, we performed a correlation between mission planning accuracy and specification writing accuracy. One requirement for the utility of the model checker is correct specifications. Without participants writing correct specifications, the model checker feedback may cause confusion because the feedback is not entirely due to the plan. In other words, the model checker could return a negative result when an incorrect specification was submitted on a correct plan or it could return a positive result when an incorrect specification was submitted on an incorrect plan. Therefore, we suspected that participants' performance in writing specifications may be correlated with their performance in mission planning.

To tentatively assess the changes made to SPEC after the usability study, we compared the QUIS-a scores from the experiment to those from the usability study, which used different versions of the SPEC display. This tentative examination used descriptive statistics and did not perform any statistical tests.

VII. RESULTS

The following results showed that the presence of SPEC led to increased mission planning accuracy as well as increased mission planning time, but there was not a significant increase in subjective workload. The scenarios were not significantly different in planning accuracy, however Scenario A took longer than Scenario B and Scenario A was rated higher in overall workload and temporal demand than Scenario B. Table II shows a summary of the experimental results. Additional analyses of participants' interactions with SPEC using quantitative, questionnaire and interview methods showed that in general participants were successful using SPEC appropriately, the second version of SPEC was an improvement over the first version, and

participants found a verification tool to be valuable for mission planning.

A. Experimental Results

We calculated mission planning performance scores for each block as the proportion of mission objectives that were met by the vehicle plans participants created. A preliminary review of descriptive statistics suggested that outliers may be present in the data. We used a quartile-based outlier identification technique, where a threshold for outliers was set based on the 1st and 3rd Quartile ± 1.5 times the interquartile range. We performed this calculation for each treatment cell individually. That is, we did this for the SPEC absent data and again for the SPEC present data. The lower threshold value for the SPEC absent data was 0.5625 and one participant's score of 0.25 was below that value. The lower threshold value for the SPEC present data was 0.75 and another participant's score of 0.625 was below that value. Both of these outlier values were obtained on the first trial suggesting perhaps these subjects needed additional training or had difficulty comprehending the instructions, despite the achievement-based training criteria and ample opportunities for questions (though this need for more training was not apparent in the total data, as indicated by the non-significant test for order effects below). To preserve the balanced nature of our design, we then sought to impute data values for these two outliers. Some researchers advocate imputing the outlier values with the respective treatment means after removal of the outliers, but a more conservative practice (that can actually increase Type-II error) is to use the overall mean after removal of the outliers [21]. We used the more conservative practice, and the overall mean value used for imputation was 0.8579.

A repeated-measures analysis of variance (ANOVA) of performance scores was done to test for order effects, as well as display condition and scenario effects. No order effects were found ($F(1, 9) = 0.35$, $p = 0.57$). Scenario, even though it was crossed with display conditions, was analyzed in order to check that we met our intention of having scenarios that were similar in difficulty. The mean mission planning performance for Scenario A was 0.78 and for Scenario B was 0.86. This difference was not significant, $F(1, 9) = 1.17$, $p = 0.31$. The main effect of display condition was significant, $F(1, 9) = 5.68$, $p < 0.05$, $\eta_p^2 = 0.39$. SPEC present trials had higher performance compared to SPEC absent trials (0.915 versus 0.801, respectively).

Mission planning time was analyzed after log transform of the completion times, to reduce the positive skew in the distribution of values. We conducted a repeated-measures ANOVA with display condition, scenario, and order as factors. There was no order effect, $F(1, 9) = 0.01$, $p = 0.98$. There was a main effect of display condition ($F(1, 9) = 42.58$, $p < 0.001$, $\eta_p^2 = 0.83$). When SPEC was present, participants took longer to plan (M $= 3.36$ log seconds, SD $= 0.17$) than when SPEC was absent (M $= 2.99$ log seconds, SD $= 0.21$). The units of log seconds can be difficult to interpret, so we have converted the means into seconds for comparison; planning with SPEC present was 2290.87 s whereas planning without SPEC was 997.24 s. There was a main effect of scenario as well ($F(1, 9) = 10.01$, $p < 0.05$, $\eta_p^2 = 0.53$). The mean (SD) planning time for Scenario A was 3.26 (0.23) log seconds and for Scenario B was 3.08 (0.27) log seconds. Expressed in seconds, Scenario A times averaged 1819.7 s and Scenario B times averaged 1202.3 s.

NASA-TLX scores of overall workload and each subscale were analyzed using repeated-measures ANOVA with factors of display condition and scenario. Only significant results are presented in full for the sake of brevity. All analyses of display condition were not significant (all $p > 0.10$). For overall workload, there was a main effect of scenario ($F(1, 9) = 5.35$, $p < 0.05$, $\eta_p^2 = 0.37$), with Scenario A being perceived as more workload than Scenario B (47.31 and 36.31, respectively). For the temporal demand subscale, there was a significant main effect of scenario ($F(1, 9) = 5.94$, $p < 0.05$, $\eta_p^2 = 0.40$), Scenario A was higher than Scenario B (38.33 and 27.50, respectively).

B. Auxiliary Results

The overall accuracy for specification writing was 81.25% (SD $= 21.8$). The data indicated that one type of objective was problematic—requiring overwatch of the Convoy while it was in a particular location (accuracy $= 41.5\%$). A correlation between SPEC writing accuracy and mission planning accuracy was significant $r = 0.71$, $p < 0.001$ (95% CI $= 0.23 - 0.91$).

The QUIS-a responses for the experiment were compared to the QUIS-a responses from the usability study. The mean difference was 1.75 (SD $= 1.43$), suggesting that participants generally rated the experiment higher than the usability study. Participant's interview responses were analyzed as before. Consistent with the usability study, participants asked for ways to specify more than one UAV in each pattern, and for improvements to the model checker feedback. Most participants found the patterns in the Pattern Library easy to understand, and participants were divided on the organization of the library. Some found the categorization to be helpful where others did not and were frustrated by manually searching amongst the categories. Last, participants were asked about design ideas for the pattern Edit Panel. Some participants (4 of 12) appreciated the added collapse all button, yet others (4 of 12) still had difficulties with the length of the variable tree. All participants liked the tool, thought it added value and provided additional validation of their plans. Furthermore, participants who used SPEC first felt that they were not as confident in their plans on their second trial, when they did not have SPEC.

VIII. Discussion

Findings showed a tradeoff between the speed and accuracy of planning associated with a model checking tool. When using SPEC, planning accuracy was higher (91.5% with SPEC versus 80.1% without SPEC) but also took substantially more time (an additional 20 min on average). In a typical speed-accuracy tradeoff situation, any additional time spent results in increased accuracy. Here, we do not know what performance would have been like if participants had spent 20 extra minutes planning

when they did not have SPEC. The extra time could have also improved performance, however it may not have because participants already were instructed to emphasize accuracy over speed. When considering the extra time spent using SPEC, it is important to recall that using SPEC did not contribute to increased perceptions of workload. Nonetheless, efficiency is paramount and future research might investigate how to reduce planning time by providing additional training and familiarization with the tool, modifying the interface to improve interaction, or some combination of the two.

With regard to the scenarios, the results indicated that our attempts to balance the difficulty of the scenarios were not entirely successful. Scenario A appeared to be more challenging than Scenario B. Participants took longer with Scenario A than Scenario B and rated Scenario A as more demanding than Scenario B. The scenario did not significantly affect planning performance, and the scenario was crossed with the availability of the SPEC tool so the interpretation of the SPEC tool results was not affected.

The analysis of specification writing accuracy showed that accuracy was high, suggesting that pattern-based specifications can be learned quickly and are a promising candidate for effective communication of intent. As future systems undertake synthesis-generated mission plans, sound methods to accurately communicate specifications will be important to safety and predictability. Specification accuracy revealed that one type of objective was problematic for participants, the related pattern was probably difficult to understand and needs revision. In addition, the relationship between specification writing accuracy and mission planning performance merits further study. This relationship may indicate either that having correct specifications to check against leads to accurate mission plans or that participants who perform mission planning well also write specifications well.

IX. Conclusion

This article described the development and testing of a tool for verifying UAV mission planning, providing the user a method to communicate their high-level requirements to the model checking software. We found that the verification tool did increase mission planning accuracy, while also increasing mission planning time. These results indicate that verification tools such as SPEC have the potential to assist human operators in precisely planning complex missions. Additional research and development may further the potential by deriving more efficient methods to communicate and troubleshoot specifications.

Acknowledgment

The authors would like to thank the anonymous reviewers for their suggestions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

References

[1] H. A. Ruff, S. Narayanan, and M. H. Draper, "Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles," *Presence: Teleoperators Virtual Environ.*, vol. 11, no. 4, pp. 335–351, 2002.

[2] J. Y. Chen, M. J. Barnes, and M. Harper-Sciarini, "Supervisory control of multiple robots: Human-performance issues and user-interface design," *IEEE Trans. Syst., Man Cybern., Part C: Appl. Rev.*, vol. 41, no. 4, pp. 435–454, Jul. 2011.

[3] M. L. Cummings, L. F. Bertucelli, J. Macbeth, and A. Surana, "Task versus vehicle-based control paradigms in multiple unmanned vehicle supervision by a single operator," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 3, pp. 353–361, Jun. 2014.

[4] M. B. Dwyer, G. S. Avrunin, and J. C. Corbett, "Patterns in property specifications for finite-state verification," in *Proc. IEEE Int. Conf. Softw. Eng.*, 1999, pp. 411–420.

[5] C. Rothwell, M. Patzek, and L. Humphrey, "Verifiable task assignment and scheduling controller," 711 Human Performance Wing Wright-Patterson AFB United States, Tech. Rep., AFRL-RH-WP-TR-2017-0045, 2017.

[6] L. Humphrey, "Model checking UAV mission plans," in *Proc. Amer. Inst. Aeronaut. Astronaut. Conf. Model. Simul. Technol.*, 2012, pp. 1–19, Paper 4723.

[7] P. Doherty, J. Kvarnström, and F. Heintz, "A temporal logic-based planning and execution monitoring framework for unmanned aircraft systems," *Auton. Agents Multi-Agent Syst.*, vol. 19, no. 3, pp. 332–377, 2009.

[8] L. R. Humphrey and M. J. Patzek, "Model checking human-automation UAV mission plans," in *Proc. Amer. Inst. Aeronaut. Astronaut. Guid., Navigat., Control Conf.*, 2013, pp. 1–15, Paper 5183.

[9] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, "Ten challenges for making automation a "team player" in joint human-agent activity," *IEEE Intell. Syst.*, vol. 19, no. 6, pp. 91–95, Nov.–Dec. 2004.

[10] A. Rowe, K. Liggett, and J. Davis, "Vigilant spirit control station: A research testbed for multi-UAS supervisory control interfaces," in *Proc. Int. Symp. Aviation Psychol.*, Dayton, OH, 2009, pp. 287–292.

[11] G. L. Feitshans, A. J. Rowe, J. E. Davis, M. Holland, and L. Berger, "Vigilant Spirit Control Station (VSCS): The face of COUNTER," in *Proc. Amer. Inst. Aeronaut. Astronaut. Guid., Navigat. Control Conf.*, 2008, pp. 1–16, Paper 6309.

[12] R. Cavada *et al.*, *NuSMV 2.5 User Manual*, 2010. [Online]. Available: http://nusmv.fbk.eu/NuSMV/userman/v25/nusmv.pdf, Accessed: Jun. 24, 2013.

[13] D. Brinksma and K. G. Larsen, "NuSMV 2: An opensource tool for symbolic model checking," in *Proc. Int. Conf. Comput. Aided Verification*. Berlin, Heidelberg: Springer, 2002, pp. 359–364.

[14] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas, "Translating structured English to robot controllers," *Adv. Robot.*, vol. 22, no. 12, pp. 1343–1359, 2008.

[15] C. Rothwell, A. Eggert, M. Patzek, G. Bearden, G. Calhoun, and L. Humphrey, "Human-computer interface concepts for verifiable mission specification, planning, and management," in *Proc. Amer. Inst. Aeronaut. Astronaut. Infotech@ Aerospace Conf.*, 2013, pp. 1–15, Paper 4804.

[16] S. Konrad and B. H. Cheng, "Facilitating the construction of specification pattern-based properties," in *Proc. IEEE Int. Conf. Requirements Eng.*, 2005, pp. 329–338.

[17] S. Konrad and B. H. Cheng, "Real-time specification patterns," in *Proc. Int. Conf. Softw. Eng.*, ACM, 2005, pp. 372–381.

[18] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proc. Special Interest Group Comput.–Human Interact. Conf. Human Factors Comput. Syst. Assoc. Comput. Machinery*, 1988, pp. 213–218.

[19] C. Baier, J.-P. Katoen, and K. G. Larsen, *Principles of Model Checking*. Cambridge, MA, USA: MIT Press, 2008.

[20] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Advances Psychol.*, vol. 52, pp. 139–183, 1988.

[21] B. G. Tabachnick, L. S. Fidell, and S. J. Osterlind, *Using Multivariate Statistics*. Boston, MA, USA: Allyn and Bacon, 2001.

Through the Looking Glass(es): Impacts of Wearable Augmented Reality Displays on Operators in a Safety-Critical System

Aaron Rowen, Martha Grabowski, and Jean-Philippe Rancy

***Abstract*—Novel information technologies such as wearable augmented reality displays (WARDs) can provide operators of safety-critical systems with crucial real-time information anywhere in the work space. The evolving capabilities and proliferation of these technologies occasion important questions about their impact on operators in safety-critical systems such as nuclear power plants, aircraft cockpits, and ships. This article describes a WARD evaluation in the safety-critical setting of marine transportation in a real-time physical simulator. A large pool ($n = 211$) of operators in conventional and WARD technology conditions showed mixed results: WARD use dampened operator performance variability, improved situation awareness (SA), self-efficacy, and trust, but also increased operator workload. These results suggest that WARDs show promise as a mobile information display technology in one safety-critical system where operator mobility is important, and also suggest that increased workload may be a cost accompanying the desirable outcomes of reduced performance variability and increased SA.**

***Index Terms*—Augmented reality, human computer interaction, maritime safety, technology evaluation, technology impact, wearable computers.**

I. Introduction

OPERATORS in safety-critical systems require timely, relevant information to support rapid, effective task response [1], particularly in settings that require them to move about the workspace away from fixed visual displays, such as in nuclear power plants [1], medicine [2], emergency response [3], search and rescue [4], offshore oil and gas [5], and marine transportation [6]. The availability of real-time, context-relevant information can aid in developing operator trust in information [7] in safety-critical systems, where effective human–computer interaction is crucial to saving lives, protecting property, and preserving the environment [8].

Wearable augmented reality displays (WARDs) superimpose information on views of the physical world [9], presenting information in context [10] to operators in a form factor that facilitates hands-free and mobile operations [2]. Earlier research evaluating display technologies in safety-critical systems relied on the task-technology fit (TTF) model [11] and Lee and Moray's model of trust, control, and function [7], with mixed results. WARD use has been observed to improve operator performance [2] and situation awareness (SA) [2], [6], while showing reduced operator performance and delayed responses due to information overload [12], visual clutter [13], [14], limited information visibility [15], and increased workload [6]. Few studies considered operator trust [16] and tradeoffs among impacts of novel technology [2], [6], [17], and many utilized small subject pools [6], [18]. Thus, despite the promise of mobile AR displays, questions persist about their impacts on operators in safety-critical systems [15].

The advent of WARDs in safety-critical systems and the mixed results in earlier studies motivate this empirical study of 211 operators in one safety-critical system, marine transportation. Operator performance and processes with WARDs and fixed, conventional displays were evaluated in a real-time physical simulator. The results showed WARD use reduced operator performance variability and improved operator SA, self-efficacy, and trust, but also increased workload, consistent with previous work [2], [6], [15], [18]. The results raise tradeoff and interaction questions and suggest that trusted information displays that reduce performance variability and improve SA may impose a workload cost.

II. Background

A. Visual Displays

Visual displays integrate information from dispersed sensors and systems so that it is comprehensible [19], [20], providing decision support in complex systems [1]. Earlier evaluations of head-up displays [21] considered impacts on operator performance, based on operator response time [12] and accuracy, showing that displays that obscure information [17] or present

Manuscript received August 20, 2018; revised March 8, 2019 and May 23, 2019; accepted September 16, 2019. Date of publication October 21, 2019; date of current version November 21, 2019. This work was partially supported by the Maritime Institute of Technology and Graduate Studies—Pacific Maritime Institute and The McDevitt Foundation at Le Moyne College. The work of M. Grabowski was supported by Google. This article was recommended by Associate Editor G. Fortino. *(Corresponding author: Aaron Rowen.)*

A. Rowen is with the Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: rowena@rpi.edu).

M. Grabowski is with the Information Systems Program, Madden School of Business at Le Moyne College, Syracuse, NY 13214 USA, and also with the Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: grabowsk@lemoyne.edu).

J.-P. Rancy is with the School of Information Studies, Syracuse University, Syracuse, NY 13210 USA (e-mail: jrancy@syr.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2944384

too much or irrelevant information [19], [21] negatively impact task performance.

At the same time, displays of available [21], relevant [22], and organized [19] information can improve SA [20], the perception and comprehension of elements in the environment, and their projection into likely future states [23], but have also been associated with increased stress and workload [22]. During complex tasks that impose high workload and stress [24], the performance and SA benefits of visual displays can be reduced or reversed [12], possibly due to competition for cognitive resources [22].

Visual displays have been shown to positively impact operator self-efficacy, the belief that operators are empowered to complete tasks when using technology [25], especially when displays are familiar and perceived as useful [26]. Reduced self-efficacy can occur with unmet expectations of display capabilities [25] or a negative predisposition to the technology.

Propensity to trust describes an individual's trust in a technology [27]. Trust in visual displays can be enhanced by relevant information appearing at an appropriate time [16], causing the technology to be perceived as useful [7]. Conversely, trust is diminished when displays show irrelevant [16] or untimely [28] information. These impacts can be moderated by gender, age, and experience [25], and may also moderate perceptions of stress in complex, safety-critical settings [29].

B. Wearable Augmented Reality Displays (WARDs)

WARDs superimpose information on views of the physical world [9], integrating information from dispersed sensors, systems, and the operational situation. AR displays can be situated in proximity to pertinent physical entities [30], providing information relevant to the location or situation [10]. Evaluations of WARD impacts have shown operator performance [2], [15] and SA [6] improvements when critical information was provided in real time [18]. At the same time, WARD use increased workload [6], and raised questions about information visibility [13], [15], understandability [17], and clutter [14], [21].

To date, WARD evaluations have utilized small, controlled experiments [6], [15] rather than evaluating large numbers of experienced subjects in operational or simulated operational environments, primarily because of the complexity of such studies [18]. Previous WARD evaluations have not examined performance variability in a large subject pool, links between operator performance and perceptions of workload and stress, or impacts among operator performance, SA, and perceptions [13], [16], despite acknowledged relationships among these factors [6], [28]. This lack of synthesis presents a simplified view of technology impacts [31] that fails to explore tradeoffs among technology impacts. Examining these tradeoffs may be important in safety-critical systems where technology benefits have been reduced or reversed in high tempo and high workload operations [12].

III. Research Model

The technology-to-performance chain (TPC) (see Fig. 1), which proposes that operator performance with technology is based on operators' perceptions of compatibility among task, technology, and operator characteristics, known as TTF [11], has been used to evaluate display impacts. Self-efficacy [25], stress [32], and trust [7] have been linked in TTF studies [33], showing that improved operator performance with novel technologies has been positively associated with self-efficacy [26] and trust [27], and negatively associated with workload [22] and stress [33]; however, the link between operator performance and fit has been less explored.

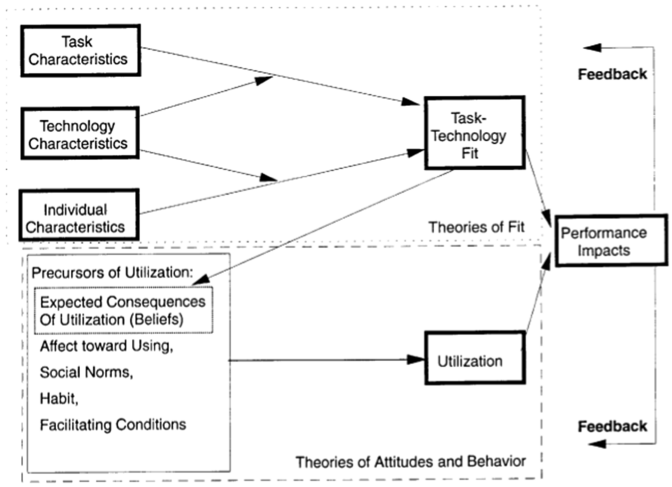


Fig. 1. Goodhue's TTPC [11].

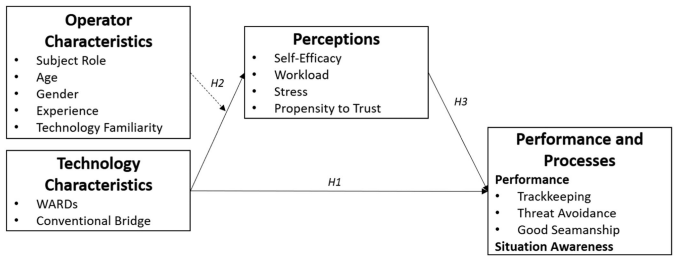


Fig. 2. Research Model: Impacts of WARDs technology on operator performance and processes. *Based on the Technology-to-Performance Chain [11].*

This article follows a similar thread, exploring how operator and technology characteristics influence operator performance, processes, and perceptions in a safety-critical system (see Fig. 2). Previous work showed that WARD introduction improved operator performance [2] and enhanced SA [6]. Some studies showed mixed results with small subject pools [2], [6], [15], [18]. Thus, our research model explores these relationships proposing.

Hypothesis 1: Operators will show improved performance and SA when using WARDs.

Previous work suggests that WARD impacts on operators may be moderated by operator age, gender, and experience [29], as well as by perceptions of self-efficacy, workload, stress, and trust [14], [15], [17]. Hypothesis 2 explores the moderating impact of operator characteristics on operator perceptions when using a WARD.

Hypothesis 2: Operator characteristics will moderate operator perceptions when using WARDs.

Fig. 3. MITAGS-PMI simulator system showing bridge equipment display configuration, as well as the view facing forward on the ship with approaching traffic.

Operator perceptions of self-efficacy, workload, stress, and trust have been linked to performance [26] and SA processes of perception and comprehension. Hypothesis 3 provides a link between Hypotheses 1 and 2, exploring the relationships among operator perceptions, performance, and SA when using WARDs.

Hypothesis 3: Operator perceptions will positively impact performance and SA when using WARDs.

Many social cognitive variables have been considered singularly in WARD studies [2], [5], [13], and links between technology and performance, workload and SA, and self-efficacy and TTF have been studied [15], [26], [34]. Despite demonstrated links between technology impacts and social cognitive factors [22], [26], [27], no model has been proposed that explores these links together. Our research model addresses these gaps by expanding the TPC model to consider social cognitive factors such as self-efficacy, workload, stress, and trust, as well as operator processes such as SA.

IV. Operational Setting: Marine Transportation

This article is set in the safety-critical setting of marine transportation, where ship operators—captains, mates, and pilots—are responsible for the safe navigation of their vessel in a complex environment requiring cognitive resources, computation, and information retrieval and processing; as well as interaction with the environment, operators on other ships, and shoreside authorities; while moving about the bridge, often out of visible range of the fixed displays shown in Fig. 3. This system is governed by the International Rules of the Road [35], by local regulations, and by norms and practices known as the "practice of good seamanship" [36].

Ship operators' ability to access timely, accurate, and relevant information is important as they move about the ship's bridge and vessel during ship navigation, docking, anchoring, and search and rescue evolutions, which often take place in inclement weather, restricted visibility, and challenging conditions. Additional operator demands come from imperfect automation, multiple competing tasks, new technologies [37], and the need to monitor multiple visual displays (see Fig. 3) [38]. Balancing these demands in a time-constrained decision environment while anticipating the ship's movements many minutes ahead of a decision [39] makes the ship navigation task cognitively complex [37].

To perform their tasks, ship operators rely on an array of information provided by visual displays (see Fig. 3) that support trackkeeping on an intended track [38], [40], threat and collision avoidance [41], SA [6], and the practice of good seamanship [35]. Display impacts are important in ship operations as failures to monitor visual displays [42] or maintain SA [43] can be costly and dangerous [44].

Effective operator performance during ship navigation tasks has been associated with adherence to the vessel's intended track (i.e., small cross track errors (XTEs) [40]), particularly in narrow channels where there is danger of grounding [44]; and good threat avoidance by coordination with other vessels (i.e., large closest points of approach (CPAs) [45]); while abiding by local and international guidelines and regulations [35]. Also important are low levels of operator workload and stress, and improved SA [38], [40]. These metrics have been considered in marine piloting studies [45], as well as in studies of embedded intelligent real-time systems for ship's piloting [41], and ship navigation AR displays [6].

V. Evaluation

This research considers the impact of WARDs on operator performance, processes, and perceptions in one safety-critical system, marine transportation, exploring the relationships presented in Fig. 2. Licensed U.S. Merchant Marine ship navigation officers performed a ship navigation task using a WARD ($n = 140$) or conventional technology ($n = 71$) in a real-time physical simulator. Data were gathered during the task, and data on operator perceptions of self-efficacy, workload, stress, and propensity to trust were gathered in post-task surveys. Hypotheses are listed in Table I, and measures and operationalizations are listed in Table II.

A. Research Vehicle and Technology

The study was conducted in a Class-A, full mission ship's bridge simulator located at the Maritime Institute of Technology

TABLE I
HYPOTHESES

H1a	Operators will show improved trackkeeping when using WARDs.
H1b	Operators will show reduced trackkeeping variability when using WARDs.
H1c	Operators will show improved threat avoidance when using WARDs.
H1d	Operators will show reduced threat avoidance variability when using WARDs.
H1e	Operators will show improved practice of good seamanship when using WARDs.
H1f	Operators will show improved SA when using WARDs.
H2a	Operator self efficacy when using WARDs will be moderated by operator characteristics.
H2b	Operator workload when using WARDs will be moderated by operator characteristics.
H2c	Operator stress when using WARDs will be moderated by operator characteristics.
H2d	Operator propensity to trust when using WARDs will be moderated by operator characteristics.
H3	Operator perceptions will positively impact performance and SA when using WARDs.

and Graduate Studies - Pacific Maritime Institute (MITAGS-PMI) in Linthicum Heights, MD, USA. Ship operators licensed as Masters (Captains), Mates, and Pilots manually piloted a simulated 725-ft, 35 000 deadweight-ton container ship through the buoyed channel leading to or from New York harbor (see Fig. 4) with the aid of a helmsman. Each 30-min transit into or out of New York harbor included vessel traffic in routine and exceptional situations; such transits require manual steering instead of automatic trackkeeping [35].

The WARD technology was an application known as GlassNav, developed by Le Moyne College researchers and implemented in Google Glass, version 2. GlassNav connected in real-time to the simulator electronic bridge equipment (radar, collision avoidance system, navigation sensors, etc.) and displayed a subset of this information superimposed on the user's view of the navigation situation (see Fig. 5).

The GlassNav displayed information valuable to the navigation task, including own ship heading (HDG) in degrees, the bearing (BRG) of a radar target in degrees, own ship course over ground in degrees, own ship speed over ground in knots, closest point of approach (CPA) to another vessel in nautical miles (nmi), and time to the closest point of approach in minutes and seconds. The status of other equipment in the system (i.e., "GPS FAIL" in Fig. 5) and deviation from the vessel's intended track, referred to as cross-track error (XTE), were also provided (see Fig. 5).

B. Subjects

Subjects ($n = 211$) were licensed U.S. Merchant Marine ship's officers voluntarily recruited from attendees of training courses at MITAGS-PMI. All subjects were qualified by the U.S. Coast Guard as Merchant Marine Officers (Officer in Charge of a Navigation Watch), had completed training on the conventional bridge equipment prior to the research transits, and had served as ship's Masters, Mates, and Pilots aboard U.S. flag vessels for an average of 13.56 years. Table III presents the operator characteristics of the subjects. No compensation for subjects was provided, and subjects volunteered for participation after hours, following their regularly scheduled and federally mandated simulator training. Subject availability resulted in an unbalanced ($n = 140$ WARD; $n = 71$ conventional) design, which was accounted for analytically, as described in Section V-F. To determine the impact of the unbalanced design on the results, additional analysis was undertaken; there was no impact on the results from the design, as described in Section V-F.

C. Procedure

Subjects were exposed to either an inbound or an outbound transit, using either the GlassNav WARD or the conventional bridge equipment. The transits were approximately 30 min in length, the length of time required to transit the approach to New York harbor, the setting for the study. During the first leg of the transit (ending at buoys 13–14 inbound or 17–18 outbound, Fig. 4), visibility was clear, currents were minimal, all bridge navigation equipment was functioning normally, and vessel traffic followed the International Rules of the Road. Following a transition leg, the final leg of the voyage (beginning at buoys 17–18 inbound, or 13–14 outbound, Fig. 4) presented subjects with reduced visibility, high currents, navigational equipment failures, and vessel traffic that behaved unpredictably and deviated from the International Rules of the Road. The extreme events increased the realism of the simulated task and exposed operators to a range of operational conditions.

Following a pretransit background survey, subjects were briefed on the purpose of the work, and gave informed consent. Subjects were then led to the simulator bridge, where they were briefed on the operational situation, instructed in the use of the WARD (when applicable), and given time to orient themselves to the setting and the equipment. Once the subject verbally acknowledged responsibility for navigation, the transit commenced. During the simulated transit, subjects interacted by radio with the Captain and with operators of other vessels (researchers outside of the simulated bridge acting in these roles). Upon completion of the simulated transit, subjects completed a posttransit survey and interview.

D. Data

Data were gathered from three sources: the simulator instruments and sensors, observation of the operators during the simulated transits using industry-standard task performance evaluation forms, and pre- and posttransit interview and survey data from the operators (see Table II). Simulator sensors measured operator performance, specifically operator trackkeeping and threat avoidance; operator processes were measured by observation and the use of validated assessment instruments; and operator characteristics were collected in pretransit surveys (see Table III), and perceptions were collected by posttransit surveys.

Operator performance items included three measures critical to maritime navigation: trackkeeping, threat avoidance, and the practice of good seamanship [36], [41], [45]. Trackkeeping involves maintaining the vessel's intended track and was assessed as the average of XTE, a simulator-captured measure of track deviation, recorded at one-second intervals by the simulator. Threat avoidance sets the course to maneuver around and avoid traffic and was also recorded by the simulator as the CPA to another vessel during traffic events. The practice of good seamanship describes qualitative ship management skills [41], evaluated by observation during the transit according to the

TABLE II
OPERATIONALIZATIONS AND MEASUREMENT OF DEPENDENT VARIABLES BY HYPOTHESIS

Hypothesis	Dependent Variable	Variable Operationalization	Data Collection Method
Hypothesis 1	Operator Performance and SA		
H1a	Trackkeeping	Mean cross-track error (XTE). Smaller XTE = better performance.	Simulator XTE
H1b	Trackkeeping variability	Mean cross-track error (XTE). Smaller variability = better performance.	Simulator XTE
H1c	Threat avoidance	Closest Point of Approach (CPA). Larger CPA = better performance.	Simulator CPA
H1d	Threat avoidance variability	Closest point of approach (CPA). Smaller variability = better performance.	Simulator CPA
H1e	Practice of good seamanship	Observations of qualitative ship management skills from the Navigation Skills Assessment Program (NSAP). Higher scores = better performance.	Transit observation with validated NSAP instrument
H1f	Situation Awareness (SA)	SA-1, perception; SA-2, comprehension from the Situation Awareness Global Assessment Technique (SAGAT) & Navigation Skills Assessment Program (NSAP). Higher scores = better SA.	Transit observation with validated NSAP instrument
Hypothesis 2	Operator Characteristics and Perceptions		
H2a	Self efficacy	New General Self Efficacy (NGSE) survey. Higher scores = better perception.	Post-transit survey
H2b	Workload	NASA Task-Load Index (TLX). Lower scores = better perception.	Post-transit survey
H2c	Stress	Short Stress State Questionnaire (SSSQ). Lower scores = better perception.	Post-transit survey
H2d	Propensity to trust	Propensity to Trust Scale (PTS). Higher scores = better perception.	Post-transit survey
Hypothesis 3	Operator Performance, Processes and Perceptions		
H3	Performance and processes	Mean cross-track error (XTE). Smaller XTE = better performance.	Simulator XTE
		Closest Point of Approach (CPA). Larger CPA = better performance.	Simulator CPA
		Observations of qualitative ship management skills from the Navigation Skills Assessment Program (NSAP). Higher scores = better performance.	Transit observation with validated NSAP instrument
		SA-1, perception; SA-2, comprehension from the Situation Awareness Global Assessment Technique (SAGAT) & Navigation Skills Assessment Program (NSAP). Higher scores = better SA.	Transit observation with validated NSAP instrument
	Perceptions	Self efficacy: New General Self Efficacy (NGSE) survey. Higher scores = better perception.	Post-transit survey
		Workload: NASA Task-Load Index (TLX). Lower scores = better perception.	Post-transit survey
		Stress: Short Stress State Questionnaire (SSSQ). Lower scores = better perception.	Post-transit survey
		Propensity to trust: Propensity to Trust Scale (PTS). Higher scores = better perception.	Post-transit survey

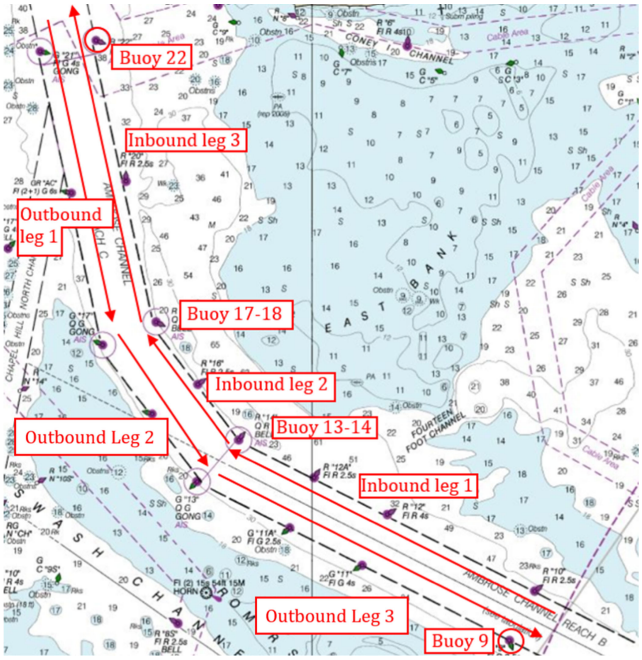


Fig. 4. Transit scenario showing transit tracks into and out of New York harbor and location of buoys.

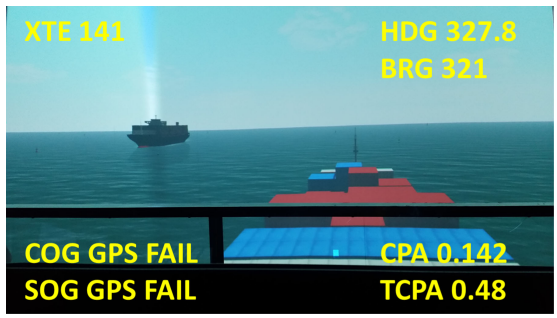


Fig. 5. GlassNav display superimposed on a view of a navigational transit.

Navigation Skills Assessment Program (NSAP), an externally validated instrument approved by the U.S. marine transportation system's regulatory agency, the U.S. Coast Guard [46].

SA was assessed by subject responses to Situation Awareness Global Assessment Technique (SAGAT) queries conducted during the simulation [23] without simulation pauses. Queries were asked and answered in real time, as routine interactions with a researcher acting as the Captain. The SAGAT queries were operationalized through the NSAP.

Operator perceptions of self-efficacy, workload, stress, and trust were assessed during post-transit surveys using validated instruments, as extensions to TTF. Self-efficacy was assessed by the New General Self-Efficacy survey (NGSE) [47]. Workload was evaluated using the NASA task-load index (TLX) [48]. Stress was measured with the Short Stress State Questionnaire (SSSQ) [49]. Propensity to trust was evaluated with the propensity to trust scale (PTS) [27].

E. Experimental Design

The study utilized a 3×2 between-subjects design: subjects in three roles (Masters, $n = 77$; Mates, $n = 115$; and Pilots, $n = 19$) were exposed to two technology treatments (WARDs, conventional bridge displays) over the course of the 18-month evaluation. The simulated transit task was balanced between two settings (inbound, $n = 107$; outbound $n = 104$). Each subject was exposed to one technology treatment, either a WARD ($n = 140$) or conventional bridge equipment ($n = 71$), in one transit. No repeat subjects or treatments were assessed, due to

TABLE III
OPERATOR CHARACTERISTICS BY SUBJECT ROLE

	Pilots		Masters		Mates		Total	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
Age	51.84	9.55	49.36	10.99	34.08	9.75	41.26	12.87
Years' Experience	23.22	9.55	20.48	11.59	7.19	6.19	13.56	11.22
	$Female$	$Male$	$Female$	$Male$	$Female$	$Male$	$Female$	$Male$
Gender	1	18	2	75	16	99	19	192
	Yes	No	Yes	No	Yes	No	Yes	No
Smartphone Use	17	2	76	1	112	3	205	6
Bluetooth Use	13	6	53	24	76	39	141	69
Wearable Use	4	15	23	54	35	80	62	149
	Low	$Casual$	$High$	Low	$Casual$	$High$	Low	$Casual$	$High$	Low	$Casual$	$High$
Technology Use	1	13	5	9	37	31	4	62	49	14	112	85

TABLE IV
TRANSIT REPLICATIONS

	Conventional Display		WARD		
Task	Inbound	Outbound	Inbound	Outbound	Total
Pilots	0	0	9	10	19
Masters	17	15	19	26	77
Mates	19	20	43	33	115
Total	36	35	71	69	211

learning effects from the simulator and during the simulated transit (see Table IV).

F. Analysis

Data for each hypothesis were collected in each voyage transit leg, for each treatment condition (WARD, conventional bridge displays). Where appropriate, analysis of data by subject role was conducted. Hypothesis 1 examined data for $n = 211$ transits: $n = 140$ with WARDs and $n = 71$ with conventional displays, and analytical methods considered this imbalance. To determine whether the unbalanced design impacted the results, additional analysis was undertaken, and a random sample of $n = 71$ WARD subjects was assessed using balanced analytical methods; no differences in the significance of the results were found. Hypotheses 1a and 1c used Kolmogorov–Smirnov (KS) tests as comparisons of nonnormal performance data of trackkeeping and threat avoidance tasks, and F-tests were used to compare performance variability in H1b and H1d. H1e used Hotelling's T^2-tests with permutations to compare nonnormal unbalanced multivariate data measuring the practice of good seamanship. H1f examined two independent measures of SA, using t-tests for each.

Analysis for Hypotheses 2 and 3 followed a different pattern, using multivariate analysis of variance (MANOVA) with Type-II sum of squares to account for unbalanced samples and confirmatory factor analysis (CFA). Due to missing data, records were omitted for 19 WARD users and 11 users of conventional bridge equipment, resulting in an $n = 181$ ($n = 121$ WARD; $n = 60$ conventional) for H2 and H3. Hypothesis 2 considered the effect of multivariate operator characteristics on operator perceptions of stress, workload, self-efficacy, and trust, and the interaction of the technology treatment used. Cosine transformations were used to approximate normal data for self-efficacy responses, and Pillai's trace was used as a robust measure for unbalanced data. Hypothesis 3 used a maximum likelihood CFA estimate to compare covariance matrices among operator perceptions and operator performance and process measures for operators using the WARD and conventional bridge treatment conditions.

Data were analyzed using R version 3.5.0 running on Windows 10. R packages used included "Hotelling," "kSamples," "lavaan," and "stats." G*Power version 3.1.9.2 was used for power calculations.

VI. RESULTS

Operators using WARDs showed significant reductions in trackkeeping variability when compared to operators utilizing conventional displays, a desirable outcome in ship navigation, particularly in a narrow channel. WARD users also showed improved SA for both perception and comprehension, another a desirable outcome (see Table V). Although workload was significantly increased with the WARD, self-efficacy and trust in the WARD information were also increased, indicating that the operators used the technology and trusted it, even in exceptional situations.

Reduced operator trackkeeping variability (H1b) when using the WARD ($F = 0.177, p = 0.004, 1 - \beta = 0.85$) is an important measure of reliable and predictable performance in ship navigation [36], [40], and a technology that increases reliability and dampens performance variability offers advantages to ship operators performing safety-critical tasks [41]. Interestingly, WARD use did not significantly improve trackkeeping performance (H1a; $D = 0.178, p = 0.088$); threat avoidance performance (H1c; $D = 0.179, p = 0.96$); threat avoidance variability (H1d), a measure of how consistently ship operators avoided collision threats with other vessels ($F = 1.15, p = 0.486, 1 - \beta = 0.84$); or the practice of good seamanship (H1e; $T^2 = 5.056, p = 0.081, 1 - \beta = 0.87$) (see Table V). The improvement in trackkeeping performance variability without an accompanying improvement in mean trackkeeping performance, although consistent with previous work examining impacts of visual displays on navigational performance [21], may be related to the limited navigational space available for maneuvering in a tightly constrained channel.

Operator SA (H1f) was improved with WARD use, in both perception (SA-1; $t = 3.302, p = 0.001$) and comprehension (SA-2; $t = 2.893, p = 0.004$) of the environment. SA is crucial to ship navigation [40], enabling an operator to appropriately respond to changes in the operational environment [23]. A link between SA and dampened performance variability would be a novel finding for a system that supports operator mobility,

TABLE V
HYPOTHESIS 1 RESULTS: OPERATOR PERFORMANCE AND SA

Measure	Data Source	Method	Test Statistic	p-value	Finding
H1a Trackkeeping	Simulator XTE	KS-test	$D = 0.178$	$p = 0.088$	Trackkeeping performance was not significantly improved with WARD use.
H1b Trackkeeping variability	Simulator XTE	F-test	$F = 0.177$	$\boldsymbol{p = 0.004}$	Trackkeeping performance variability was significantly improved with WARD use.
H1c Threat avoidance	Simulator CPA	KS-test	$D = 0.179$	$p = 0.096$	Threat avoidance performance was not significantly improved with WARD use.
H1d Threat avoidance variability	Simulator CPA	F-test	$F = 1.15$	$p = 0.486$	Threat avoidance performance variability was not significantly improved with WARD use.
H1e Practice of Good Seamanship	NSAP	Hotelling's T^2-test	$T^2 = 5.056$	$p = 0.081$	The practice of good seamanship was not significantly improved with WARD use.
H1f Perception (SA-1)	NSAP & SAGAT	t-test	$t = 3.302$	$\boldsymbol{p = 0.001}$	SA-1, perception, was significantly improved with WARD use.
H1f Comprehension (SA-2)	NSAP & SAGAT	t-test	$t = 2.893$	$\boldsymbol{p = 0.004}$	SA-2, comprehension, was significantly improved with WARD use.

TABLE VI
H2: MANOVA OF OPERATOR CHARACTERISTICS ($n = 181$)

Operator characteristic	Df	Pillai's trace	F-value	Num Df	Error Df	Pr(>F)
Role	1	0.207	2.921	13	145	8.56e-4
Age	1	0.153	2.024	13	145	0.022
Gender	1	0.101	1.243	13	145	0.255
Sea Days	1	0.052	0.607	13	145	0.845
Experience	1	0.072	0.869	13	145	0.587
Technology	1	0.106	1.321	13	145	0.207
WARD	1	0.168	2.257	13	145	0.009
Role:WARD	1	0.106	1.321	13	145	0.207
Age:WARD	1	0.225	3.242	13	145	2.53e-4
Gender:WARD	1	0.129	1.649	13	145	0.078
Sea Days:WARD	1	0.1311	1.683	13	145	0.071
Experience:WARD	1	0.061	0.723	13	145	0.739
Technology:WARD	1	0.063	0.756	13	145	0.705
Residuals	157					

perhaps suggesting a fit between the WARD technology and operational requirements.

Operators reported significantly increased self-efficacy, workload, and propensity to trust when using the WARD, based on a MANOVA (Table VI, $1 - \beta = 0.95$). Self-efficacy (H2a), measured as an operator's confidence in their trackkeeping ability, was significantly improved ($F = 2.593, p = 0.003$) with WARD use, as was operator propensity to trust the information displayed in the WARD (H2d; $F = 2.029, p = 0.022$). At the same time, operators reported significantly increased workload (H2b), specifically higher effort ($F = 1.895, p = 0.034$) and frustration ($F = 2.217, p = 0.012$), and did not report significantly reduced stress (H2c) when using the WARD ($F = 1.434, p = 0.149$) (see Table VII), which could suggest that increased workload may be incurred by operators using WARDs.

Significant role-related results were observed as Pilots showed better mean trackkeeping performance than either Masters ($t = 2.993, p = 0.004, 1 - \beta = 0.75$) or Mates ($t = 3.551, p = 9.42 \times 10^{-4}, 1 - \beta = 0.78$) when using the WARD. This result might be expected, as the primary role of a Pilot is to provide trackkeeping and expert knowledge in a particular waterway [39]. This suggests that Pilots' improved performance with WARDs might be attributed to their role rather than to technology familiarity, a factor without significant differences among the three groups ($F = 0.007, p = 0.933$).

The results of a CFA showed operator perceptions did not significantly impact operator performance or SA (H3, Table VIII). The model showed a weak estimated covariance between operator perceptions, performance, and processes (cov $= -0.33, p = 0.995$), and poor model fit statistics, including poor variance-covariance matrix fit ($\chi^2 = 837.643, df = 168, p = 1.72 \times 10^{-89}$), a Comparative Fit Index of 0.644 (acceptable at ≥ 0.95), a Tucker–Lewis Index of 0.598 (acceptable at ≥ 0.95), a root mean squared error approximation of 0.148 (acceptable at ≤ 0.06), and a standardized root mean square residual of 0.161 (acceptable at ≤ 0.08).

VII. DISCUSSION

In this study, WARDs presenting real-time information to mobile operators improved operator performance variability, SA, self-efficacy, and trust, at the cost of increased workload. At the same time, WARD use did not improve operator trackkeeping, threat avoidance variability, or the practice of good seamanship. Reductions in trackkeeping variability without improvements in mean trackkeeping have been reported previously, possibly indicating that real-time WARD information helped operators without impacting trackkeeping and threat avoidance performance that was already within acceptable safety limits. Trackkeeping and threat avoidance may conflict when the limits of a narrow channel impede the ability to maintain distance from a passing vessel; acceptable mean CPAs (≥ 0.1 nmi) observed in this work suggest this conflict was not a consideration in this scenario.

WARD use also significantly improved SA, an important finding affirmed by previous work [18] set in aviation or vehicle driving tasks where operators are stationary. To maintain SA in ship navigation, operators must move about the bridge and the vessel, gathering information from dispersed displays and the environment. The SA improvements observed suggest that the WARD aided users in perception and comprehension of information presented in a mobile form factor while performing tasks requiring mobility, obviating the need to return to fixed displays. Visual displays have previously been observed to improve SA [18], but this is a novel result for mobile wearable AR displays.

Operators using WARDs also reported improved perceptions of self-efficacy and propensity to trust, even while reporting increased workload. Improved self-efficacy and trust have been associated with improved performance [25], and increased workload has been associated with improved SA [6]. These results underscore the need to examine performance, processes, and perceptions together [24]. Improved self-efficacy suggests that operators found the WARDs and the information they presented

TABLE VII
HYPOTHESIS 2 RESULTS: OPERATOR CHARACTERISTICS AND PERCEPTIONS

Measure	Data Source	Method	Test Statistic	p-value	Finding
H2a Self efficacy	NGSE	MANOVA	$F = 2.593$	$\mathbf{p = 0.003}$	Self efficacy was significantly improved with WARD use.
H2b Workload: Effort	NASA-TLX-5	MANOVA	$F = 1.895$	$\mathbf{p = 0.034}$	Workload: effort was significantly increased with WARD use.
H2b Workload: Frustration	NASA-TLX-6	MANOVA	$F = 2.217$	$\mathbf{p = 0.012}$	Workload: frustration was significantly increased with WARD use.
H2c Stress	SSSQ	MANOVA	$F = 1.434$	$p = 0.149$	Stress was not significantly improved with WARD use.
H2d Propensity to Trust	PTS	MANOVA	$F = 2.029$	$\mathbf{p = 0.022}$	Propensity to trust was significantly improved with WARD use.

TABLE VIII
HYPOTHESIS 3 RESULTS: OPERATOR PERCEPTIONS AND PERFORMANCE

Measure	Data Source	Method	Test Statistic	p-value	Finding
H3 Covariance	Performance: Simulator XTE & CPA; NSAP Processes: NSAP Perceptions: NGSE, SSSQ, NASA-TLX & PTS	CFA	$\beta = -0.33$	$p = 0.995$	Perceptions of self efficacy, workload, stress, and trust did not significantly covary with Performance and SA.

useful [28] and valuable to the task [47]. Increased workload suggests that operators devoted resources to understanding WARD information [48], indicating that WARDs were used and not merely worn. High propensity to trust was observed in WARD users, indicating subjects with varied experience were willing to trust WARDs in this setting. As with SA, self-efficacy and trust have been associated with improved performance, and these are regarded as positive effects.

Although no operator characteristics had direct affects on operator performance, links were observed between operator characteristics and perceptions. Older subjects reported higher workload and stress increases with WARDs, an intuitive expectation for which empirical evidence now exists. Older WARD users also reported higher stress, a finding supported by earlier work [33]. These results indicate that younger operators with higher propensity to trust may benefit more from WARDs than older users showing less, but still improved, propensity to trust.

Impacts of operator perceptions on performance and SA were not confirmed by the CFA estimate (H3). Together with evidence from H2, this could imply that these relationships are overshadowed by more direct WARD impacts on operator perceptions. Weak estimated covariance and poor fit statistics indicate that the observed variance-covariance matrix may contain relationships unexplored in this CFA, suggesting additional relationships among operator perceptions [26] and operator characteristics [29]. These could be evaluated using residual correlations in future analyses. These results support previous work concluding that relationships among operator perceptions, and their impacts on performance are indirect and require further study [24], [28].

VIII. CONCLUSIONS, FUTURE WORK

This article examined WARD impacts on operators in the safety-critical system of marine transportation, showing mixed results that align with previous studies of visual displays in similar tasks [6], [18], but showing novel empirical results for wearable AR displays. The improvements observed in WARD users may relate to the novel capabilities of the technology to present valuable task-related information to an operator [22] in real-time, from anywhere in the work space, improving operator SA during tasks that required mobility. The observed SA improvement may have important implications in other safety-critical systems that require mobility, such as nuclear power and emergency response [1], [3]. The negative impact of increased workload with WARD use may be the cost of significantly improved performance variability and SA, two highly desirable outcomes in this safety-critical system [38].

Because the focus of this initial study was the impact of WARDs on operator performance and SA, WARD influences on perceptions were not studied as distinct variables. Given the mixed results of this study, further exploration of each variable in Fig. 2 could provide more insight into the impact of WARDs on operator performance, processes, and perceptions.

Acknowledged links among operator perceptions of workload and SA, as well as between SA and performance, have been identified as salient areas of research [34]. Other work has identified trust in technology as an important emerging area of study with links to performance and perceptions of workload and technology use [28]. An improved understanding of the relationships among operator perceptions of tasks and technology, and among operator performance and processes, would be beneficial to designers and operators of novel technologies such as WARDs.

ACKNOWLEDGMENT

The authors would like to thank Capt. R. Becker, Capt. G. Paine, Capt. E. Friend, Capt. S. Conway, LCDR A. Birch (USCG, Ret.), Capt. M. Cosenza (USCG, Ret.), Capt. B. Kimball, C. Gianelloni, H. Cheong, C. Schaffer, and R. Weiner of the Maritime Institute of Technology and Graduate Studies—Pacific Maritime Institute; A. St. Cerny and G. Ferrando of Transas; and Prof. S. Dunn and T. Willemain of Rensselaer Polytechnic Institute for their assistance with this work.

REFERENCES

[1] K. J. Vicente, E. M. Roth, and R. J. Mumaw, "How do operators monitor a complex, dynamic work domain? The impact of control room technology," *Int. J. Human-Comput. Stud.*, vol. 54, no. 6, pp. 831–856, 2001.

[2] J. Stewart and M. Billinghurst, "A wearable navigation display can improve attentiveness to the surgical field," *Int. J. Comput. Assisted Radiol. Surg.*, vol. 11, no. 6, pp. 1193–1200, 2016.

[3] C. Reuter, T. Ludwig, and V. Pipek, "Ad hoc participation in situation assessment: Supporting mobile collaboration in emergencies," *ACM Trans. Comput.-Human Interact.*, vol. 21, no. 5, 2014, Art. no. 26.

[4] C.-Y. Shen, D.-N. Yang, and M.-S. Chen, "Collaborative and distributed search system with mobile devices," *IEEE Trans. Mobile Comput.*, vol. 11, no. 10, pp. 1478–1493, Oct. 2012.

[5] D. Blauhut and K. L. Seip, "An empirical study of mobile-device use at Norwegian oil and gas processing plants," *Cognition, Technol. Work*, vol. 20, pp. 325–336, 2018.

[6] T. C. Hong, H. S. Y. Andrew, and C. W. L. Kenny, "Assessing the situation awareness of operators using maritime augmented reality system (MARS)," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, 2015, vol. 59, no. 1, pp. 1722–1726.

[7] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[8] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[9] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators Virtual Environ.*, vol. 6, no. 4, pp. 355–385, 1997.

[10] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht, "Towards pervasive augmented reality: Context-awareness in augmented reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 6, pp. 1706–1724, Jun. 2017.

[11] D. L. Goodhue and R. L. Thompson, "Task-technology fit and individual performance," *Manage. Inf. Syst. Quart.*, vol. 19, no. 2, pp. 213–236, 1995.

[12] J. Baumeister *et al.*, "Cognitive cost of using augmented reality displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 11, pp. 2378–2388, Nov. 2017.

[13] F. Biocca, C. Owen, A. Tang, and C. Bohil, "Attention issues in spatial information systems: Directing mobile users' visual attention using augmented reality," *J. Manage. Inf. Syst.*, vol. 23, no. 4, pp. 163–184, 2007.

[14] N. M. Moacdieh and N. Sarter, "The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 886–895, Dec. 2017.

[15] B. J. Dixon, M. J. Daly, H. Chan, A. D. Vescan, I. J. Witterick, and J. C. Irish, "Surgeons blinded by enhanced navigation: The effect of augmented reality on attention," *Surgical Endoscopy*, vol. 27, no. 2, pp. 454–461, 2013.

[16] M. Yeh and C. D. Wickens, "Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration," *Human Factors*, vol. 43, no. 3, pp. 355–365, 2001.

[17] J. Gabbard, D. G. Mehra, and J. E. Swan II, "Effects of AR display context switching and focal distance switching on human performance," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 6, pp. 2228–2241, Jun. 1, 2019, doi: 10.1109/TVCG.2018.2832633.

[18] N. A. Stanton, K. L. Plant, A. P. Roberts, and C. K. Allison, "Use of highways in the sky and a virtual pad for landing head up display symbology to enable improved helicopter pilots' situation awareness and workload in degraded visual conditions," *Ergonomics*, vol. 62, pp. 255–267, 2017, doi: 10.1080/00140139.2017.1414301.

[19] K. Chen, Z. Li, and G. A. Jamieson, "Influence of information layout on diagnosis performance," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 3, pp. 316–323, Jun. 2018.

[20] S. W. Kortschot, G. A. Jamieson, and C. Wheeler, "Efficacy of group-view displays in nuclear control rooms," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 4, pp. 408–414, Aug. 2018.

[21] P. M. Ververs and C. D. Wickens, "Head-up displays: Effect of clutter, display intensity, and display location on pilot performance," *Int. J. Aviation Psychol.*, vol. 8, no. 4, pp. 377–403, 1998.

[22] C. D. Wickens, "Multiple resources and mental workload," *Human Factors*, vol. 50, no. 3, pp. 449–455, 2008.

[23] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 65–84, 1995.

[24] H. Mansikka, K. Virtanen, and D. Harris, "Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 1, pp. 1–9, Feb. 2019.

[25] R. Agarwal, V. Sambamurthy, and R. M. Stair, "The evolving relationship between general and specific computer self-efficacy–an empirical assessment," *Inf. Syst. Res.*, vol. 11, no. 4, pp. 418–430, 2000.

[26] T.-C. Lin and C.-C. Huang, "Understanding knowledge management system usage antecedents: An integration of social cognitive theory and task technology fit," *Inf. Manage.*, vol. 45, no. 6, pp. 410–417, 2008.

[27] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human Factors*, vol. 55, no. 3, pp. 520–534, 2013.

[28] F. Ekman, M. Johansson, and J. Sochor, "Creating appropriate trust in automated vehicle systems: A framework for HMI design," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 1, pp. 95–101, Feb. 2018.

[29] M. A. Rupp, J. R. Michaelis, D. S. McConnell, and J. A. Smither, "The role of individual differences on perceptions of wearable fitness device trust, usability, and motivational impact," *Appl. Ergonom.*, vol. 70, pp. 77–87, 2018.

[30] W. Willett, Y. Jansen, and P. Dragicevic, "Embedded data representations," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 461–470, Jan. 2017.

[31] I. Benbasat and H. Barki, "Quo vadis TAM?," *J. Assoc. Inf. Syst.*, vol. 8, no. 4, pp. 211–218, 2007.

[32] P. A. Hancock, "A dynamic model of stress and sustained attention," *Human Factors*, vol. 31, no. 5, pp. 519–537, 1989.

[33] Q. Shu, Q. Tu, and K. Wang, "The impact of computer self-efficacy and technology dependence on computer-related technostress: A social cognitive theory perspective," *Int. J. Human-Comput. Interact.*, vol. 27, no. 10, pp. 923–939, 2011.

[34] M. A. Vidulich and P. S. Tsang, "The confluence of situation awareness and mental workload for adaptable human– machine systems," *J. Cogn. Eng. Decis. Making*, vol. 9, no. 1, pp. 95–97, 2015.

[35] International Maritime Organization, "International Regulations for Preventing Collisions at Sea (COLREGs)," 1972. [Online]. Available: https://www.navcen.uscg.gov/pdf/navRules/navrules.pdf. Accessed on: Sep. 19, 2019.

[36] N. Bowditch, *The American Practical Navigator*. Bethesda, MD, USA: Nat. Imagery and Mapping Center, 2002.

[37] J. D. Lee and T. F. Sanquist, "Augmenting the operator function model with cognitive operations: Assessing the cognitive demands of technological innovation in ship navigation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 3, pp. 273–285, May 2000.

[38] J. Sauer, D. G. Wastell, G. R. J. Hockey, C. M. Crawshaw, M. Ishak, and J. C. Downing, "Effects of display design on performance in a simulated ship navigation environment," *Ergonomics*, vol. 45, no. 5, pp. 329–347, 2002.

[39] National Research Council, *Minding the Helm: Marine Navigation and Piloting*. Washington, DC, USA: National Academy Press, 1994.

[40] K. S. Gould, B. K. Røed, E.-R. Saus, V. F. Koefoed, R. S. Bridger, and B. E. Moen, "Effects of navigation method on workload and performance in simulated high-speed ship navigation," *Appl. Ergonom.*, vol. 40, no. 1, pp. 103–114, 2009.

[41] M. Grabowski and S. D. Sanborn, "Human performance and embedded intelligent technology in safety-critical systems," *Int. J. Human-Comput. Stud.*, vol. 58, no. 6, pp. 637–670, 2003.

[42] Marine Accident Investigation Branch, "Collision between Huayang Endeavour and Seafrontier approximately 5 nm west of Sandettie Bank, English Channel 1 July 2017," United Kingdom Dept. Transport, London, U.K., Tech. Rep., 2018. [Online]. Available: https://assets.publishing.service.gov.uk/media/5ad86d01e5274a76c13dfdc1 /MAIBInvReport07_2018.pdf. Accessed on: Sep. 19, 2019.

[43] National Transportation Safety Board, "Allision of the Liberian Freighter Bright Field with the Poydras Street Wharf, Riverwalk Marketplace, and New Orleans Hilton Hotel in New Orleans, LO, USA, Dec. 14, 1996. Marine Accident Report NTSB/MAR-98/01," United States Dept. Transport., Tech. Rep. Marine Accident NTSB/MAR-98/01, 1998. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/MAR9801.pdf. Accessed on: Sep. 19, 2019.

[44] National Transportation Safety Board, "Grounding of the U.S. Tankship Exxon Valdez on Bligh Reef, Prince William Sound near Valdez, Alaska March 24, 1989. Marine Accident Report NTSB/MAR-90/04," United States Dept. of Transportation, Tech. Rep. Marine Accident NTSB/MAR-90/04, 1990. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/MAR9004.pdf. Accessed on: Sep. 19, 2019.

[45] L. Orlandi and B. Brooks, "Measuring mental workload and physiological reactions in marine pilots: Building bridges towards redlines of performance," *Appl. Ergonom.*, vol. 69, pp. 74–92, 2018.

[46] D. Murphy, "The NSAP program at PMI," 2015. [Online]. Available: https://maritime-executive.com/features/the-nsap-program-at-pmi. Accessed on: Sep. 19, 2019.

[47] G. Chen, S. M. Gully, and D. Eden, "Validation of a new general self-efficacy scale," *Org. Res. Methods*, vol. 4, no. 1, pp. 62–83, 2001.

[48] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.

[49] W. S. Helton and K. Näswall, "Short stress state questionnaire," *Eur. J. Psychological Assessment*, vol. 31, no. 1, pp. 20–30, 2015.

Modeling Human Pilot Behavior for Aircraft With a Smart Inceptor

Shuting Xu, Wenqian Tan, and Xiangju Qu

***Abstract*—This article presents a model of the human pilot for an aircraft with a smart inceptor, as a means to mitigate human–vehicle system loss–of–control. A key feature of the pilot model is the capability to reflect the interaction between the human–vehicle interface and the human pilot. The proposed human pilot model is primarily composed of a perception module, adaptation module, and execution module. The visual and tactile cues perceived from the human–vehicle interface including a smart inceptor are utilized to model the perception module. The adaptation module, including the central nervous system, is introduced to describe the adaptive behavior of the human pilot. In addition, the execution module contains the neuromuscular system of the human pilot so that the dynamics of the neuromuscular system and the smart inceptor are considered. The human pilot model is assessed by the time domain and wavelet–based analysis in comparison with the experiments of pilot–in–the–loop flight simulation. The simulation results indicate that the tracking results of the time histories given by this model are in good agreement with the behavioral characteristics of a test pilot. The potential applications of the human pilot model include the design of the smart inceptor and evaluation of aircraft loss–of–control events.**

***Index Terms*—Flight simulation, human control model, human–vehicle system, interface, manual control.**

I. Introduction

AIRCRAFT loss–of–control remains a key contributing factor to fatal airline accidents [1]. One of the primary reasons for loss–of–control events is unfavorable interaction of pilot–aircraft systems, which is associated with the human pilot's attempt to strictly control the aircraft, often in response to some triggering event in the environment (e.g., turbulence or severe crosswinds) or changing aircraft dynamic responses (e.g., flight control system failures or unexpected transitions). Due to the unpredictability, suddenness and time urgency of such failure conditions, the information obtained by the human pilot in an emergency may be incomplete and incomprehensive at the beginning period of the trigger event, which may lead to the problem of unfavorable pilot–aircraft system interactions. Pilot–aircraft system loss–of–control in the form of pilot–induced oscillations (PIO) has long been a persistent aviation safety problem [2].

As discussed in [3]–[15], research into the human pilot control behavior in pilot–aircraft system loss–of–control events has proceeded rapidly. In the case of sudden changes in aircraft dynamics, Hess developed a pursuit tracking model of the human pilot and clarified the detection of time–varying vehicle dynamics by a human pilot [9], [28]. The advantage of this model was that the sudden changes were rejected. It is worth mentioning that the outline of this model was derived from the pursuit human pilot model presented in [10] and [11]. Zaal used a maximum likelihood estimation procedure to investigate the adaptation of human pilot control behavior to time–varying aircraft dynamics [12]; the method provided accurate parameter estimations. In the presence of flight control anomalies, [13] used the adaptive human pilot model proposed in [9] and applied it for the assessment of the flight control system subjected to system failures. Hess [14] expanded the model's applicability to accommodate significant and sudden variations in elements of the flight control system. In the investigation of boundary–triggered pilot–induced oscillations, a human pilot model of [15] was developed for use in examining boundary avoidance tracking. However, the human pilot models abovementioned failed to take the pilot–aircraft interaction into account.

Currently, regarding the problem of pilot–aircraft system loss–of–control, Klyde *et al.* [1] proposed a smart adaptive flight effective cue (SAFE-Cue system), namely a smart inceptor implemented in the aircraft. As a pilot–aircraft interaction interface, the smart inceptor can be used to eliminate the tendencies of pilot–aircraft system oscillations. The SAFE-Cue system provides force feedback to the human pilot via a smart inceptor based on a system error between the actual response and a nominal system response. The SAFE-Cue alerts the human pilot in the presence of damage or failures and provides guidance via force feedback cues, to ensure the stability and performance of a pilot–aircraft system.

The previous investigation of the SAFE-Cue system involved evaluation using the experiments of pilot–in–the–loop flight simulation [2]. It is worth noting that the SAFE-Cue system allowed the test pilot to focus on the current task rather than simply maintaining control. While the experimental results are not standardized and vary with individuals, a theoretical human

Manuscript received May 14, 2018; revised December 15, 2018; accepted September 7, 2019. Date of publication November 14, 2019; date of current version November 21, 2019. This work was supported in part by the National Natural Science Foundation of China with the project reference number of 11502008 and in part by the Aeronautical Science Foundation of China with the project reference number of 2017ZA51002. This article was recommended by Associate Editor Z.-H. Mao. *(Corresponding author: Wenqian Tan.)*

The authors are with the School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China (e-mail: xushuting@buaa.edu.cn; tanwenqian@buaa.edu.cn; uq@buaa.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2944376

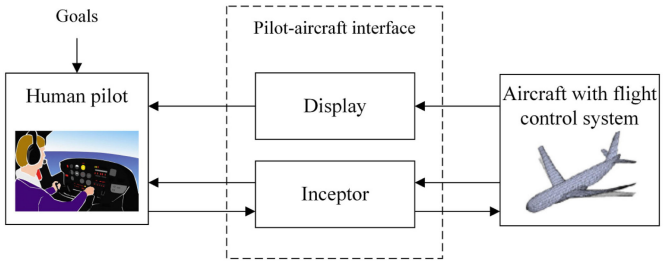


Fig. 1. Outline of pilot–aircraft system.

pilot model is valuable in view of its uniformity and standardization. The model can be made more objective to test the effect of the smart inceptor.

In this article, we developed a human pilot model for an aircraft with a smart inceptor. The aim of this study is to predict the characteristics of a pilot–aircraft system including a smart inceptor for mitigating loss–of–control. A key feature of the pilot model is its capability to reflect the interaction between the pilot behavior and the smart inceptor. To simulate the actions of a real human pilot, the behavioral mechanism is considered before establishing a suitable human pilot model. The perception module, adaptation module, and execution module are introduced into this model. In the perception module, not only the visual cues provided by the display interface, but also the tactile cues obtained by the smart inceptor are utilized to analyze the perception behavior of the human pilot. In the adaptation module, the adaptive parameters can be changed based on the force feedback from the smart inceptor. In the execution module, the neuromuscular system can help gain a better understanding of the interaction between the smart inceptor and the human pilot control behavior. The neuromuscular model is based partly on models of the neuromuscular system available in the literature. These modules provide a feasible way to explore the human pilot model of an aircraft with a smart inceptor for mitigating loss–of–control and are a major contribution of this article.

The remaining article is structured as follows. The outline of the pilot–aircraft system is presented first. Next, the principle of modeling the human pilot is proposed, and three modules are described in detail. Subsequently, the model of the aircraft system with the SAFE-Cue system is established. This is followed by a description of the experimental design and presentation of the results. Finally, the discussion and conclusions are presented.

II. Pilot–Aircraft System

The pilot–aircraft system consists of the human pilot, the aircraft system (including flight control system) and the pilot–aircraft interface. These systems interact and coordinate to complete a given task. Fig. 1 provides a simplified representation of the human pilot activity in the aircraft flight control. As shown, the visual perception and tactile feedback are considered in this article on account of the pilot–aircraft interface. To simplify the pilot model, the motion feedback of the vehicular is neglected. Not only does the human pilot perceive the flight information from the display interface, but he/she also obtains force feedback cues via a smart inceptor, such as the SAFE-Cue system. Next, the human pilot drives the smart inceptor to perform the control. As a pilot–aircraft interface, the smart inceptor is a primary means through which the human pilot passes commands to the flight control system. It is also an essential method to provide guidance via force feedback cues for the human pilot.

In this article, the human pilot behavior is separated into three elements: perception module, adaptation module, and execution module, as shown in Fig. 2. The system is motivated by an error that is the difference between the command signal and the actual output to be displayed in the flight display window. The display delivers the information to the visual perception modality. The force feedback cues from the SAFE-Cue system are perceived by the human pilot via the tactile perception. The information is processed by the central nervous system to drive a process through which the desired control action can be generated. The control action is then executed via the forces generated by the muscles in the neuromuscular system and exerted on the SAFE-Cue system. The SAFE-Cue system also provides feedback signals through displacements to the neuromuscular system. The output signals from the SAFE-Cue system are used to help the actuation system realize the appropriate movement of the control surfaces. These cues effectively close the feedback control loops. Therefore, the human pilot model components can now be grouped together as the perception, adaptation, and execution modules.

III. Human Pilot Model

A human pilot model is developed considering the interaction between the SAFE-Cue system and the human pilot. The perception module, adaptation module, and execution module of this human pilot model are described in this section.

A. Perception Module

The perception system is the primary source for guidance and control information for the human pilot [16]–[20]. Visual and tactile perceptions are considered in this study on account of the pilot–aircraft interface. The visual cues include the display of the pitch angle, flight path angle, altitude, speed, and outputs of other instruments of the aircraft. Similarly, tactile cues mainly include the feedback force via the SAFE-Cue system. As the human perception is subject to physiological conditions, it is impossible to create a precise unit mapping of the perception information in the brain. The flight test shows that the human pilot cannot perform control actions with small deviations. This indicates that there exists a threshold in the perception mechanism of the human. Therefore, the perception module should reflect the nonlinearity of the threshold.

In the analysis of the human perception, the perception mechanism can be regarded as an interaction module of the neurons of the receptor and the original signal in the neural network. Two modules are developed to represent the visual perception (see Fig. 3) and the tactile perception (see Fig. 4). The input signals include the absolute value of the command and the threshold b. $w(1,1)$ represents the weight of the input signals. Here, the value is 1. When the absolute value of the input signals is less than the

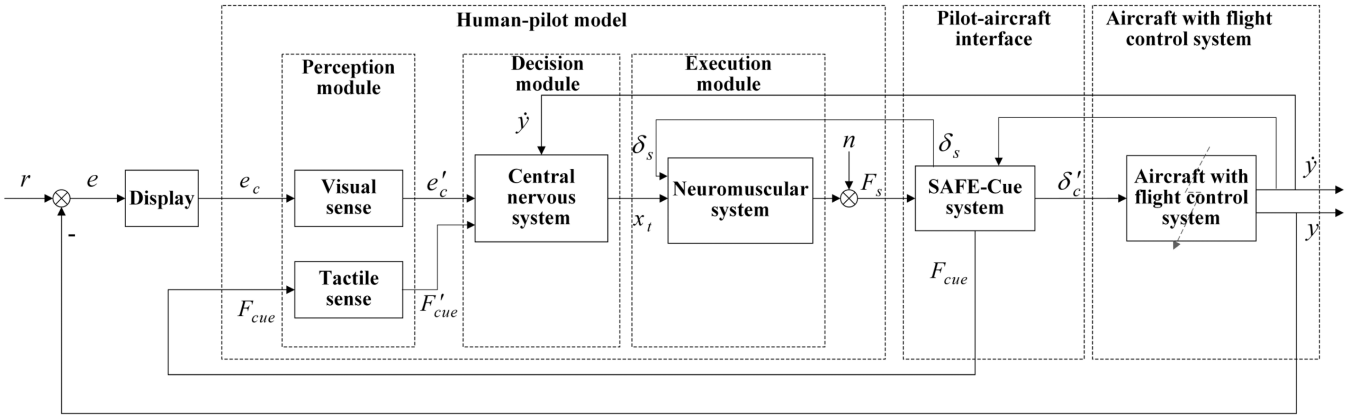


Fig. 2. Block diagram of pilot–aircraft system.

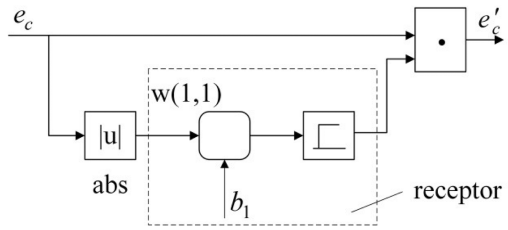


Fig. 3. Visual perception module.

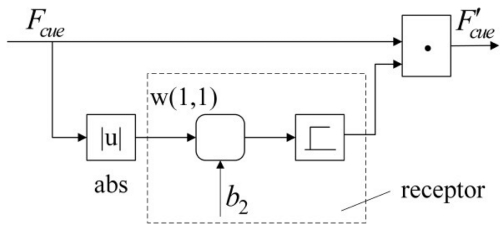


Fig. 4. Tactile perception module.

threshold b, the output of the receptor is 0; otherwise, the output of the receptor is 1. The final output is the product between the receptor output and the command input. Therefore, the effects of the perception module represent the capacity of perception of the human pilot when the threshold is considered.

For the perception module, physiological experiments indicated that with the reasonable illumination and contrast, the eyes of a human could distinguish the black line to 0.01 drad (1 rad = 10 drad) [21]. The investigation of controllable pressure sensitivity showed that the average normal pressure threshold of the human palm is approximately 0.158 g [22]. Therefore, the threshold $b1$ of the visual perception is 0.01 drad, and the threshold $b1$ of the tactile perception is 0.158 g.

B. Adaptation Module

The adaptation module of the human pilot involves activities in which the human pilot needs to accommodate the smart inceptor. The human pilot adaptation behavior involves changing the control action according to the input of the current input information. The human pilot adaptation behavior is shown in Fig. 5.

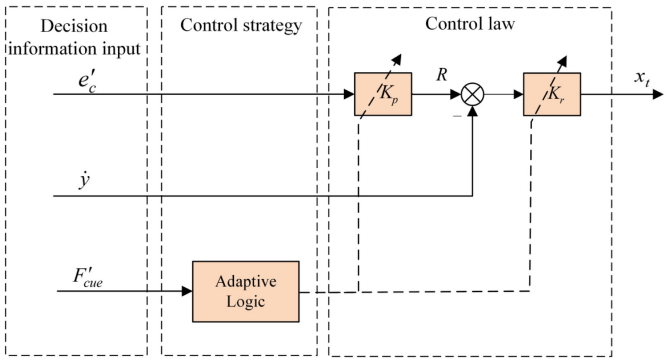


Fig. 5. Human pilot adaptation module.

In this article, according to the adaptive control theory, the input information primarily includes the error information e'_c obtained from the visual perception, the feedback force F'_{cue} from the tactile perception, and the rate information $\dot{y}$ of the aircraft (such as pitch angle rate and bank angle rate).

The human pilot control strategy can be changed based on the force feedback F'_{cue} of the SAFE-Cue system. Herein, we use the Hess adaptive logic [9] as the control strategy. The logic controlling variables Kr and Kp were developed and presented in Hess's study [9] as follows. First, taking the effect of force feedback into account, the criterion signal x is defined as

$$x = F_{\text{cue}}. \tag{1}$$

Here, the criterion signal x is related to the force feedback of the SAFE-Cue system, reflecting the changes of the system error q_{err}. It is the key function of the SAFE-Cue system. The time when the failure occurs is defined as $t = t_c$; the trigger factor K_{trigger} can then be defined as

$$K_{\text{trigger}} = \begin{cases} 0, & \text{if } \sqrt{|x|} < 3 \cdot \text{rms}[\sqrt{|x|}] \quad \text{or} \quad t < t_c \\ 1, & \text{if } \sqrt{|x|} \geq 3 \cdot \text{rms}[\sqrt{|x|}] \quad \text{and} \quad t \geq t_c \end{cases}. \tag{2}$$

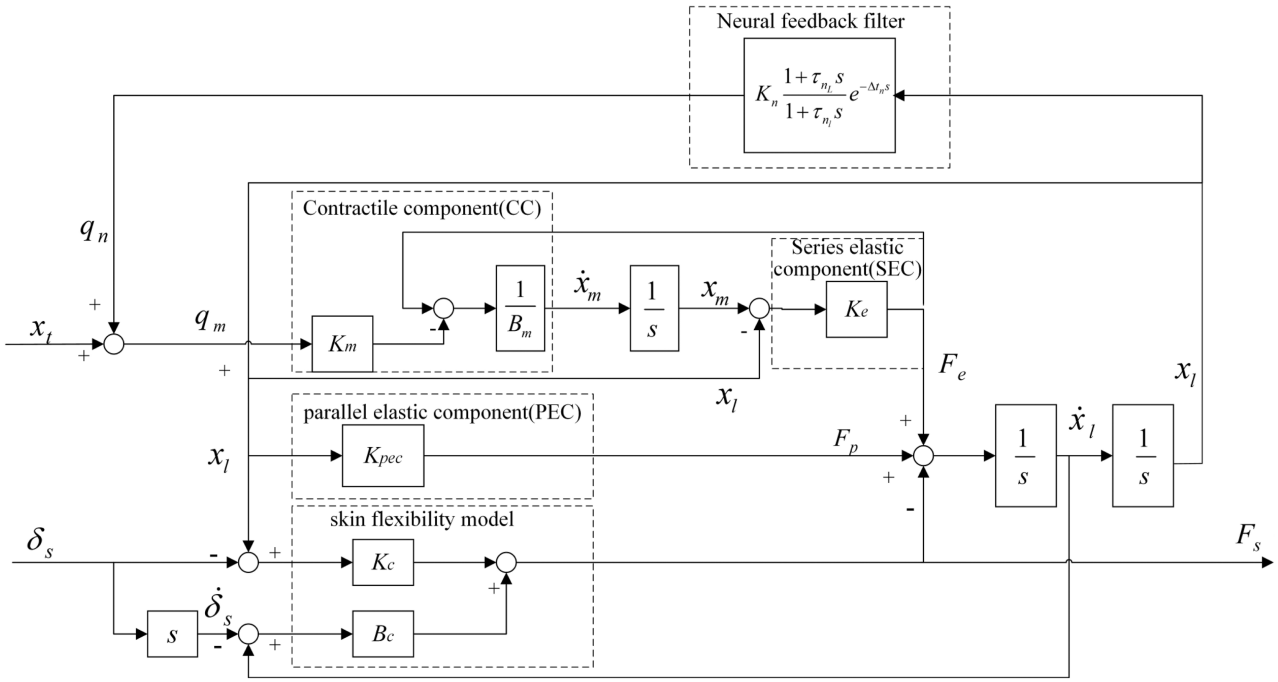


Fig. 6. Neuromuscular system module.

Equation (2) illustrates the trigger principle of the adaptive logic. According to human physiological characteristics, human judgment obeys normal distribution. A factor of 3 was chosen in (2) as it represents an instantaneous "3σ" value. It is apparent that an excessively small trigger value leads to normal system disturbances (e.g., turbulence), which could initiate undesirable human pilot model adaptation, whereas an excessively large value inhibits adaptation when it is required.

For the control law, the changes in K_r and K_p are defined as ΔK_r and ΔK_p, respectively, and formulated as follows:

$$\Delta K_r = x \cdot K_{\text{trigger}} \tag{3}$$

$$\Delta \mathrm{K_p} = 0.35 \cdot \Delta \mathrm{K_r},\ \Delta \mathrm{K_r} > 0$$

$$\Delta \mathrm{K_p} = 0,\ \Delta \mathrm{K_r} \leq 0. \tag{4}$$

Finally, through the above adaptive logic analysis and theoretical evaluations, the human pilot control action can be obtained by the generated output.

C. Execution Module

In the execution module, the effects of tactile feedback depend on the neuromuscular system of the arm and hand as the human pilot is holding the smart inceptor. The model of the neuromuscular system is based partly on models available in the literature [23]–[27]. The proposed model is combined with models for skin flexibility and limb inertia and a neural feedback for the control of the neuromuscular system. The neural feedback enables the human pilot to modify the effective properties. For the SAFE-Cue force feedback system, we design the model for a condition in which the human pilot does not need reflexive feedback paths, which correspond to the neuromuscular system properties as measured in a relaxation task.

Fig. 6 shows the components of the neuromuscular system module. The model contains the skin flexibility model (damping and elasticity components B_c and K_c, respectively), a parallel elastic component (elasticity K_{pec}), a contractile component (CC) (damping and elasticity components B_m and K_m, respectively), the series elastic element (elasticityK_e), and neural feedback. The neuromuscular system parameters adopted herein were obtained from the literature [25], [26], as given in Table I. These parameters describe the intrinsic properties of the neuromuscular system.

In addition, a remnant signal *n* is simulated using a low–pass–filtered white–noise signal to describe the pilot remnant. The noise power is based on the signal-to-noise ratio pertaining to full attention observation, which has a value of 0.01. The low–pass filter is as follows [30]:

$$H_n = \frac{0.2(3.0s+1)}{(1.5s+1)(0.4s+1)}. \tag{5}$$

IV. Aircraft System With SAFE-Cue System

The SAFE-Cue system provides force feedback to the human pilot via an active control inceptor with the corresponding gain

$$\frac{q}{\delta_e} = \frac{-0.04133s(s+0.009561)(s+0.6347)}{(s^2+2\times 0.002782\times 0.06904s+0.06904^2)(s^2+2\times 0.4709\times 1.599s+1.599^2)}. \tag{6}$$

TABLE I
PARAMETER VALUES OF NEUROMUSCULAR SYSTEM

Parameter	K_e $Nm\cdot rad^{-1}$	K_{pec} $Nm\cdot rad^{-1}$	K_c $Nm\cdot rad^{-1}$	I_l $Kg\cdot m^2$	B_m $Nm\cdot rad^{-1}$	B_c $Nm\cdot rad^{-1}$	K_m $Nm\cdot rad^{-1}$	τ_{n_L} s	τ_{n_l} s	Δt_n s	K_n rad^{-1}
Value	50	7	400	0.07	0.45	15	1	0	0.491	0.04	1

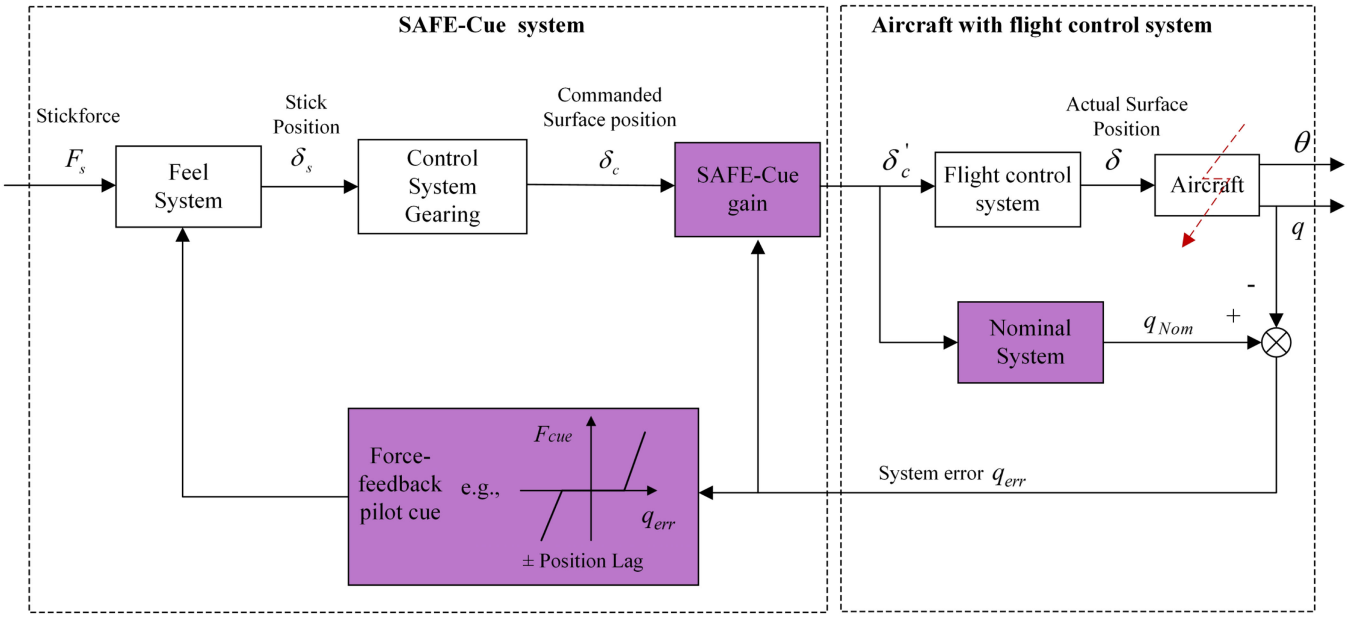


Fig. 7. Block diagram of aircraft system with SAFE-Cue system.

adjustments based on a system error between the actual response and a nominal system response. The SAFE-Cue alerts the human pilot in case of occurrence of damage or failures and provides guidance via force feedback cues, to ensure the pilot–aircraft system stability and performance. The SAFE-Cue system can be applied to any flight control system implementation as a means to mitigate loss–of–control. The block diagram for the aircraft system with the SAFE-Cue system is shown in Fig. 7.

A. SAFE-Cue System Description

The SAFE-Cue system includes the feel system, SAFE-Cue gain and SAFE-Cue force.

The longitudinal feel system dynamics for the inceptor are modeled as those for a second order system. The characteristic parameters of the feel system are given in Table II [29]. The resulting transfer function is as follows:

Specifically, $\frac{\delta_s}{F_s} = \frac{1}{s^2+2\times 0.707\times 20+20^2}$.

The SAFE-Cue gain is a distinct structure that adjusts the gain applied to the inceptor position input. The magnitude of the gain depends on the system error, and it is activated when the system error exceeds a threshold value. The gain is then reduced linearly as a function of the system error until it reaches a prescribed minimum value.

The SAFE-Cue force mechanization takes advantage of the force feedback to form a force cue sent through the inceptor to the human pilot. The force feedback also depends on the system error, and it is activated when the system error exceeds a prescribed threshold. The selected SAFE-Cue gain and SAFE-Cue force mechanizations are shown in Fig. 8.

TABLE II
CHARACTERISTIC PARAMETERS OF THE FEEL SYSTEM

Parameter	Value
Spring Gradient(N/m)	400
Damping($N\cdot s\cdot m^{-1}$)	28.28
Mass(kg)	1
Natural Frequency(rad/s)	20
Damping Ratio	0.707
Control system gearing	1

B. Aircraft System Description

The NASA generic transport model [2] was employed in this study. The selected flight condition involved an altitude of 15 000 ft. and a calibrated airspeed of 260 knots. The controlled element dynamics are as (6) shown at the bottom of previous page.

The flight control system downstream of the SAFE-Cue system is given in Table III. The system includes the rate limit, surface actuator, and surface position limit.

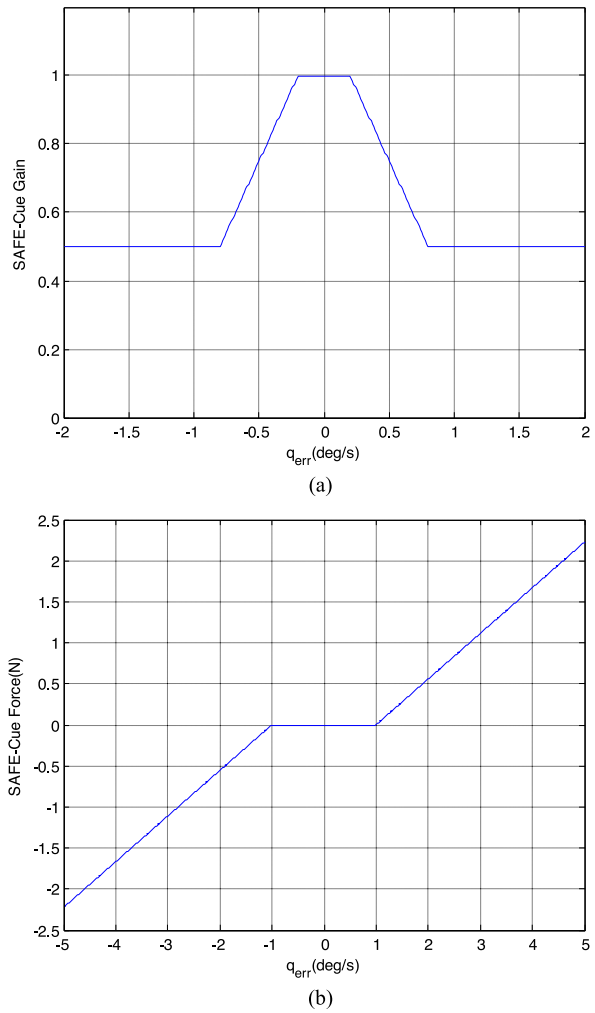
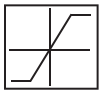


Fig. 8. SAFE-cue gain and SAFE-cue force. (a) SAFE-cue gain mechanization. (b) SAFE-cue force mechanization.

TABLE III
PARAMETERS OF FLIGHT CONTROL SYSTEM

Flight control system	Form	Parameter values
Rate limit(*deg/s*)		$V_L = 15°/s$(failure); $60°/s$(baseline)
Surface actuator	$\frac{\omega_n^2}{s^2+2\xi\omega_n+\omega_n^2}$	$\omega_n = 44\ rad/s$, $\xi = 0.707$
Surface position limit	$\delta_{e_{max}}$	$30°$

V. EXPERIMENTS AND RESULTS

A pilot–in–the–loop flight experiment in pitch was conducted to evaluate the SAFE-Cue system developed by Systems Technology, Inc. (STI) for the NASA VSST Project [2]. As reference, we introduce the results of a comparison between the human pilot model simulation and the pilot–in–the–loop experiment.

TABLE IV
FAILURE SCENARIOS

Element	Value
Time of failure initiation(*s*)	20
Actuator effectiveness	75%
rate limit (*deg/s*)	15

The results of the time domain and wavelet–based analysis are presented in this section.

A. Experiment Description

A team led by Systems Technology, Inc., for the NASA VSST Project [2], developed the pilot–in–the–loop flight experiment in a simulator to assess the feasibility of the SAFE-Cue system. The experimental setup used an STI fixed-base pilot–in–the–loop flight simulator to strengthen the capabilities of STI in the area of real–time flight simulation and pilot–aircraft system identification. The simulator shown is a win32 console application designed to interface with MATLAB for data input and output. The tracking task display was designed to replicate the symbology used in the Calspan Learjet in–flight simulators, which is a phase 2 flight test aircraft. A McFadden Series 292 A two–axis (pitch and roll) fighter stick control loader was integrated with the STI PC-based flight simulator. The McFadden control loader was used previously to successfully develop and evaluate candidate Smart–Cue designs. The McFadden feel system is a key component in that it provides key proprioceptive cues that enhance the fidelity of the simulation.

In the experiment, a pitch tracking task was used to ensure that the human pilot tracked the displayed attitude command and attempted to keep errors within the specified tolerances. The sum–of–sines (SOS) command signals were selected as they forced the human pilot to continuously apply a control input. Failure scenarios were created as a means to evaluate the capabilities of the SAFE-Cue system, as given in Table IV.

In this article, we select the failure scenarios and simulation task that are consistent with the pilot-in-the-loop experiment described in [2]. The details of the comparison results are presented in the following.

B. Time Domain Analysis

Fig. 9 compares the tracking performance of the pitch sum–of–sines histories for the human pilot model simulation and pilot–in–the–loop experiment. In comparison to that in the experiment, the simulation tracking is smoother. The phase lags to exhibit similar tendencies in both the simulation and experiment. It can also be seen that the responses of the simulation and the experiment are consistent with those of the tracking task. These indicate that the human pilot model and actual human pilot are able to approach the desired performance.

The system error signals for the simulation and experimental results are shown in Fig. 10. The system errors remain within approximately $\pm$5deg/s for both the cases, which represents a reasonable range. Fig. 11 shows the curves of K_p and K_r for the

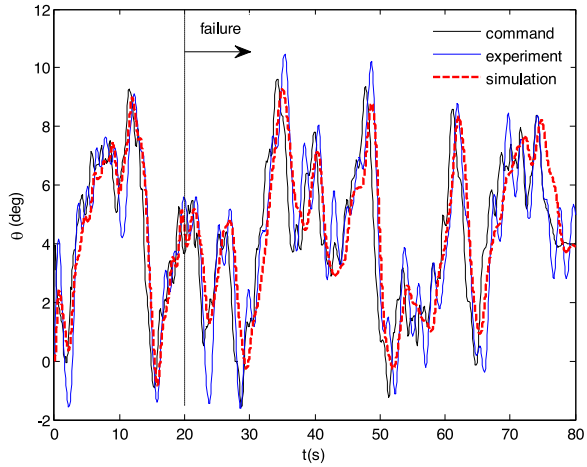


Fig. 9. Time domain tracking performance.

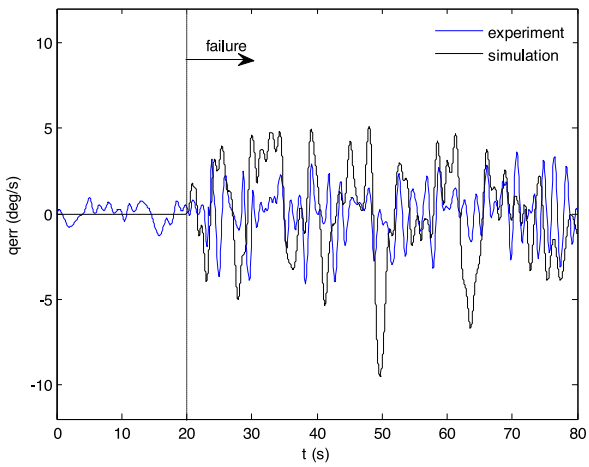


Fig. 10. System error q_{err}.

human pilot model. Due to the force feedback of the SAFE-Cue system, the gain of the human pilot model decreases, which reduces the manipulation, thereby improving the rate–limit effectively.

C. Wavelet–Based Analysis

Fig. 12 shows the results for the wavelet–based analysis of the system error. The scalograms for these cases reflect the difference in the signal activity. Fig. 13 shows the time slices of the system error scalograms for time nodes of 21, 22, 24, 26, and 40 s. The time of 21, 22, 24, and 26 s are those in which the human pilot changed his/her control behavior in the presence of failures. A total of 40 s is the time at which the aircraft achieves stable response. It can be seen that the frequencies corresponding to the peaks of the wavelet–based transforms pertaining to the human pilot model are similar to those of the experiment in the frequency range of 1–3 rad/s, around the region of the crossover frequency. The differences in the peaks are expected to adjust the gains of the human pilot in the human pilot model to better match the experimental results.

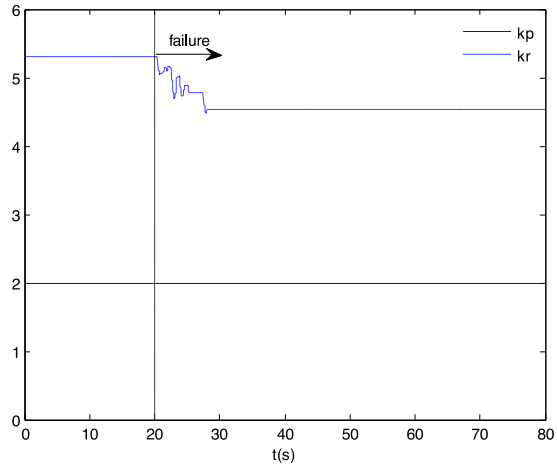


Fig. 11. K_p and K_r.

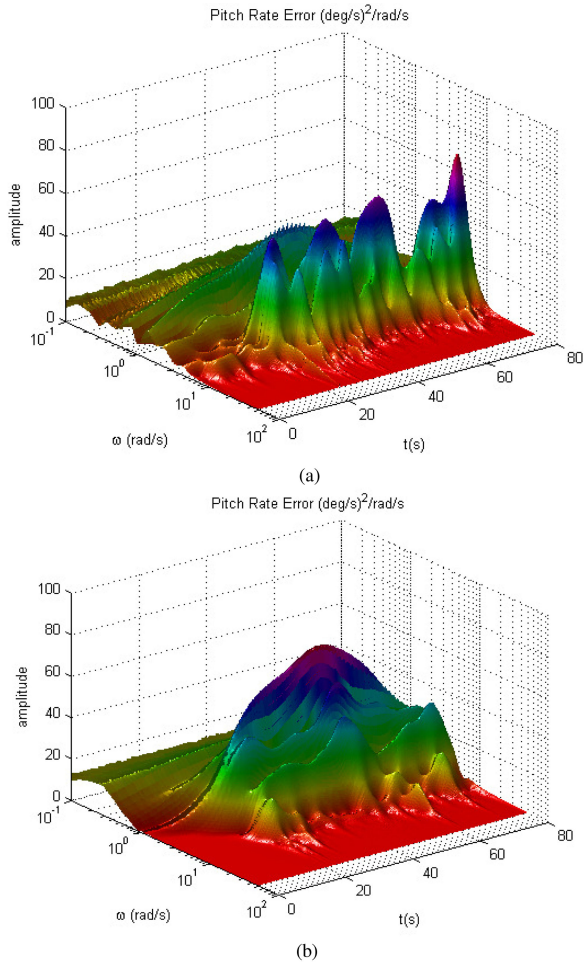


Fig. 12. Scalograms of pitch rate system error. (a) Experimental result. (b) Simulation result.

VI. Application and Validation of the Pilot Model on Calspan Learjet Aircraft

The Calspan Learjet aircraft [1] was selected for the validation of the pilot model. The selected flight condition included an altitude of 15 000 ft. and a calibrated airspeed of 250 kt. The

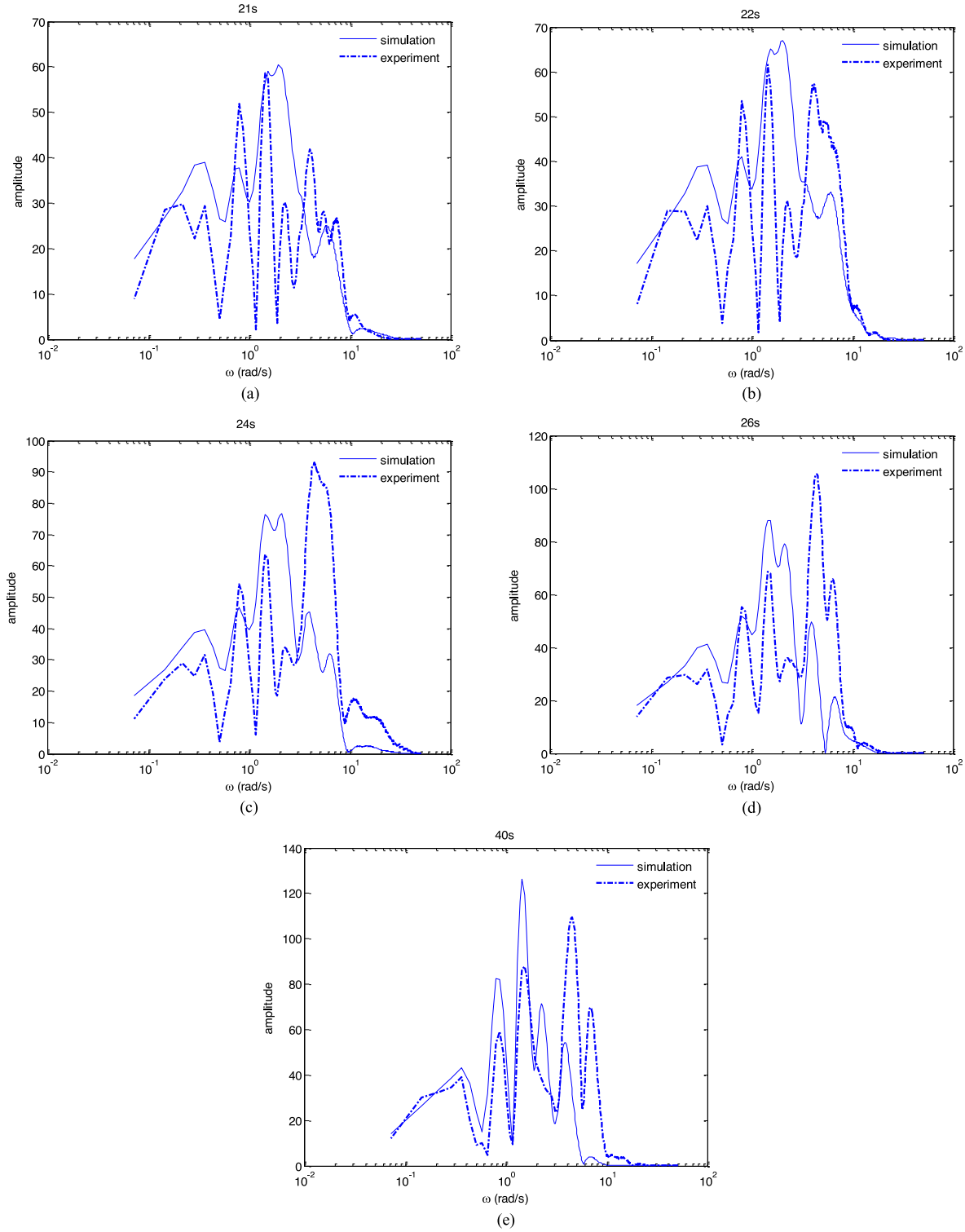


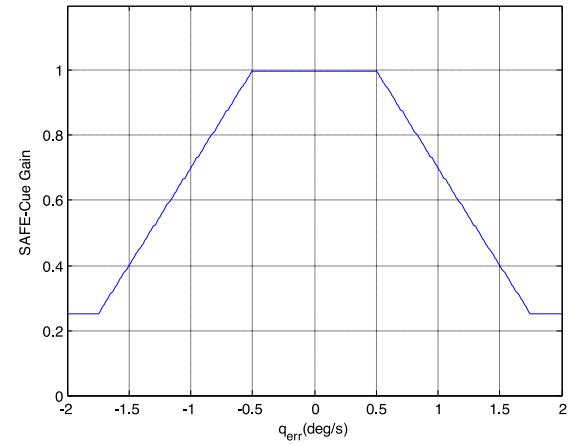
Fig. 13. Scalograms of time slices of system error. (a) 21 s. (b) 22 s. (c) 24 s. (d) 26 s. (e) 40 s.

controlled element dynamics are as follows:

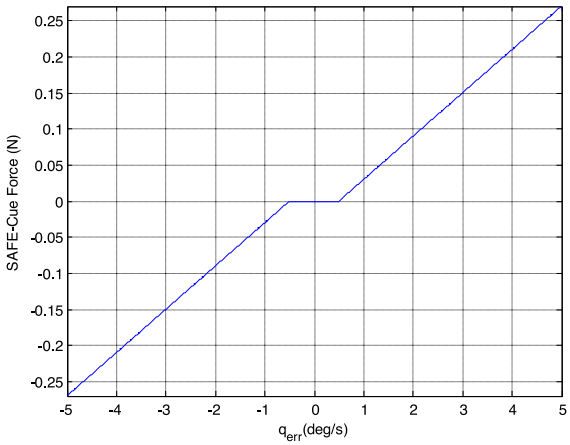
$$\frac{q}{\delta_s} = \frac{211.8(0.5328)(2)^2(7.5)}{(0.7646)(1.184)[0.5149, 2.833][0.7071, 88.61]} e^{-0.13s}. \tag{7}$$

The flight control system is given in Table II. The selected SAFE-Cue gain and SAFE-Cue force mechanizations are shown in Fig. 14.

To validate the human pilot model, the simulation and flight test data results were compared. A flight test conducted by Systems Technology, Inc. [1] was considered. The failures (see



(a)



(b)

Fig. 14. SAFE-Cue gain and SAFE-cue force. (a) SAFE-Cue gain mechanization. (b) SAFE-Cue force mechanization.

Table IV) were introduced approximately 20 s after the start of the pitch-tracking SOS task.

The tracking performance is shown in Fig. 15. It can be seen that the simulation and flight test results can track the command well. The system errors for the simulation and flight test in Fig. 16 are small, within approximately ± 1.5 deg/s. Fig. 17 shows the curves of K_p and K_r for the human pilot model. Due to the signal feedback of smart inceptor, the gain of the pilot model decreases, thereby mitigating the unfavorable characteristics of failure.

The scalograms for the system are shown in Fig. 18. Fig. 19 shows the time slices of the system error scalograms. Similar to in the previous case, the frequencies corresponding to the peaks of the wavelet–based transforms pertaining to the simulation results approach those in the flight test results, around the region of the crossover frequency.

VII. Discussion

The human pilot model herein assumes some fixed and some adaptive parameters. Among them, the parameters of the perception and execution modules are fixed, and they are obtained

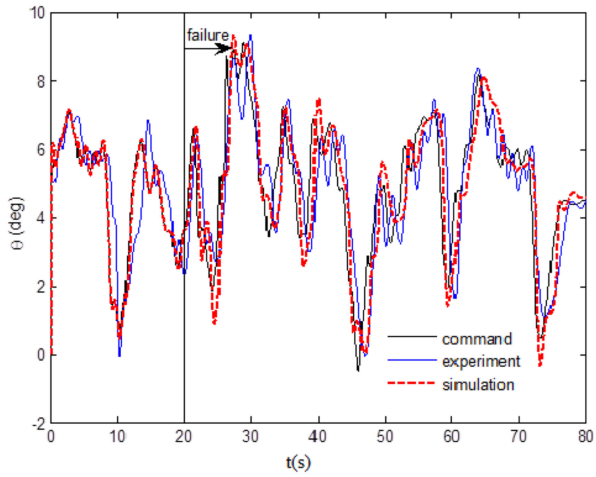


Fig. 15. Time domain tracking performance.

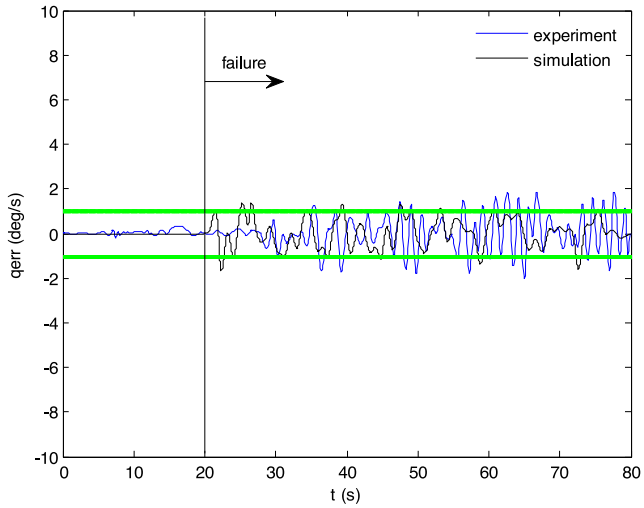


Fig. 16. System error q_{err}.

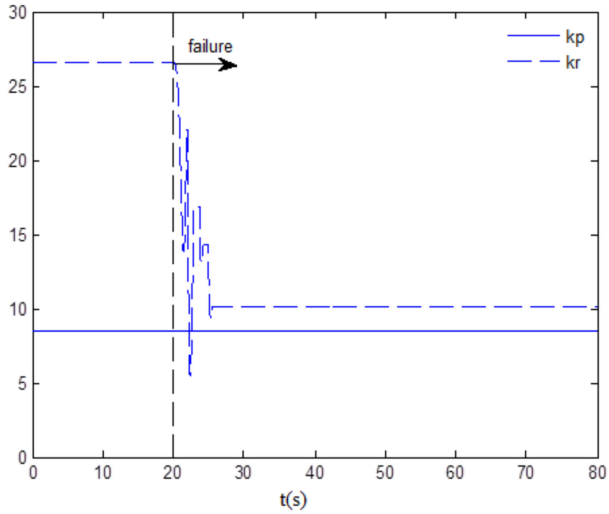


Fig. 17. K_p and K_r.

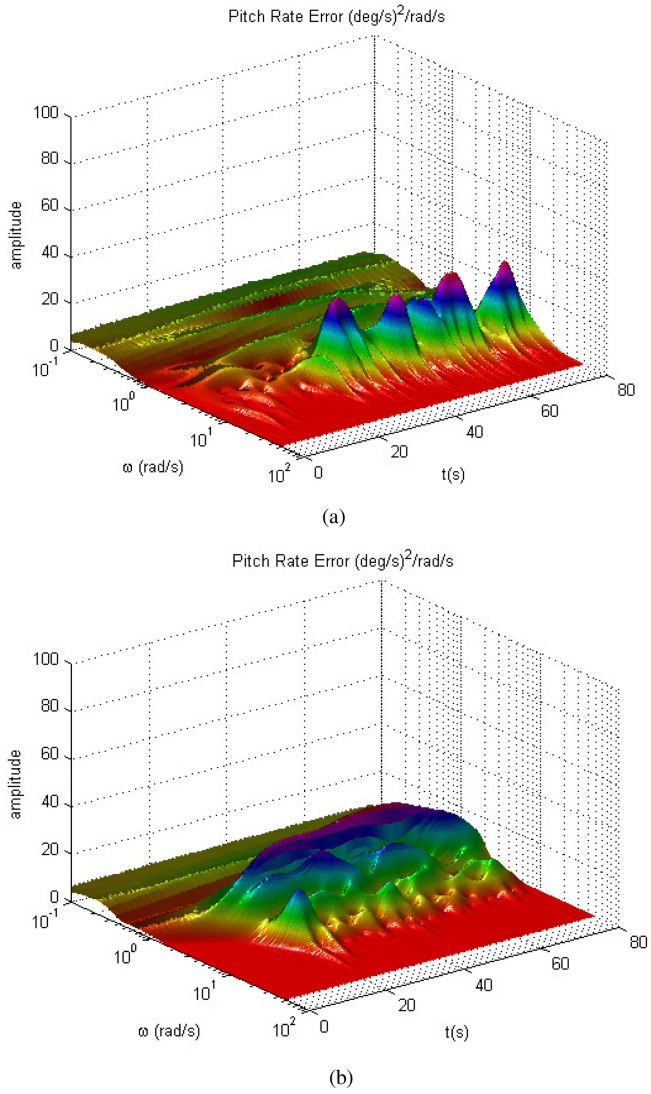


Fig. 18. Scalograms of pitch rate system error. (a) Experimental result. (b) Simulation result.

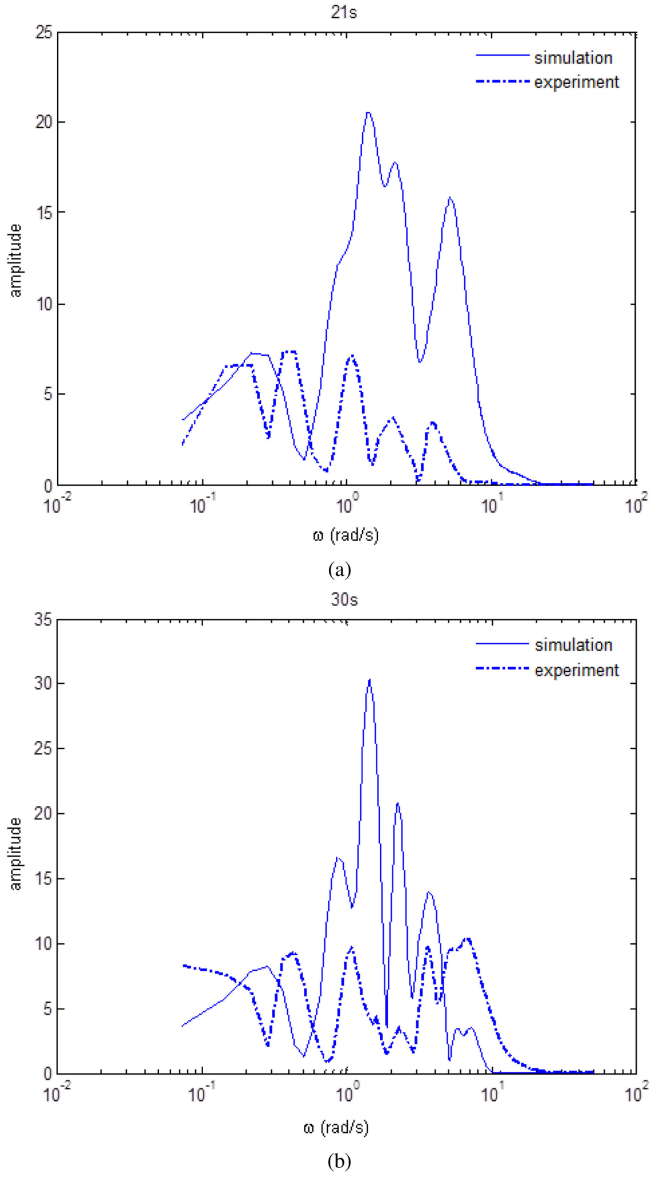


Fig. 19. Scalograms of time slices of system error. (a) 21 s. (b) 30 s.

by conducting experiments. The adaptive parameters K_p and K_r of the adaptation module reflect the adaptation of the human pilot.

Fig. 11 shows the time histories of K^r and K^P. When failure occurs ($t = 20$ s), the human pilot adaptations occur in under 2 s and take almost 8 s to complete. With the adaptation of the human pilot as adjusted by gains K_p and K_r and the smart inceptor, pilot–aircraft system performance is improved.

It must be emphasized that the human pilot model developed herein cannot be considered to be in its unique and final form. However, the framework of the pilot model, which includes the three modules and the connection between the smart inceptor and the human pilot, can be used for other smart inceptors to address additional loss–of–control cases. It is offered as a model of the manner in which a human pilot may adapt to the smart inceptor, although the adaptive parameters of the pilot model can be only suitable for this SAFE-Cue system employed herein. Further, the limitation of this model is that the adaptation of the human pilot is based on the smart inceptor, which indicates that an aircraft using a smart control system implemented with the smart inceptor is required.

VIII. Conclusion

In this article, a human pilot behavior model was developed according to the behavior characteristics of a human pilot of an aircraft with a smart inceptor.

This model is essentially different from some previous human pilot models that fail to take into account the smart inceptor. First, the perception module of the human pilot model considered the tactile perception via the smart inceptor. Second, the adaptation module changed the manipulation strategies based on the smart inceptor force feedback. Additionally, the execution module evaluated the neuromuscular system of the human pilot

so that the correlation of the neuromuscular system and the smart inceptor was considered. Consequently, this human pilot model was able to timely receive and respond to the smart inceptor.

The pilot–aircraft system was established by combining the human pilot model, the smart inceptor, and the aircraft system. The comparison of the simulation and the experimental results of tracking performance demonstrated that the responses are consistent. In addition, the frequencies corresponding to the peaks of the wavelet–based transforms determined by the human pilot model were close to the ones in the experiment around the region of the crossover frequency. The limits of the wavelet–based transforms were the differences of the peaks. It is expected to adjust the human pilot's gain of the human pilot model to better match the experimental results.

However, further investigation of human pilot modeling is needed. First, the experimental result presented in this article was specific to an experienced test pilot; thus, certain subjectivity is inevitable in the simulations. For this purpose, more experienced test pilots need to be introduced into the modeling. Second, the assessment of a wavelet–based PIO metric with the human pilot model needs to be explored further.

REFERENCES

[1] D. H. Klyde, A. K. Lampton, N. D. Richards, and B. Cogan, "Flight–Test Evaluation of a Loss–of–Control Mitigation System," *J. Guid., Control Dyn.*, vol. 40, no. 4, pp. 981–997, 2017.

[2] D. H. Klyde, C. Y. Liang, D. J. Alvarez, N. Richards, R. J. Adams, and B. Cogan, "Mitigating unfavorable pilot interactions with adaptive controllers in the presence of failures/damage," in *Proc. AIAA Atmospheric Flight Mech. Conf. Exhib.*, 2011, pp. 1–17.

[3] S. Xu, W. Tan, A. V. Efremov, L. Sun, and X. Qu, "Review of control models for human Pilot behavior," *Annu. Rev. Control*, vol. 44, no. 1, pp. 274–291, 2017.

[4] S. Xu, W. Tan, L. Sun and X. Qu, "Survey on theory and method of pilot–aircraft system with intelligent control," in *Proc. 3rd IEEE Int. Conf. Control Sci., Syst. Eng.*, 2017, pp. 92–96.

[5] M. Mulder *et al.*, "Manual control cybernetics: State–of–the–Art and current trends," *IEEE Trans. Human–Mach. Syst.*, vol. 48, no. 5, pp. 468–485, Oct. 2018.

[6] J. J. Potter and W. Singhose, "Improving manual tracking of systems with oscillatory dynamics," *IEEE Trans. Human–Mach. Syst.*, vol. 43, no. 1, pp. 46–52, Jan. 2013.

[7] L. R. Young, "On adaptive manual control," *IEEE Trans. Man–Mach. Syst.*, vol. MMS-10, no. 4, pp. 292–331, Dec. 1969.

[8] A. V. Phatak and G. A. Bekey, "Model of the adaptive behavior of the human operator in response to a sudden change in the control situation," *IEEE Trans. Man–Mach. Syst.*, vol. MMS–10, no. 3, pp. 72–80, Sep. 1969.

[9] R. A. Hess, "Modeling pilot control behavior with sudden changes in vehicle dynamics," *J. Aircraft*, vol. 46, no. 5, pp. 1584–1592, 2009.

[10] R. A. Hess, "Simplified approach for modelling pilot pursuit control behaviour in Multi–loop flight control tasks," *Proc. Inst. Mech. Eng., Part G, J. Aerosp. Eng*., vol. 220, no. 2, pp. 85–102, 2006

[11] R. A. Hess, "Analytical assessment of performance, handling qualities, and added dynamics in rotorcraft flight control," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 39, no. 1, pp. 262–271, Jan. 2009.

[12] P. M. Zaal, "Manual control adaptation to changing vehicle dynamics in roll–pitch control tasks," *J. Guid., Control Dyn.*, vol. 39, no. 5, pp. 1046–1058, 2016.

[13] R. A. Hess, "A model for pilot control behavior in analyzing potential loss–of–control events," *Proc. Inst. Mech. Eng., Part G, J. Aerosp. Eng.*, vol. 228, no. 10, pp. 1845–1856, 2014.

[14] R. A. Hess, "Modeling human pilot adaptation to flight control anomalies and changing task demands," *J. Guid., Control, Dyn.*, vol. 39, no. 3, pp. 655–666, 2016.

[15] L. Lu and M. Jump, "Multiloop pilot model for boundary–triggered pilot–induced oscillation investigations," *J. Guid., Control, Dyn.*, vol. 37, no. 6, pp. 1863–1879, 2014.

[16] C. Chen, W. Tan, H. Li, and X. Qu, "A fuzzy human pilot model of longitudinal control for carrier landing task," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 453–466, Feb. 2018.

[17] K. Van Der El, D. M. Pool, M. M. van Paassen, and M. Mulder, "Effects of linear perspective on human use of preview in manual control," *IEEE Trans. Human–Mach. Syst.*, vol. 48, no. 5, pp. 496–508, Oct. 2018.

[18] D. Cleij, J. Venrooij, P. Pretto, D. M. Pool, MMulder, and H. H. Bülthoff, "Continuous subjective rating of perceived motion incongruence during driving simulation," *IEEE Trans. Human–Mach. Syst.*, vol. 48, no. 1, pp. 17–29, Feb. 2018.

[19] M. C. Dorneich *et al.*, "Interaction of automation visibility and information quality in flight deck information automation," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 6, pp. 915–926, Dec. 2017.

[20] R. J. Kuiper, D. J. Heck, I. A. Kuling, and D. A. Abbink, "Evaluation of haptic and visual cues for repulsive or attractive guidance in nonholonomic steering tasks," *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 5, pp. 672–683, Oct. 2016.

[21] X. Qu, H. Wei and J. Guan, "Modeling of Feeling Institution in Pilot Structural Model (in Chinese)," *Space Med. Med. Eng.*, vol. 14, no. 2, pp. 124–127, 2001.

[22] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile Sensing–from Humans to Humanoids," *IEEE Trans. Robot.*, vol. 26, no. 1, pp. 1–20, Feb. 2010.

[23] D. T. McRuer, R. E. Magdaleno, and G. P. Moore, "A neuromuscular actuation system model," *IEEE Trans. Man–Mach. Syst.*, vol. MMS-9, no. 3, pp. 61–71, Sep. 1968.

[24] J. Smisek, E. Sunil, M. M. van Paassen, D. A. Abbink, and M. Mulder, "Neuromuscular–system– based tuning of a haptic shared control interface for UAV teleoperation," *IEEE Trans. Human–Mach. Syst.*, vol. 47, no. 4, pp. 449–461, Aug. 2016.

[25] M. M. van Paassen, "Biophysics in aircraft control: A model of the neuromuscular system of the pilot's arm," TU Delft, Delft Univ. Technol., Delft, The Netherlands, 1994.

[26] M. M. van Paassen, J. C. Van Der Vaart, and J. A. Mulder, "Model of the neuromuscular dynamics of the human pilot's arm," *J. Aircraft*, vol. 41, no. 6, pp. 1482–1490, 2004.

[27] J. Venrooij, D. A. Abbink, M. Mulder, M. M. Van Paassen, and M. Mulder, "A method to determine the relationship between biodynamic feedthrough and neuromuscular admittance," *IEEE Trans. Syst., Man Cybern., B, Cybern.*, vol. 41, no. 4, pp. 1158–1169, Aug. 2011.

[28] R. A. Hess, "Modeling pilot detection of time–varying aircraft dynamics," *J. Aircraft*, vol. 49, no. 6, pp. 2100–2104, 2012.

[29] D. H. Klyde and C. Y. Liang, "Flight assessment of pilot behavior with smart–cue and smart–gain concepts active," in *Proc. AIAA Atmospheric Flight Mech. Conf. Exhib.*, 2009, pp. 1–23.

[30] F. M. Nieuwenhuizen, P. M. Zaal, M. Mulder, M. M. Van Paassen, and J. A. Mulder, "Modeling human multichannel perception and control using linear time–invariant models," *J. Guid., Control Dyn.*, vol. 31, no. 4, pp. 999–1013, 2008.

A Flight Simulator Study of an Energy Control System for Manual Flight

Karolin Schreiter, Simon Müller, Robert Luckner, and Dietrich Manzey

***Abstract*—This article describes a flight simulator study to validate a new command control system for longitudinal load factor n_x. The system is called nxControl and actuates thrust, speedbrakes, and wheel brakes that directly influence the total energy state of the aircraft. It aims at manually flying with higher precision and lower workload. After a brief overview of the functionality of nxControl and its cockpit interfaces, the key results of earlier proof-of-concept simulator studies are summarized. The focus lies on the comprehensive flight simulator study with 24 airline pilots to conclusively evaluate the complete system. The task was a highly demanding approach trajectory containing segmented-continuous-descent in gusty wind conditions, touch-down, deceleration, and taxi as well as engine failure at low altitude. With nxControl, the pilots achieved higher precision in airspeed and energy control with lower workload compared to conventional manual thrust control. In addition, nxControl achieved safety benefits in case of an engine failure.**

***Index Terms*—Flight precision, human–machine interface, manual flight control system, total energy angle.**

NOMENCLATURE

D, T, W	= Drag, thrust force, and weight, N.
F	= Snedecor's F distribution.
V	= Airspeed, m/s.
V_K	= Flight-path speed, m/s.
g	= Acceleration due to gravity, m/s^2.
n_x, n_z	= Load factor in longitudinal and normal direction, 1.
p	= Probability value, 1.
t	= Student's *t*-distribution.
α_W	= Rotation angle from air-path to flight-path axis system caused by wind, deg.
γ	= Flight-path angle, deg.
γ_E	= Total energy angle, deg.

Subscripts

k	= Flight-path axis.
tot	= Total.

Manuscript received October 12, 2018; revised February 27, 2019 and June 3, 2019; accepted August 10, 2019. Date of publication September 18, 2019; date of current version November 21, 2019. This work was supported by the DFG (German Research Foundation) under Grants LU 1397/3-2 and MA 3759/3-2. This paper was recommended by Ron A. Hess. *(Corresponding author: Karolin Schreiter.)*

K. Schreiter and R. Luckner are with the Department of Flight Mechanics, Flight Control, and Aeroelasticity, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: karolin.schreiter@tu-berlin.de; robert.luckner@tu-berlin.de).

S. Müller and D. Manzey are with the Department of Work, Engineering, and Organizational Psychology, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: simon.mueller@tu-berlin.de; dietrich.manzey@tu-berlin.de).

Digital Object Identifier 10.1109/THMS.2019.2938138

I. INTRODUCTION

GROWING air traffic increases the demands on flight precision and the necessity to cope with complex flight trajectories. If such complex trajectories have to be flown manually, the demands on manual flying skills and pilots' workload will increase considerably. Today's fly-by-wire control laws of commercial airliners already support manual flight, enabling the pilot to directly command flight parameters like pitch rate, roll rate, or load factors. Flight control laws then adjust the primary control surfaces according to the command values and the current flight state. With such control laws, pilots are able to control precisely the aircraft attitude. In addition, pilot workload is reduced as disturbances are automatically compensated; see [1]. However, such support is lacking for manual control of the flight path and airspeed by means of thrust, spoilers, and wheel brakes. Here, pilots set engine parameters, spoiler deflection, or brake pressure (on ground) and have then to wait for the aircraft reaction to readjust their inputs until the result is as intended. Doing this, the pilots have to anticipate the impact of the thrust, as the aircraft reacts slower to thrust inputs compared to the aerodynamic control surfaces. A system called nxControl was developed to support this energy management in manual flight in order to lower these demands on pilots. Analogous to the flight control laws for pitch and roll, nxControl enhances the manual control of all control devices that affect the energy state of the aircraft in flight and on ground. Several flight simulator studies were conducted with the system to determine its benefits for today's demands. This article focuses on a new experiment that investigated the effects of the system in a highly complex flight task with strict demands on flight precision, as it is expected in the future air traffic.

A. nxControl System Description

The nxControl system enables pilots to directly control the impact of thrust, spoilers, and wheel brakes on the aircraft's flight state. The pilot commands the longitudinal load factor in flight-path direction $n_{xk,\mathrm{tot}}$ (further abbreviated to n_x) by means of the so-called nxLever that replaces the throttle levers. The manual control of the flight attitude and flight path with sidestick (or yoke) and pedals remains unchanged. In the investigation, Airbus-like manual flight control laws were used where pilots command vertical and lateral load factor and roll rate. Before detailing the technical aspects of the system, the flight mechanical fundamentals are summarized. The basis is the longitudinal force equation of motion. The load factor n_x

depends, by definition [2], on the external forces (thrust T, drag D, and tilting of lift due to wind) related to weight W. For symmetric flight (zero sideslip angle, bank angle, and change of azimuth angle), n_x is defined as

$$n_{xk,\mathrm{tot}} = \frac{T-D}{W} + n_{zk,\mathrm{tot}} \sin\alpha_W \quad = \frac{\dot{V}_K}{g} + \sin\gamma \qquad (1)$$

where $n_{zk,\mathrm{tot}}$ is the vertical load factor, α_W is the wind angle of attack (between airspeed and flight-path velocity V_K), $\dot{V}_K$ is the flight-path acceleration, g is the gravitational constant, and γ is the flight-path angle. The first expression for n_x represents the sum of the normalized external forces. The second expression relates to the sum of inertial forces and weight. The force equilibrium requires that both sums are equal. By commanding n_x, the pilot directly controls a change in total energy–expressed by $\dot{V}_K$ and γ in Eq. (1)–and the controller adjusts the control devices for thrust and drag. How this energy change is distributed to acceleration and flight path depends on the actual flight path.

From an energy point of view, n_x is equal to the sine of the total energy angle γ_E, also called specific excess thrust or potential flight path angle. The relationship between n_x and γ_E can be derived by differentiating the total energy and dividing it with W and V_K, as for instance [3] shows

$$\sin\gamma_E = \frac{\dot{V}_K}{g} + \sin\gamma \quad = \quad n_x. \qquad (2)$$

That is why n_x and γ_E can be used synonymously. An n_x command equals a change of total energy and can be distributed to change either kinetic or potential energy or simultaneously both. Using the total energy for flight control was introduced by Lambregts in his concept for the total energy control system (TECS) [4]. His approach focused on altitude and airspeed command modes for autopilot and autothrust. He also investigated manual control, where the pilot commands flight-path angle (rate command/angle hold) and TECS takes care of the airspeed [5]. In conventional manual flight, pilots command thrust with the engines and drag with the airbrakes which results in a longitudinal load factor according to (1). The aircraft reaction depends on the actual flight state, and continuous adjustments are necessary to control airspeed and flight path angle. The nxControl system enables the pilots to directly command the impact of engines and spoilers by commanding n_x or γ_E. They command γ_E with a thrust-lever-like inceptor (nxLever) at the center pedestal (see Fig. 1). The lever is equipped with a passive haptic feedback, which is achieved by a pressure pin on the bottom of the handle and notches at characteristic command values for n_x. The notches represent minimum thrust, -15°, -10°, 0°, maximum continuous thrust (MCT), and take-off/go-around (TO/GA) thrust. This allows the pilot to find these positions without visual control. It further supports the pilot's awareness for certain commands. In addition, the use of speedbrakes in flight and thrust reverser on ground requires conscious activation. For this purpose, the nxLever is equipped with two switches to activate these critical control elements. This ensures that the pilot is always aware of their status.

To monitor the impact of the control system, energy cues on flight displays are beneficial. Energy cues on head-down

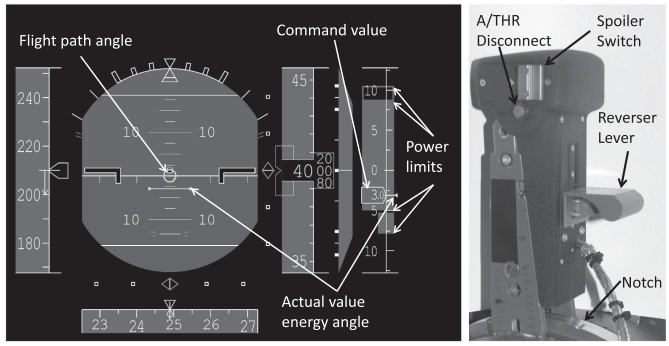


Fig. 1. Human–machine interface of nxControl.

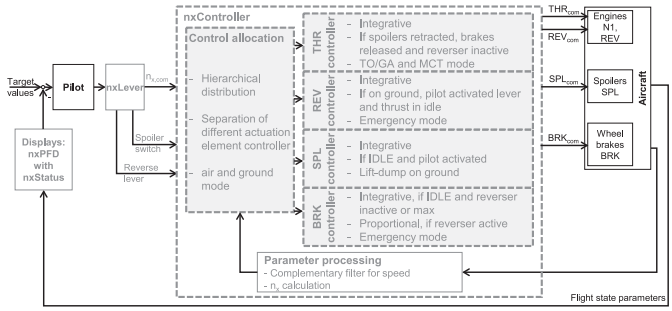


Fig. 2. Overview of the nxControl system with the pilot in the control loop.

primary flight displays were already used in different research works, e.g., with perspective flight-path information [3], [6] or with standard PFD symbology but an energy-scaled altitude indicator [7]. The nxControl system includes two new elements on a modified primary flight display called nxPFD (see Fig. 1), where the pilot can monitor the command and its impact on the aircraft. The first one is a horizontal line integrated in the artificial horizon, representing γ_E. The total energy angle in alignment with the indicated flight-path angle provides a direct cue of the relationship of altitude and speed change. The second one is a specific nxStatus scale that is added to the right side of the vertical speed indicator. The primary objective of this scale is to support the awareness of pilots concerning the effects of given inputs on the controller behavior. Specifically, the nxStatus scale indicates the commanded value as well as their limitations. In flight, the limitations represent maximum and minimum n_x values that can be commanded with thrust and speedbrakes. On ground, they show the n_x limits for thrust, wheel brakes, and thrust reverser commands. The feedback flight control laws of nxControl adjust the actuators, i.e., engines, spoilers, and wheel brakes, according to the selected n_x command value (see Fig. 2). The actuators are used in a hierarchical way (daisy chain control allocation) with forward thrust as primary actuator. If the engines are in idle, spoilers are used during flight or wheel brakes and thrust reverser on ground to reduce n_x. The system is described in more detail in [8]–[10].

The general behavior of the nxController was designed to mimic the typical pilot's control strategies in energy management as best as possible. For achieving this objective, several pre-studies and interviews with experienced pilots were

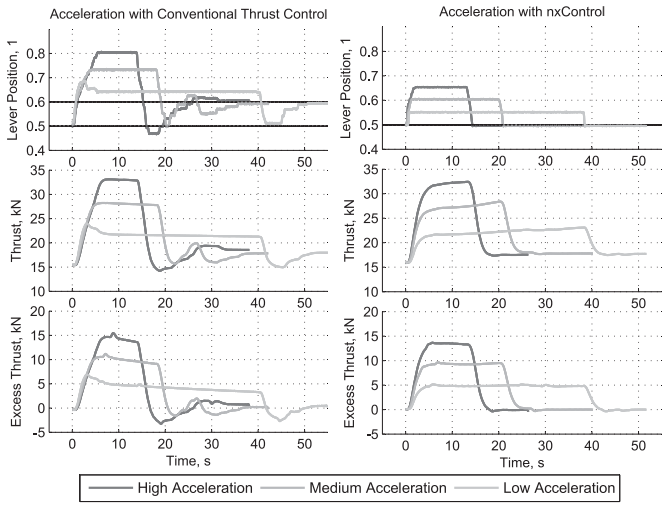


Fig. 3. Comparison of input characteristics for speed change between conventional thrust control and nxControl.

conducted. The influence of the nxControl system on manual flight control can be illustrated for a typical flight task. Fig. 3 shows the time histories for a speed increase from 220 to 240 kt with three different mean accelerations. Altitude was maintained with sidestick pitch commands. With conventional thrust control (Fig. 3, left), the pilot has to push the thrust lever forward to raise thrust force. The resulting excess thrust (difference of thrust and drag) causes the targeted acceleration according to (1). Drag force increases with higher speed, which causes a gradual decrease in excess thrust and acceleration. When the target speed is reached, the pilot must return the thrust lever and adjust a new trim position. As it can be seen, several adjusting inputs are necessary to find the required trim position and where the excess thrust or, in other words, the energy rate becomes zero. With nxControl (Fig. 3, right), the pilot has to push the thrust lever forward to raise thrust. A glance on the n_x indicator confirms that the commanded excess thrust causes the targeted acceleration. Drag force increases with higher speed and the controller compensates this by increasing thrust accordingly. The excess thrust and acceleration value remain constant over the acceleration phase. When target airspeed is reached, the pilot must return the nxLever to the zero position at the middle notch. The controller will hold the current energy state with zero excess thrust and energy rate. The general handling, thus, remains as close as possible to the conventional handling characteristics. However, the n_x command relieves the pilot from mentally transforming engine parameters, spoiler deflections, or brake pressure to the energetic state of the aircraft. Instead of the various control elements, the pilot directly commands the required aircraft reaction. Disturbances affecting total energy are compensated automatically. The distribution of the total energy to kinetic and potential energy by altering the flight-path angle with the pitch control device remains conventional. The displayed current and commanded energy angles provide a direct relation to flight-path angle and acceleration. The nxPFD, thus, provides support for energy distribution. With those characteristics, the nxControl system aims at more precise energy management in manual flight and shall reduce workload.

B. Previous Research

Previous flight simulator studies already showed the positive effect of the nxControl system on flight precision and workload in standard tasks like approaches with Instrument Landing System (ILS) and advanced approach tasks with demanding required navigation performance (RNP). During airwork [11] or ILS straight-in approaches [12], the participating airline pilots achieved all important flight parameters with the same precision when using nxControl compared to the well-trained conventional thrust and spoiler control, even though they used nxControl for the first time. As expected, this performance was reached with significant less thrust lever activity while maintaining the usual control strategy. This suggests that the new system does not affect the basic flying techniques in general, but enables pilots to achieve flying goals with less effort. For advanced approach patterns, e.g., steep and curved approaches in the mountain area of Salzburg Airport, the precision of speed control and workload involved was even better with nxControl compared to the conventional manual flight [8], [9]. Importantly, it was found that the use of nxControl system did not impair the pilots' situation awareness compared to conventional manual flight. The pilots rated their workload subjectively higher due to the additional energy information [9]. In addition, also the suitability of nxControl for coping with air/ground transition was shown [10]. This included a proof of the new control concept for standard take-off, landing, and go-around procedures, as well as off-nominal scenarios with engine failures before and after decision speed. In summary, the previous results showed that the nxControl system can be used in all flight phases. Workload and precision benefits can be expected when pilots are required to fly complex trajectories involving high demands on energy control.

C. Current Research

The current research capitalizes on the previous studies by investigating the performance consequences of nxControl in even more challenging flight tasks to be expected in the future. Because of a predicted increase of air traffic during the next decade, major airports will have to cope with a significantly higher air traffic density. This will result in more complex approach trajectories, which have to be flown with an even higher precision than today. In order to simulate such demands, pilots were required to fly a complex noise abatement approach procedure with conventional thrust control versus a prototype of the nxControl system in a flight simulator. The task was to fly an approach to Frankfurt Airport, which reduces noise by means of three factors: lateral curved guidance around populated area, segmented continuous descent, and late aircraft configuration. In addition, engine failures below stabilization height were introduced to investigate safety issues. A standard ILS approach without any disturbances was used as a reference scenario. Certified airline pilots flew these scenarios with both the nxControl system and the conventional thrust control.

No noticeable differences between nxControl and conventional thrust control were expected in the standard approach, but clear benefits of nxControl in terms of higher precision and

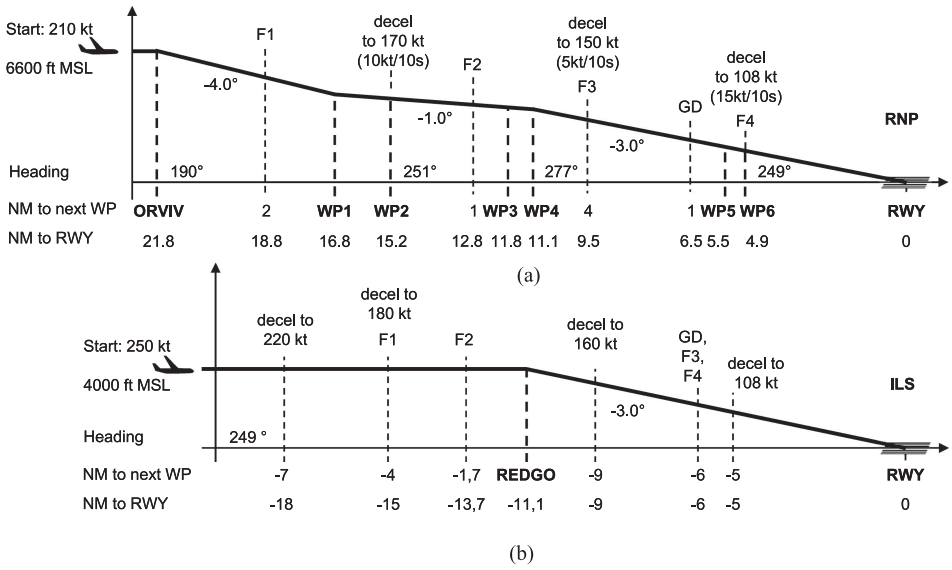


Fig. 4. Vertical approach patterns of (a) RNP and (b) ILS task.

less control effort were expected to emerge in the segmented-continuous-descent. Each approach ended with landing and leaving on a designated runway exit to validate the air-to-ground transition.

II. Experimental Method

A. Participants

A total of 24 commercial airline pilots (all male, 7 captains, 17 first officers) participated. Their age ranged from 27 to 52 years with an average of 37.2 years. Their flight experience ranged from 900 to 17 000 h. All pilots had an Airbus type-rating (20 A320, 4 A330/340/350) and were familiar with the Airbus displays, the sidestick, and the typical Airbus control laws.

B. Flight Simulator and Experimental Setup

The experiment was conducted in the fixed-base flight simulator SEPHIR (Simulator for Educational Projects and Highly Innovative Research) [13] configured as VFW-614 ATD. The cockpit was equipped with displays and sidestick (including flight control laws) similar to some modern commercial transport aircraft. Despite minor differences (e.g., the VFW-614 ATD's higher aerodynamic drag), the flight characteristics, handling, and cockpit configuration closely resembled an Airbus A320.

Two configurations were used, referred to in the following as *conventional* and *nxControl*. In the conventional configuration, the simulator was equipped for manual raw-data flight with standard PFD and conventional thrust control via commands for fan rotation speeds. This configuration served as reference for assessing the effects of the nxControl system. In the nxControl configuration, the simulator was supplemented with the enhanced energy information on the PFD, the energy angle command with the nxLever, and the underlying nxControl law (see Figs. 1 and 2).

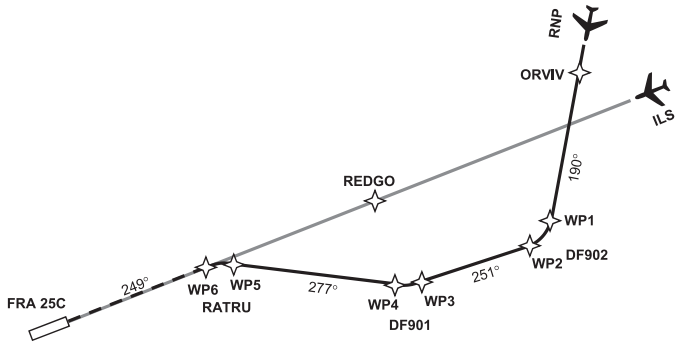


Fig. 5. Lateral approach patterns of ILS and RNP task.

C. Tasks

1) A Curved Segmented-Continuous-Descent RNP Approach represents a potential complex approach pattern of the future. Fig. 4(a) shows its vertical profile. The approach started at 1 NM to ORVIV at 6600 ft above mean sea level (MSL) with an airspeed of 210 kt. The flight-path angles from the top of descent at ORVIV to the runway touchdown zone change in three steps ($-4°$, $-1°$, and $-3°$). This vertical profile was inspired by a procedure described in [14]. The lateral profile of the RNP task (Fig. 5) was based on an existing RNAV GPS night procedure to FRA 25C with several turns to fly around residential areas. The lateral profile contained three turns with track changes of $61°$, $26°$, and $-28°$ beginning at waypoint WP1, WP3, and WP5, respectively. The turns were designed for bank angles of $25°$, $15°$, and $-15°$. The navigation precision was increased from standard RNP 0.3 to RNP 0.1 to simulate future precision requirements leading to tolerances of ± 100 ft vertical and ± 0.1 NM lateral. The lateral and vertical RNP deviations were indicated by the PFD's RNP markers. One dot corresponded to the desired lateral and vertical tolerances. The procedure for aircraft configuration and speed reduction was related to distances to waypoints ahead

(see Fig. 4). It requested a late aircraft configuration of flaps (F1–F4) and gear (GD) for high noise reductions. The three speed reduction phases required constant decelerations of 10, 5, and 15 kt per 10 s, respectively. The deceleration rates could be estimated either by the length of the speed trend (in kt per 10 s) or by the difference between flight-path and energy angle in the nxControl configuration (10 kt per 10 s equals 3°). The participants had to follow this approach procedure as precise as possible to ease analysis. Wind of 20 kt from 320° with light turbulence disturbed the flight. Visibility was considerably impaired by clouds. Turbulence and clouds lasted until 1000 ft height above ground level (AGL).

2) A Straight-In ILS Approach to Frankfurt FRA 25C with standard flight procedures provided the reference scenario, see Fig. 4(b). The participants had to follow the given profile and procedure as closely as possible. The glide slope should be maintained with a precision less than 1 dot deviation lateral and vertical that are indicated by the glide slope and localizer marker. The task began at 4000 ft above MSL at 250 kt and 20 NM to the threshold. The flight procedure precisely defined when to change airspeed or to configure flaps and landing gear. The sequence corresponds to standard instrument procedures. It was customized to enhance the observability of the participant's energy adjustments in three subtasks. First, to maintain a speed of 180 kt while intercepting the 3° glide slope requires a precise thrust adjustment. Second, deceleration to 160 kt and maintaining 160 kt was required until 5 NM. This requirement closely corresponded to an approach procedure at London Heathrow Airport. Third, configuration change to landing configuration (gear down, flaps F3, and flaps F4) should be performed sequentially at constant speed at 6 NM. This sequence demanded substantial thrust activities to compensate for the additional drag. There was no wind and the outside view was considerably impaired by clouds until 1000 ft height AGL.

3) Engine Failure To address failure conditions with nxControl, engine failures below stabilization height at 850 ft AGL were investigated. This supplemented earlier engine-failure tests during take-off. An engine failure in final approach is interesting as speed and altitude tolerances are small to ensure flight safety, and as the nxController supports pilots by compensating the loss of thrust and maintaining the commanded energy angle.

4) Landing and Taxi was included in all approaches. The task was to touch down on the runway, decelerate with brakes and thrust reverser to approximately 20 kt taxi speed to leave the runway at the first exit, and to stop at the stop bar on the runway exit.

D. Experimental Measures

1) Flight Precision was assessed based on the deviations from given target values (actual-target). The deviations of the total energy height H_E was used to assess how well the participants could manage the total energy E_{tot}. The deviations of airspeed V and altitude H were examined as representatives of kinetic and potential energy. The lateral deviation was observed to detect potential impacts on flight-path precision although nxControl does not directly influence lateral dynamics. For purely descriptive analysis, the simple deviation measures were used. Root-mean-square error (RMSE) scores aggregated over time were used for the statistical analyses in order to weight larger deviations higher than smaller ones.

2) Workload was assessed with respect to two different aspects. First, the input activity (either at the thrust lever or the nxLever) was determined as an objective measure of pilot's workload for thrust control. The lever activity was defined as a change in lever position above a threshold of 0.5% of the maximum lever range (about 2 mm at the handle) within a time interval of 2 s. It was normalized by the number of intervals. The higher this measure, the more often a pilot had to invest cognitive and physical effort to assess and adjust the energy state, and thus, the higher the workload to be invested for energy management. Second, the NASA-TLX [15] was used to assess the subjectively perceived workload. It consists of six rating scales, addressing different aspects of workload. Overall workload measures per participant and flight task were obtained by averaging the ratings across these scales [16].

E. Procedure

The experiment started with a short briefing. This involved the familiarization with simulator in the conventional configuration, including an explanation of the differences to an Airbus cockpit. In order to support manual raw data flying in conventional configuration, the pilots received a pitch-and-power table for the VFW-614 ATD. They performed the ILS task including landing and taxiing and the RNP task. The test engineer performed the tasks of the pilot not flying.

The main experimental session was split into two blocks characterized by the simulator configuration (i.e., conventional and nxControl). Each pilot performed both experimental blocks. The order was counterbalanced across all participants. At the beginning of the nxControl block, the system and its functionality were explained in detail and trained on the basis of short example flight tasks and the ILS task. This training of about 1.5 h was only necessary in the nxControl block as pilots are well familiar with conventional flying. The following sequence of four experimental trials was identical in both blocks.

1) ILS task, including landing and NASA-TLX;
2) practicing the RNP procedure in undisturbed atmosphere;
3) RNP task #1, including landing with engine failure;
4) RNP task #2, including landing and NASA-TLX.

After the experimental trials, a debriefing interview addressed the general use of nxControl, engine failure, and the air–ground transitions. A detailed listing of all questions used in this structured interview can be found in [17]. Additionally, the pilots were asked to predict their workload during the RNP task, assuming that they would have gained sufficient operational experience with nxControl. The NASA-TLX was used for this purpose. In total, the experiment took about 4 h for each participant.

F. Data Analysis

The first right turn of the RNP task was particularly demanding and difficult to execute, due to the relatively high speed at

the beginning of the turn, the narrow turn radius, the change in relative wind direction, and the change in required flight-path angle at the same time. Pilots who were inattentive at this point missed the beginning of the turn or did not bank quickly and sufficiently enough. As a consequence, they had considerable effort to recapture the specified lateral flight-path. This effort had a negative impact on the subsequent flight and added a cumulating error to our precision measures. Since this issue was rather dependent on the pilot's attention or concentration instead of the configuration, trials with more than 0.2 NM lateral deviation in the first turn were declared as not representative. Therefore, 13 of all 96 RNP trials were excluded from the analyses. The remaining data were analyzed with respect to six different aspects.

The first analysis comprised a descriptive evaluation of the data derived from the trials of the RNP task addressing the impact of nxControl on precision and workload during different sections of the approach. The time histories of various flight parameters of each pilot and configuration were mapped on the distance to runway to synchronize the actions of the target procedure and to compensate differences between pilots in speed management. Then the mapped time plots of all pilots were aggregated to median plots ("median pilot") that were used for a visual comparison of the effects of both configurations. The median plots were analyzed for six approach sections as defined in Fig. 6 to determine the influence of the nxControl system on individual parts of the RNP task.

The second analysis statistically compared flight precision and workload for both configurations while performing the ILS and RNP tasks. Flight precision data (RMSE) and workload data (lever activity and NASA-TLX) were analyzed by a 2 (configuration [conventional, nxControl]) $\times$ 2 (task [ILS, RNP]) repeated measures analysis of variance (ANOVA). This approach was selected, due to an assumed normal distribution of these variables' data in the basic population. The ANOVA's F-test determined the probability that the observed differences in the mean values of the experimental measures for each test condition reflected only random variation, i.e., a variation caused neither by configuration nor by task. A probability $p \leq 5\%$ was defined as the threshold to reject the assumption of pure random effect (significance level). The mean values of both RNP approaches were used for the ANOVA. As the ILS procedure did not specify any particular deceleration targets, all deceleration phases were excluded from the comparative analysis.

Third, the performance during the deceleration phases of the RNP task with nxControl versus conventional control was compared in a separate nondirectional t-test analysis.

The fourth analysis addressed the handling of the engine failure during the first trial of the RNP task. Since the engine failed at low altitude and at an approach speed close to stall, it was important for the pilots to maintain speed and not to descent below glide slope. Performance in conventional and nxControl configuration was compared both descriptively and statistically. The descriptive comparison included an inspection of deviations of airspeed and altitude and of the amount of lever activity as an indicator of workload. The time plots of all participants as well as the median plots were analyzed to detect detailed differences between conventional and nxControl configuration. The statistical evaluation then addressed a comparison of the aggregated precision data (RMSE) and workload (at lever activity) during the phases with engine failure via t-test.

Fifth, the impact of the nxControl system during transition from air to ground was descriptively analyzed in the nxControl configuration. The landing phase was bounded from 200 m before touch-down to 100 m after runway exit. The data of all participants in the second experimental RNP trial and their median were considered. The interesting values for this task were the pilots' load factor command and thrust reverser activation as well as the resulting airspeed. The underlying controller commands of forward thrust, thrust reverser, spoiler, and wheel brakes give insight into the functionality of the control law in ground mode (GMDE). A comparison to the conventional configuration was not possible, as manual braking with the pedals was physically not available in the simulator. Thus, the pilot behavior in the conventional configuration was not representative for the conventional air–ground transition.

Sixth, a final analysis addressed the pilots' predictions of presumable workload imposed by the RNP tasks in nxControl configuration, assuming that they had more operational experience. Although these predictions are somewhat speculative, they provide insights into the potential of the nxControl system as perceived by the pilots. Predictions on all six NASA-TLX scales were compared to the actual scores of the experimental RNP trials in the conventional and the nxControl configuration. One-way ANOVAs were used for this purpose, combined with post hoc contrasts according to the Dunn–Šidák procedure [18].

III. Results

A. Descriptive Evaluation of the RNP Task

The median plots of the precision parameters in Fig. 6 are the basis of the analysis of the six sections marked above. The entire descriptive results of the precision parameters of the RNP task can be found in [17].

Section I: To intercept the glide slope, it was necessary to reduce energy supply. Therefore, the pilots reduced the lever position in both configurations. With nxControl, the pilots used the given glide slope value as command value. In the conventional case, the pilots reduced the lever input gradually to gain the appropriate thrust setting close to idle position. As the pilots started to reduce thrust close to the intercept point, the necessary energy reduction was delayed and the error in energy height increased after the intercept point in both configurations. Due to the gradual lever setting in the conventional case, this error was higher than in nxControl configuration. In conventional configuration, the pilots tried to reduce the energy error with idle thrust (backward lever position). In nxControl configuration, they commanded a value below the value of the actual flight-path angle. The altitude error was nearly identical for both configurations. The difference in total energy was observable in the higher airspeed deviation, in the conventional configuration. After setting flaps F1 at 18.8 NM, the additional drag supported energy reduction. With nxControl, the pilots made negligible lever inputs to correct their prior energy error and maintained

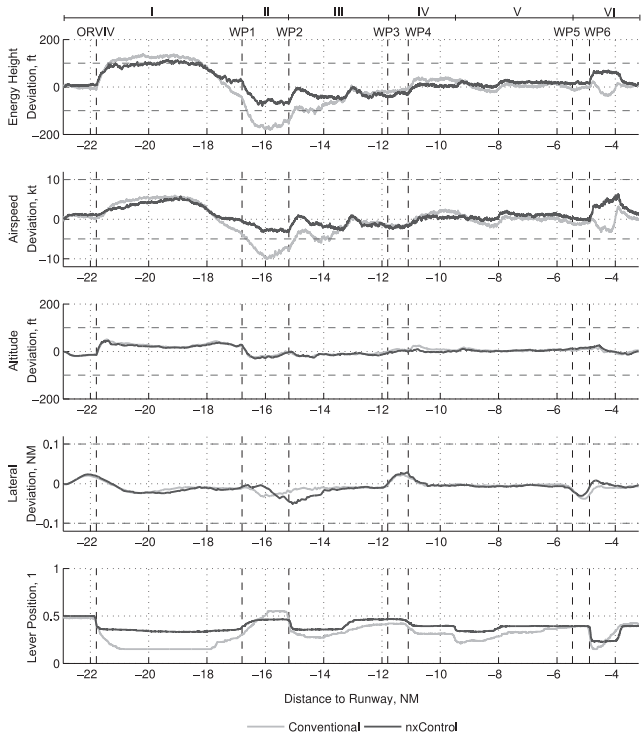


Fig. 6. Median plots of flight parameters for conventional and nxControl configuration in RNP task.

a command value corresponding to the actual flight-path angle. With the conventional configuration, the pilots gradually changed lever setting that led to considerably undershooting the target speed.

Section II: This section turned out to be the most challenging part of the scenario, as a rather steep turn disturbed by crosswind together with change in glide slope at constant airspeed had to be controlled. The thrust lever movements show for both configurations that the pilots started to increase the setting after waypoint WP1, when they simultaneously began to turn and change glide slope. The lagging inputs led to an unintended reduction in airspeed. In the conventional case, the pilots gradually commanded a much higher fan rotation speed than before. Therefore, they obtained the lever setting for constant speed later than with the nxControl configuration, in which the pilots used the glide slope value as command. That delay produced a substantial error in airspeed while speed could be nearly maintained with nxControl. However, the pilots stated that the unfamiliar setting of the n_x command attracted their attention significantly. This led to a rising deviation of the lateral flight-path toward the end of the turn that could only be corrected during the following straight segment (Section III). As before, the altitude deviation did not show noticeable differences between nxControl and conventional configuration.

Section III: At WP2, the pilots had to reduce speed with the given deceleration rate to the target speed of 170 kt. Meeting the requested deceleration would result in constant deviations of speed and energy height. With nxControl, the pilots could calculate a particular command value for energy angle due to the direct relation of deceleration rates to energy angle. It can be seen that the pilots set and maintained a certain lever position for the deceleration, but the energy height was reduced more than requested. That means that they used a command value lower than calculated. In the conventional configuration, the energy height deviation stayed nearly constant, but the pilots adjusted their lever inputs throughout the deceleration phase. Especially to set the final target speed (approx. at 13.5 NM to runway), the lever movements of the conventional configuration constantly changed, whereas the nxLever was directly positioned. The speed deviation was lower than with the conventional configuration at any time. At the end of the deceleration (approx. at 13 NM to runway), the pilots had eliminated all parameter deviations. At that point, they had to set flaps F2. With nxControl, no further action was necessary to maintain the flight state, but gradual thrust lever movements were observable in the conventional configuration to compensate for the additional drag. The precision parameters evolved similarly after setting F2 for both configurations with minor negative deviations in energy height.

Section IV: The second turn changed the relative wind direction. The pilots adjusted thrust lever inputs in the conventional configuration while no input was needed in nxControl configuration, where the controller compensated the disturbance. That caused a constant energy deviation for both configurations. At the end of the turn (at WP4), the target glide slope increased to the final value of -3°. The pilots lowered the lever settings in both configurations, whereas the transition to the final value took longer for the conventional configuration and caused higher energy deviations.

Section V: The following constant deceleration phase was disturbed by flaps deployment to F3. With nxControl, the pilots set a command value corresponding to the required deceleration rate. The controller set the engines taking drag changes into account. Therefore, no additional nxLever input was necessary. Conversely, the pilots adjusted the thrust lever input of the conventional configuration during the whole deceleration phase. The lagged reaction of the pilots led to higher deceleration rates as requested, inducing higher deviations in energy height and airspeed than with nxControl. At the end of deceleration (approx. at 8 NM to runway), the pilots commanded the energy angle equal to the flight-path angle with nxControl, using the energy angle indication. In the conventional configuration, again multiple lever movements were observable. The extension of the landing gear at 6.5 NM to runway disturbed the energy state. With the conventional configuration, the pilots adjusted their lever input according to higher drag in a gradual manner leading to a slight energy drop. With nxControl, the control law maintained energy deviation constant without pilot input.

Section VI: The last turn again changed the relative wind direction. In both configurations, the pilots chose constant lever settings. That is why the energy height in the conventional configuration decreased, while the nxControl law again compensated the disturbance in energy and led to lower error in energy height. At the end of the turn, the pilots had to decelerate and increase flaps setting to F4 simultaneously. In the conventional case, they lowered the lever input earlier and stronger. Therefore, the deceleration rate was higher than requested and actual

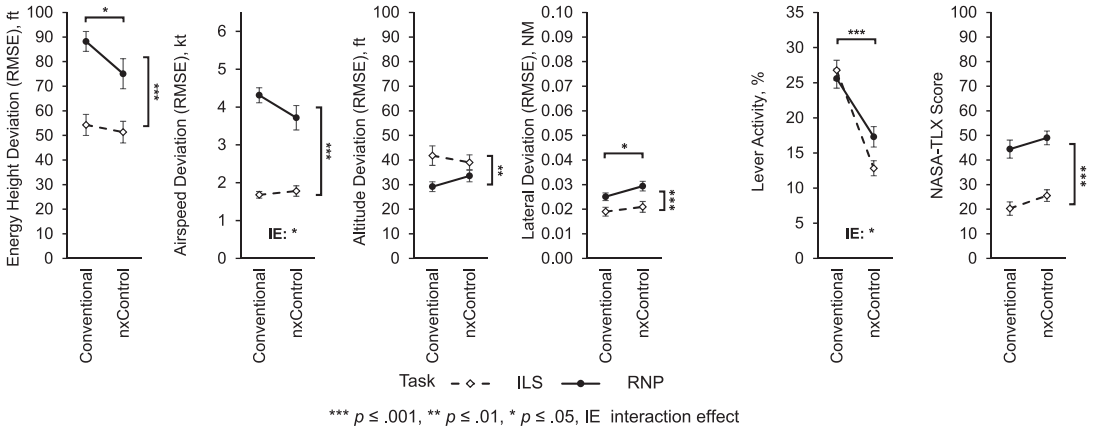


Fig. 7. Statistical comparison of experimental measures for conventional and nxControl configuration in RNP and ILS tasks.

TABLE I
STATISTICAL COMPARISON OF CONVENTIONAL VERSUS NXCONTROL IN RNP AND ILS TASKS

ANOVA Results	Configuration			Task			Config. × Task (IE)		
	$F(1,21)$	p		$F(1,21)$	p		$F(1,21)$	p	
Energy height deviation (RMSE), ft	5.30	.032	*	46.90	<.001	***	1.64	.214	
Airspeed deviaton (RMSE), kt	2.39	.137		108.74	<.001	***	7.84	.011	*
Altitude deviation (RMSE), ft	0.13	.727		8.56	.008	**	3.47	.077	
Lateral deviation (RMSE), NM	4.69	.042	*	16.70	<.001	***	0.74	.399	
Lever activity, %	73.73	<.001	***	2.09	.163		7.34	.013	*
NASA-TLX Score	3.35	.084		114.61	<.001	***	0.09	.764	

Significance levels of t-test: ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$.

energy height and airspeed fell below the target values. With nxControl, the pilots set and maintained an n_x-command value according to the given deceleration rate right after waypoint WP6. Due to the delays of the engine dynamics, the necessary deceleration rate was delayed. Therefore, the deviation in energy height and airspeed increased in positive direction. The constant deviation in energy height shows that a constant energy angle was obtained, but it was not constantly transferred to airspeed as deviation of altitude was corrected at the same time. At the end of the deceleration phase, the pilots brought the lever command back to the value for the -3° flight path. With nxControl, the lever movements were more direct compared to the conventional lever movements.

B. Statistical Evaluation of ILS and RNP Task

The mean precision scores (RMSE) and the mean workload scores achieved in both configurations are shown separately for each tasks (ILS and RNP) in Fig. 7. The error bars represent the standard error. For each experimental measure, the results of the ANOVA are given in Table I. The columns represent the main effects of configuration and task as well as their interaction effect, i.e., effects in configuration that are dependent on the type of task (e.g., precision improvements in RNP but not in ILS task).

Considering the effect of task, most measures show significant differences between the ILS and RNP tasks. This was expected as the RNP task is the more demanding task by design. The precision parameters of energy height, airspeed deviation, and lateral deviation, as well as the subjective workload were significantly higher than in the ILS task. The altitude deviation showed a reverse effect; it is higher in the ILS task than in the RNP task. Thrust lever activity is similar for both tasks.

The ANOVA revealed several effects caused by configuration. The precision parameter energy height and the workload parameter lever activity showed significantly lower values for nxControl compared to the conventional configuration. The lateral deviations were significantly higher for nxControl. The other measures did not show configuration effects.

Interaction effects were found for speed and lever activity. In the RNP task, the airspeed error was reduced in configuration nxControl, whereas it remained approximately unchanged in the ILS task. The lever activity of conventional configuration was similar in RNP and ILS tasks. With nxControl, the lever activity decreased more in the ILS task than in the RNP task.

Table II shows the results of the deceleration phases of the RNP task. The t-test revealed significant effects (indicated by one, two, or three asterisks) in the measures energy height, $t(21) = 2.25, p = .035$, lateral deviation, $t(21) = -3.28, p = .004$, and lever activity, $t(21) = 4.22, p < .001$. The error of the energy height and the lever activity were significantly lower when using the nxControl system compared to conventional configuration. Yet, at the same time, a slightly raised lateral flight-path error was observable in configuration nxControl compared to conventional.

TABLE II
MEAN VALUE COMPARISON OF CONVENTIONAL VERSUS NXCONTROL FOR DECELERATION PHASES AND ENGINE FAILURE IN RNP TASK

Difference of Mean Values nxControl – Conv.	RNP Deceleration		Engine Failure	
Energy height deviation (RMSE), ft	-14.81	*	-5.79	
Airspeed deviaton (RMSE), kt	-0.014		-0.025	
Altitude deviation (RMSE), ft	4.04		-1.55	
Lateral deviation (RMSE), NM	0.014	**	0.001	
Lever activity, %	-9.3	***	-23.1	***

Significance levels of t-test: $^{***}p \leq .001$, $^{**}p \leq .01$, $^{*}p \leq .05$.

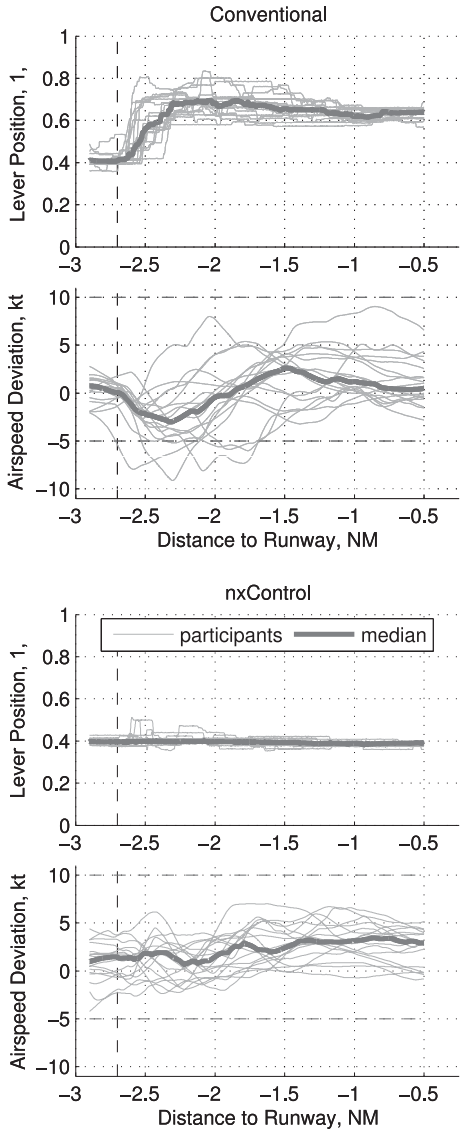


Fig. 8. Lever movements and airspeed after engine failure.

C. Engine Failure

Table II shows the difference between the mean values of nxControl and conventional with indicators for significant effects. The precision parameters of the two configurations had only small and not statistically relevant differences. Nevertheless, there was a clear and statistical significant lower lever activity with nxControl, $t(17) = 6.76, p \leq .001$. Fig. 8 shows the histories of airspeed and lever positions of all participants after engine failure. The approach tolerances for speed are marked (+10 kt, −5 kt). In the conventional case, the lever commands gradually increased directly after the engine failure and they were reduced beginning approx. 0.5 NM until almost 2 NM later. The airspeed dropped immediately after the engine failure and was recovered about 1.5 NM later. The median plot shows values close to the lower tolerance limit. Five participants undershot the lower speed limit. In contrast, Fig. 8 shows almost constant values for lever movements and airspeed with nxControl. Only a few participants moved the nxLever shortly after engine failure with a noticeable value. All participants maintained speed within the given tolerances. Comparing altitude development after engine failure, no considerable differences emerged between conventional and nxControl configuration.

After inspecting the history plots of airspeed and lever position, it became obvious that interindividual variation in flight performances seems to decrease with nxControl compared to conventional. This impression was confirmed by an analysis of the interquartile range of the flight parameters such as airspeed, lever activity, energy height, and altitude, all $p \leq .001$.

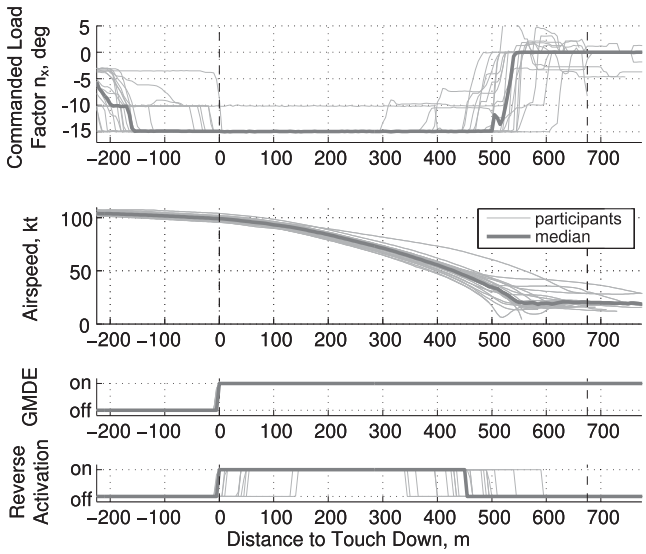


Fig. 9. Transition air to ground after second RNP trial with nxControl.

D. Air–Ground Transition

Fig. 9 shows selected flight parameters during the landing phase versus the distance to touch down starting 200 m before touch down. The commanded load factor values vary between the participants. Nevertheless, the median values can be regarded as representative behavior and are basis for the following results description. During flare, the pilots reduced the n_x command value from approx. -3° to -10° and later -15° on average. These values corresponded to the two lower notches in the lever path. In consequence, airspeed began to reduce. After touchdown, the pilots maintained lever position and activated thrust reverser. The aircraft decelerated constantly. At 50 kt, the n_x controller deactivated the thrust reverser command. When the

aircraft reached the target taxi speed of 20 kt, the pilots moved the nxLever to the middle position to maintain the speed. The pilots reached the runway exit with constant speed.

IV. Discussion

The results of the ILS and RNP tasks showed beneficial effects of nxControl for energetic precision parameters and physical workload when compared to conventional manual flight. As expected, the energy height measure showed clear benefits of the nxControl system for overall energy management. The pilots deviated less and with a smaller amount from the given optimal energy profile. This was particularly noticeable during flight-path changes and decelerations. The replacement of the conventional thrust lever command by a command that is directly related to energy aids energy control and prevents deviations from required energy levels. When considering airspeed, the nxControl system was more beneficial in the demanding RNP task, especially at changes of flight-path and wind conditions. This was expected by the results of the previous studies [8], [9], [12]. It is assumed that the pilots used the flight-path angle as reference for the energy angle command. In addition, the automatic compensation prevented energy errors due to disturbances (wind, flaps setting, landing gear extension) and thus relieved the pilots. The vertical flight-path precision did not show significant differences comparing both configurations. As the vertical control via sidestick remained the same for both configurations, this might not be surprising. However, there was a marginal, yet not significant, tendency toward higher altitude errors in the more demanding RNP task. It should be taken into account that the altitude error in the RNP task was only about one-third of the total energy error in terms of energy height. In contrast, the altitude error covers more than 75% of the total energy error in the ILS task. Hence, the airspeed error is more substantial than the altitude error in the RNP task and the altitude error more substantial in the ILS task. Both more substantial parameters showed tendencies to smaller errors with nxControl, just like the total energy. Surprisingly, significant differences were found in lateral deviation comparing both configurations, especially emerging at the end of the highly demanding first turn of the RNP task. The pilot comments following the task imply that they focused more on precise control of the energy change with nxControl. With conventional configuration, the focus was on lateral control, but energy control was inferior. Even though the altitude and lateral deviations were rather small errors, the tendency for higher errors with nxControl needs to be monitored, as similar tendencies have been found in the prior study with the Salzburg RNP approach [8]. As the pilots were not familiar with nxControl, we assume that they had to focus on the new system and the necessary inputs, and thus the lateral and vertical control via sidestick was less prioritized. These effects might vanish when the pilots are sufficiently trained and have gained experience with the system.

Fewer input actions at the input lever were observed when using the nxControl system. This became visible in almost every situation when the energetic flight state needed to be changed, i.e., change of flight-path angle and predefined deceleration phases. As nxControl provided reference to the actual flight-path angle and visualized command and impact, the command inputs were more direct and goal-oriented. In addition, in situations including energy disturbance by wind or flap deployment, the automatic compensation of energy disturbance due to the given command value relieved the pilots of incremental input corrections. Since the RNP task generally required more energetic changes, the reduction of lever activity was unsurprisingly less pronounced than in the ILS task. Fewer control inputs were assumed to be correlated to lower cognitive and physical workload. However, the subjective workload rating did not validate this finding. We suppose that the novelty of the system affected the results of the workload measures. A longer familiarization phase may reduce this influence. The pilots predicted much lower NASA-TLX scores for the curved continuous-descent RNP approach, when they would implement nxControl in their daily use. During the debriefing, the pilots commented positively on the future use of nxControl, as it would enable easier manual flight. They also expected advantages for the pilot monitoring when identifying the intentions of the pilot flying or monitoring the power setting.

After engine failure, significant differences in lever activity of the two configurations became apparent. In the conventional configuration, the pilots had to compensate the thrust loss of one engine manually. The nxControl system automatically adjusted thrust according to the commanded energy angle. No lever adjustment was needed. Nevertheless, some pilots briefly pushed the nxLever forward and explained this behavior as habit from conventional control. In addition, the direct adjustment of thrust setting in nxControl prevented safety critical loss of speed. In conventional configuration, the pilots reacted delayed due to the surprising situation and had to find the correct thrust lever setting, which required several adjustments. The pilots stated that nxControl allowed focusing on lateral control instead of taking care of both thrust and lateral control. Moreover, the pilots performance became more homogeneous, i.e., the quality of the reaction on engine failures does not depend on the individual flying skills that much with nxControl. In general, the vast majority of pilots rated the behavior of the nxControl law as good, as speed was maintained better with no or very little input. They assumed that automatic power-up of one engine in this specific situation may lead to undesirable lateral deviations. These comments should be considered for further improvements of the control concept. Nevertheless, the suspected higher lateral deviation was not apparent in the simulator data.

All pilots accomplished the transition from flight to ground mode despite different control techniques as no procedure was defined. The pilots handled the system well, especially during transition to ground mode until reaching taxi speed. Potential for optimization exists for taxiing. Sometimes, operating errors occurred when the lever should be moved to the middle position to command zero deceleration at taxi speed. It is assumed that the unusual handling made it more difficult for the pilots as they usually control thrust from the most backward lever position. In general, the pilots were rather satisfied with the behavior of nxControl on the ground. Nevertheless, they saw hardly any advantage for taxiing, as the unusual handling

characteristics resulted in an equal or higher effort. Due to the very different characteristics between flight and ground movements, pilots would tolerate the switch to conventional control on ground. The pilots also noted that the current nxControl ground mode concept may not fit today's braking systems in terms of brake temperature and wear. Nevertheless, it could be suitable for newer electronic taxi system concepts such as motorized wheels.

V. Conclusion

The results of this study show that nxControl offers benefits for speed and total energy precision in the complex segmented-continuous-descent approach scenario. As expected, these benefits did not emerge in the standard ILS approach, where little to no difference in performance was observed. The control activity of the thrust lever was reduced for all tasks with nxControl. This was expected, because the required energy can be selected more precisely and adequately without additional adjustments. Moreover, the controller automatically compensates for energetic disturbances. We suppose that the evident reduction in control activity also implies a reduction in physical and cognitive workload. Most pilots in our study responded very positively to the nxControl system. They could envision using it in line operations as they expect a simplification of their flight tasks. They get the ability to directly control the physical impact of drag and thrust. This simplification in manual flight seems to make the nxControl system well suited to meet the higher demands on flight precision in complex trajectories of future air traffic with an appropriate workload.

The study revealed benefits of nxControl in safety-critical situations, such as engine failure. nxControl was able to reduce the total loss of energy after an engine failure and maintain stable flight conditions. It dramatically reduced the workload in terms of lever activity and avoided critical speed drops. This workload relief enabled the pilots to focus on lateral flight-path control. No safety issues caused by nxControl were found, as in our earlier studies of engine failures [10].

The transitions from air to ground demonstrated that the nxControl concept was successfully adopted for on-ground operations, which supports the findings of earlier studies [10]. The nxControl ground mode worked as intended and was especially useful for targeted deceleration to a certain runway exit. Despite potential for refinement of the ground mode, the results showed that the nxControl concept is valuable for all flight phases from departure gate to arrival gate.

Some limitations of the present research must be considered. First, the comparison in the study is biased, since nxControl was compared to conventional manual flying that was highly trained over years of experience. Second, the tested procedure is fictional and it is pending whether its trajectory is representative for future trajectories. Third, we restricted the testing of critical situations to engine failures. Fourth, the study was performed on a simulator without motion. Fifth, the simulated aircraft was a VFW-614 ATD instead of an A320. Sixth, the ground mode was not tested in a direct comparison to conventional manual control, due to technical constraints.

Nevertheless, after five flight simulator studies with 78 pilots, the nxControl system has demonstrated such a degree of maturity that tests in a certified full flight simulator and flight tests should follow. The flight tests should investigate the system performance under various operational conditions taking atmospheric disturbances and coordination with other air traffic into account. Flight testing requires the integration of the nxControl functions into the avionic systems of a test aircraft, assumingly without unsolvable issues. In full flight simulator investigations, the pilots have the opportunity for significant longer training with nxControl before it is investigated how far flight precision, flight safety, and workload in today's and in future high-precision scenarios can be enhanced. Furthermore, it needs to be scrutinized how the long-term use of nxControl affects basic flying skills. Especially, it needs confirmation that pilots still can handle potential degradations of the flight-control system.

However, the main question "Is nxControl the right system for the cockpits of future commercial transport airplanes?" involves further considerations. The trend to a higher use of automation, including single-pilot operations, requires a new definition of the pilot's role in future cockpits. Moreover, new questions arise. Is it possible to automate the system so far that a single pilot would be sufficient to fly an airliner? Is it possible to enhance the availability of automatic flight control systems so far that degradations to manual flying become extremely improbable? Will manual flying become unnecessary or will it remain necessary that pilots exercise and maintain basic manual flying skills for the case of system degradation? If manual piloting skills remain necessary in the future, we are sure, the nxControl concept can significantly contribute.

In conclusion, we were able to investigate whether a control system based on n_x is reasonable and feasible, despite the limitations. With the proposed implementation, we showed that such a concept provides benefits for flight precision and workload in manual flight. The nxControl system completes the fly-by-wire concept with a control law for energy change while preserving handling heuristics for manual thrust control. With future higher flight-trajectory demands, nxControl would enable manual flight in daily operations, keeping pilots in the control loop. The system visualizes necessary energy cues and supports their manual flying skills.

References

[1] C. Favre, "Fly-by-wire for commercial aircraft: The airbus experience," *Int. J. Control*, vol. 59, no. 1, pp. 139–157, 1994, doi: 10.1080/00207179408923072.

[2] International Organization for Standardization, "Flight dynamics - Concepts, quantities and symbols—Part 1: Aircraft motion relative to the air, ISO1151-1," (ISO Standard No. 1151-1:1988), 1988. [Online]. Available: https://www.iso.org/standard/5699.html

[3] M. H. Amelink, M. Mulder, V. Paassen, and J. Flach, "Theoretical foundations for a total energy-based perspective flight-path display," *Int. J. Aviation Psychol.*, vol. 15, no. 3, pp. 205–231, 2005, doi: 10.1207/s15327108ijap1503_1.

[4] A. A. Lambregts, "Integrated system design for flight and propulsion control using total energy principles," in *Proc. AIAA Aircraft Des. Syst. Technol. Meeting*, 1983, pp. 1–12, AIAA-83-2561, doi: 10.2514/6.1983-2561.

[5] A. A. Lambregts, "TECS generalized airplane control system design – an update," in *Proc. Advances Aerosp. Guid. Navigation Control*, Q. Chu, B. Mulder, D. Choukroun, E. Van Kampen, C. de Visser, and G. Looye, Eds. Berlin, Germany: Springer, 2013, pp. 503–534, doi: 10.1007/978-3-642-38253-6_30.

[6] M. M. Van Paassen, C. Borst, J. Ellerbroek, M. Mulder, and J. M. Flach, "Ecological interface design for vehicle locomotion control," *IEEE Trans. Human-Mach. Interact.*, vol. 48, no. 5, pp. 541–555, Oct. 2018.

[7] T. Lambregts, R. Rademaker, and E. Theunissen, "A new ecological primary flight display concept," in *Proc. IEEE/AIAA 27th Digit. Avionics Syst. Conf.*, 2008, pp. 4.A.1–1–4.A.1–20, doi: 10.1109/DASC.2008.4702820.

[8] K. Schreiter, S. Müller, R. Luckner, and D. Manzey, "Demand control law for total energy angle tested at manual approaches," *J. Guid. Navigat. Control*, vol. 41, no. 6, pp. 1443–1448, Jun. 2018, doi: 10.2514/1.G003194.

[9] S. Müller, K. Schreiter, R. Luckner, and D. Manzey, "Manual flying and energy awareness: Beneficial effects of energy displays combined with a new approach of augmented thrust control," *Aviation Psychol. Appl. Human Factors*, vol. 7, no. 1, pp. 18–27, Jul. 2017, doi: 10.1027/2192-0923/a000111.

[10] K. Schreiter, S. Müller, R. Luckner, and D. Manzey, "nxControl: Ground mode for manual flight control laws with longitudinal load factor command," in *Advances in Aerospace Guidance, Navigation and Control*, B. Dołęga, R. Głębocki, D. Kordos, and M. Żuga, Eds. Switzerland, Cham: Springer, 2017, pp. 203–223.

[11] S. Müller, K. Schreiter, D. Manzey, and R. Luckner, "nxControl instead of pitch-and-power: A concept for enhanced manual flight control," *CEAS Aeronautical J.*, vol. 7, no. 1, pp. 107–119, 2016, doi: 10.1007/s13272-015-0169-9.

[12] K. Schreiter, S. Müller, R. Luckner, and D. Manzey, "Verbesserung von Flugpräzision und Arbeitsbeanspruchung bei manuellen RNP-Anflügen durch Vorgaberegler und Anzeigen für den Energiewinkel (nxControl)," in *D. Luft- und Raumfahrtkongress 2016, Braunschweig*. D. Ges. für Luft- und Raumfahrt, 2016.

[13] TU Berlin, Sephir - Simulator for educational projects and highly innovative research. Accessed: Aug. 15, 2017. [Online]. Available: http://www.fmra.tu-berlin.de/menue/forschung/ausstattung/sephir/

[14] R. Koenig, J. Heider, and M. Maierhofer, "Aircraft flight procedure design with respect to noise abatement as well as economical and pilot workload aspects," in *Proc. Inter-Noise Congr. Expo. Noise Control Eng.*, International Institute of Noise Control Engineering, 2005, pp. 512–521.

[15] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Human Mental Workload*, vol. 52, pp. 139–183, 1988.

[16] J. C. Byers, A. C. Bittner, and S. G. Hill, "Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary?," *Advances Ind. Ergonom. Saf.*, pp. 481–485, 1989.

[17] TU Berlin, Flight parameters and questions of the structured debriefing interview following RNP approaches to Frankfurt. Accessed Sep. 26, 2018. Available: https://www.fmra.tu-berlin.de/fileadmin/fg162/Grafiken/Projekte/DFGnxControl/RNP-Results-Addition.pdf

[18] Z. Šidák, "Rectangular confidence regions for the means of multivariate normal distributions," *J. Amer. Statistical Assoc.*, vol. 62, no. 318, pp. 626–633, 1967, doi: 10.1080/01621459.1967.10482935.

Effects of Gain and Index of Difficulty on Mouse Movement Time and Fitts' Law

Yik Hang Pang, Errol R. Hoffmann, and Ravindra S. Goonetilleke, *Member, IEEE*

***Abstract*—The mouse, being the major means of inputting and controlling data on a computer, should be set right for best performance. Mouse gain, defined as the ratio of the movement distance on the computer screen to the movement distance of the mouse, controls the movement time when both speed and accuracy are important. Reported optimum gain values vary widely from about 2 to 15. An experiment having a targeting task was used to test 16 participants using a wide range of gain (1.2, 2.4, 14.5, and 38.9). An optimum gain of 2.4 was found for a display movement amplitude (A) of 100 mm, whereas the optimum gain was 14.5 for amplitudes of 200 and 400 mm. Depending on the gain and the movement amplitude, Fitts' law should be modified to accurately model movement time (MT) in the form of MT $= A + b$(ID) $- c(A*$ID). Only at low gains could a critical index of difficulty be determined, that being about 3.5.**

***Index Terms*—Fitts' law, index of difficulty (ID), mouse gain, movement time (MT).**

I. Introduction

CONTROL-DISPLAY (CD) gain (G) is defined as the distance moved on the display relative to the movement of the manipulated device or control. Depending on the device used and the gain set, user performance can be enhanced or hindered. Touch-screen, pen-based and eye-gaze technologies are gaining popularity over mice for pointing and selection possibly due to faster movement and better accuracy [1]–[3]. These new technologies do provide added user benefits such as direct contact, but the mouse remains as a major means of inputting and controlling events on a computer as it can be tuned to a user's needs. Hence, there is a need to know the optimum settings of mice, especially since the technology used in mice have changed from a ball rolling on a surface to ones that have optical sensors.

Investigations reporting the effects of CD gain for input devices date back to Gibbs [4]. The one area in which there is some agreement in data is that performance, measured in terms of movement time to capture a target, has a "U" shape, with both low and high gains taking longer to perform the task than middle-level gains. The explanation for this is generally given as the movement time being the sum of two components: the distance-covering phase in which the cursor has to be moved to the region of the target and the accuracy phase, where the cursor has to be accurately controlled to hit the target. At low gains, the distance to the target is slow to achieve, but accuracy is easily controlled. At high gains, the time to reach the target is very short, but the limb finds difficulty in the accuracy phase due to the extreme sensitivity of the control. Thus, there is an optimum gain between those of the low and high values [5]. With the availability of touch-screen technology in tablets and computers, users prefer somewhat "direct" contact in situations where accuracy is required [6]. One parameter that makes the mouse disadvantaged in high-accuracy situations is possibly its gain.

The optimum values for the mouse gain has been investigated by many researchers [5]–[11] (see Table I) mainly in a Fitts form of task [12], [13] but also in tracking tasks [6]. The optimum values of gain range from about 2 to around 15, depending on various experimental factors, such as the resolution of the screen (which may cause a problem with quantization), the size of the mouse pad (causing the use of "clutching," which is holding, lifting, and moving the mouse over the pad so that there is no display cursor movement), the reporting sensitivity of the mouse, and the type of task being performed. The most detailed research is that of Casiez *et al*. [5], where they investigated constant gain and also gains with an acceleration function that is standard in most operating systems. We restricted our research to cases of constant gain within movements, with no acceleration or gain changes with mouse movement. Casiez *et al*. also found differences in mouse gain for differing monitor resolution. With a high-resolution monitor, the movement times appeared to level out at gains of about 15, compared to values of 4 to 8 for the lower resolution monitor. In both cases, however, there was little variation in movement time as gain increased above these values. The movement time (MT) versus gain relationship was largely "L" shaped with increasing gain. However, the effect of gain on ballistic and visual control movements has not been investigated even though the applicability of Fitts' law in computer applications is abundant.

Fitts' law [12], [13] is given as follows:

$$\mathrm{MT} = a + b\,(\mathrm{ID}) \;\text{ where ID} = \log_2\left(\frac{2A}{W}\right) \tag{1}$$

where A is the amplitude of movement, W is the target width, a and b are constants. Kvalseth [14] proposed an alternative for

Manuscript received July 16, 2018; revised March 27, 2019 and May 13, 2019; accepted June 30, 2019. Date of publication September 9, 2019; date of current version November 21, 2019. *(Corresponding author: Ravindra S. Goonetilleke.)*

Y. H. Pang was with the Human Performance Laboratory, Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Hong Kong. He is now with Walnut Technology, www.walnutt.com (e-mail: henri.peng@gmail.com).

E. R. Hoffmann was with the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Hong Kong (e-mail: erroldot@tpg.com.au).

R. S. Goonetilleke is with the Division of Integrative Systems and Design, and the Human Performance Laboratory of the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Hong Kong (e-mail: ravindra@ust.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2931743

TABLE I
OPTIMUM MOUSE GAINS REPORTED IN PRIOR RESEARCH

Authors	Type of task	Experimental conditions	Optimum mouse gain	Comments
Jellinek and Card [7]	Fitts	G=1,2,4,8, 16,32 ID=1.6 to 5	2	Clutching at low gain Quantization at high gain
Lin et al. [8]	Fitts	G=.5, 1, 2, 4 ID=1, 2.5, 4, 5.4	2	
Johnsgard [9]	Fitts	G=1,2,3 ID = 1 to 4	No optimum over range of gain	
Bohan et al. [10]	Fitts	G=1,2,4,8	4 with wrist moves; 2 for fingers.	Limb components used in movements considered.
Sandfeld and Jensen [11]	Fitts	G=2,4,8 ID=5	4 for young participants	
Casiez et al. [5]	Fitts	G=1,2,4,6, 8,12 ID=4.5 to 8.5 (5 values)	4 to 8 (Expt 1) Approx 15 (Expt2)	Expt 2 used a large hi-resolution screen
Senanayake & Goonetilleke [6]	Tracking	G=2.3, 5,10,15 (A/W)=3. 6, 6.25, 8.3,14,25, 33	12 (ballistic) 9 (visually-controlled)	

MT, $(= aA^bW^c)$, which is supposed to be a more generalized form with direct dependencies on both *A* and *W*.

Equation (1) is valid for aimed movement with continuous visual control. At low values of index of difficulty (ID), movements of the mouse are likely to be made in a ballistic manner without continuous visual feedback. If ballistic movements can be made, considerable reductions in the capture time of a target can be achieved. For such movements, the time is independent of the width of the target and is solely dependent on the distance moved. Then, MT fits the equation proposed by Gan and Hoffmann [15]

$$\mathrm{MT} = a + b\sqrt{A}. \tag{2}$$

For those cases, the stopping error measured as the variability of aiming in the direction of movement [16] will give further validation on the effect of distance moved. The value of ID at which aimed movements requiring continuous visual control changes to ballistic movements is known as the "critical" ID ($\mathrm{ID_{crit}}$). For the experiment in [15]

$$\mathrm{ID_{crit}} = 2.58 + 0.101\sqrt{A} \tag{3}$$

where *A* is in millimeters. This $\mathrm{ID_{crit}}$ is very useful in determining the change-over from ballistic to visual control so that the appropriate equation [(1) or (2)] can be used.

$\mathrm{ID_{crit}}$ has been reviewed for many different body components and motions by Hoffmann [17]. In that review, it was found that there was a lack of information for the $\mathrm{ID_{crit}}$ when the mouse is controlled by an arm motion rather than just the wrist. With a relatively large range of ID, it is possible to determine the $\mathrm{ID_{crit}}$ below which participants are able to perform a movement in a ballistic manner. Thus, the objective of this study is to examine the effect of gain on $\mathrm{ID_{crit}}$, stopping error, and the formulations of the two regimes of aimed movement.

II. METHOD

In this article, we use Fitts' paradigm [12], [13] to test the effects of changes in gain between mouse input and screen output. The two equations (1) and (2) for movement time will form the basis for the modeling of the time taken to move with varying amplitude of movement (*A*), ID, and gain.

A. Participants

Sixteen volunteer university students from the Hong Kong University of Science and Technology were recruited to participate in this study. They were between 20 and 31 years (mean age: 23.19, SD: 2.64) in age. Eight were male and eight female. They were fully informed of the purpose of the experiments and took part under the ethical guidelines of the Hong Kong University of Science and Technology. All were competent computer users, having an average usage of 6 to 7 h per day. None had any physical difficulties that may have affected the experiment and had normal or corrected to normal vision.

B. Apparatus

The basic components were a desktop computer with a CPU clock rate of 2.19 GHz with Windows 8 operating system, and custom software developed with C++ programming language. The monitor was a capacitive touch screen monitor (58.0 cm diagonally) for display. A Logitech G9X high-precision (5700 dpi) optical mouse and a mouse pad with dimensions of 300 mm in breadth and 800 mm width were used for this experiment. The mouse acceleration function was turned OFF so that there was a direct linear relationship between mouse and screen movements.

The mouse motion settings in Windows 8 can be adjusted on a scale of 0 to 10. Scale (slider) values of 1, 2, 5, and 10 were chosen for this experiment. In this study, CD gain (*G*) was constant within movements although gain was varied between conditions. CD gain cannot be set directly, but can be measured or calculated for a given slider value (*L*) and related sensitivity ratio of the mouse (*R*), mouse DPI (*D*), screen size (L_s for screen length, W_s for height in mm), and resolution (P_h for number of pixels horizontally, P_v for number of pixels vertically). Then, pixel size $= L_s/P_h = W_s/P_v$. *R* is a function of *L*. For $L = 10$, $R = 1$. At $L = 1$, $R = 0.03125$, etc.

$$G = \frac{R \times D \times L_s}{P_h} = \frac{R \times D \times W_s}{P_v}. \tag{4}$$

The above equations were validated with actual measurements of mouse movement and cursor movement on the display. With the mouse dpi settings and sampling rate, this gave actual CD gains of 1.2, 2.4, 14.5, and 38.6.

A range of ten ID values, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, and 6.0, were used along with the four display movement amplitudes of 40, 90, 160, and 250 mm. These IDs and amplitudes were to allow an investigation of the effects of arm inertia: at the low gain values, the movement of the arm would be similar to that for arm movements in [15], while at high gains the effect

of arm inertia would be reduced due to the smaller movement amplitudes and accelerations. This allows the dominant effect of inertial and cognitive to be evaluated when manipulating a mouse.

For each ID, the target widths were calculated at the four amplitudes from (1). A mouse pad of sufficient length was used so that, at the lowest gain of 1.2, the mouse could remain in contact with the pad during all movements. In other words, it was not necessary for the participant to use clutching of the mouse during such movements where the participant has to lift the mouse and move it without a change in display cursor.

C. Procedure

Participants sat at a computer desk and with an adjustable chair had their elbow height about 15 cm above the table height. The monitor was around 30 cm in front of the participant. The mouse pad was placed in between subject and monitor. The four gains were presented to participants in a Latin square design on four different days. Within each gain, the 40 conditions of amplitude and ID were given in a different random order for each of the participants. Participants were allowed practice on each condition until they felt confident that they could perform the tasks without error. The practice was followed by ten measured trials, and breaks were given to participants whenever requested. If more than one target was missed during the ten measured trials, the full condition was repeated at the end of the set of trials of that condition ensuring 100% accuracy.

The task involved each participant moving the display cursor from a start point that turned green as soon as it was clicked to a rectangular target of 100 mm height at the tested distance, using a high-precision mouse. Each participant was asked to move as quickly and accurately as possible to click the approximate center area of the target. Once the participant clicked the target area, the movement path was shown, and the software program recorded the movement path, location of the endpoint, and the time taken. If the participant failed to click inside the target area, that trial was repeated at the end of the experiment.

III. Results

Participants were allowed practice until they felt confident that they could perform the tasks without error. That does not imply that they had stable performance in the experimental trials. An initial analysis of the movement times as a function of trial number showed that the first two trials had significantly higher movement time than later trials, which showed no further effect of trial number. Thus, the first two trials were ignored, and data analysis was made on the last eight trials of the set.

Mean movement time data for gains of 1.2, 2.4, 14.5, and 38.6 for each ID are shown in Fig. 1. Of significance is the effect of gain as seen in Fig. 1. At low gain, the data appear similar to those for simple hand/arm movements, with a strong separation of the effects of amplitude and ID at low ID values [15]. This disappears at high gains where the data at low ID all tend to merge to a common line. At the higher gains, the effect of amplitude also becomes less, apart from the smallest amplitude of 40 mm, where performance is significantly different to the higher amplitudes.

Repeated-measures ANOVA of the movement times with factors of amplitude A (4), gain G (4), ID (10), and participants

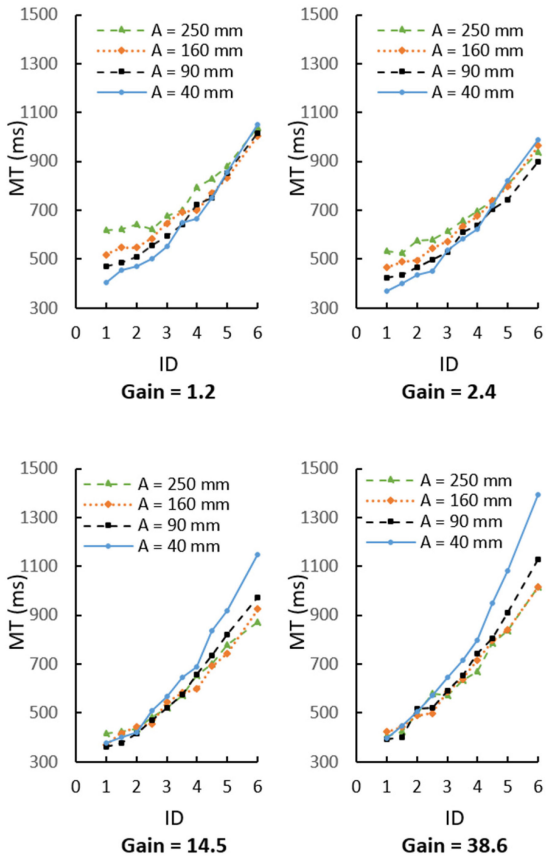


Fig. 1. Mouse movement time as a function of index of difficulty and movement amplitude for gains of 1.2, 2.4, 14.5, and 38.6.

(16) showed main effects of A [$F(3, 45) = 10.14$, $p < .001$, $\eta_p^2 = 0.072$]; G [$F(3,45) = 192.9$, $p < .001$, $\eta_p^2 = 0.330$] and ID [$F(9,135) = 465.91$, $p < .001$, $\eta_p^2 = 0.932$]. There were also significant interactions between all three factors, $A \times G$ [$F(9,405) = 48.63$, $p < .001$, $\eta_p^2 = 0.329$]; $A \times$ ID [$F(27, 405) = 19.14$, $p < .001$, $\eta_p^2 = 0.307$] and ID $\times G$ [$F(9, 405) = 10.60$, $p < .001$, $\eta_p^2 = 0.253$]. The triple interaction between the three experimental factors was also significant [$F(81, 1215) = 1.54$, $p = .002$, $\eta_p^2 = 0.093$]. According to Cohen's categories [18], the effect size of ID is large, those for G, $A \times G$, $A \times$ ID are small to medium, while those of A, ID $\times G$ and ID $\times G \times A$ are small. The interactions were analyzed further and reported later.

The triple interaction is difficult to interpret due to the large number of levels of factors involved. However, some indication of the triple interaction can be seen in Fig. 1. At small gains, close to those for purely manual input without an intervening mouse, the data appear similar to those of Gan and Hoffmann [15] for arm movements, with a clear separation of the lines of varying amplitude at low ID values. In fact, at low ID, the times for movement are approximately linear in the square-root of movement amplitude, as in arm movement data (see Fig. 2).

For the low gain ($G = 1.2$) and lowest ID $= 1.0$, regression in terms of (2) gives

$$\mathrm{MT}_{(G=1.2,\mathrm{ID}=1)} = 258 + 22.2\sqrt{A}; \quad r^2 = 0.94, p = 0.02. \tag{5}$$

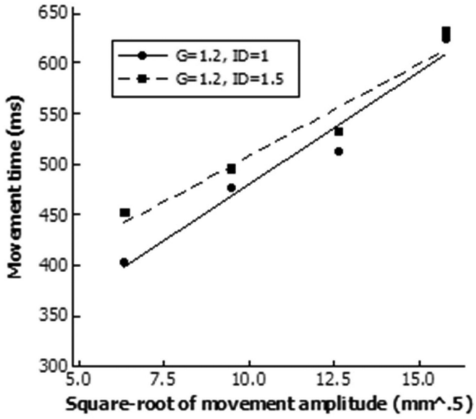


Fig. 2. Movement time at the lowest ID = 1.0 and lowest gain ($G = 1.2$) where ballistic movements were made, with conditions similar to the arm movements of Gan and Hoffmann [15].

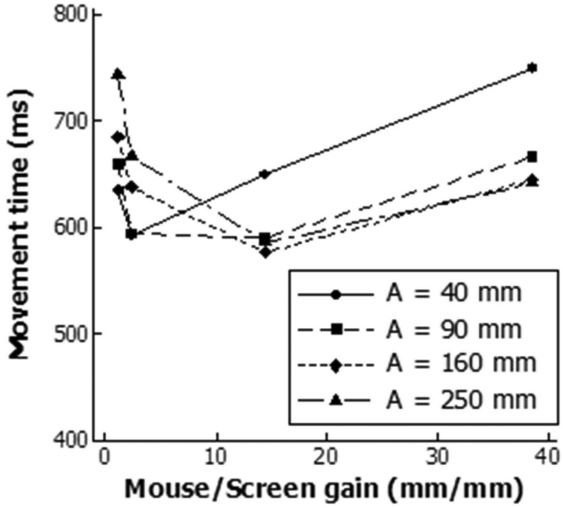


Fig. 3. Amplitude $*$ gain interaction illustrating the optimum gains for the various movement amplitudes.

A. Movement Times for ID ≥ 3

Movements at ID values greater than about three are generally made with ongoing visual control [15]. The linearity of MT with ID for each of the control/display gains is seen in Fig. 1. Given the different formulations involving ID, amplitude, and square-root amplitude, stepwise regressions of the data for the four gains, including terms for ID, square-root of movement amplitude and the interaction between amplitude and ID, gave the results in (6)–(9). The ($A * \mathrm{ID}$) interaction was included as a result of its significance in the MT ANOVA. Stepwise regression is where the predictor variables enter the model one at a time based on the strength of the relationship between the response and the predictor variables. The added explained variance is obtained from the change in the r^2 value. Such a procedure is more robust compared to a simultaneous multiple regression analysis where the model can be underspecified and misleading

$$\mathrm{MT}_{1.2} = 116 + 136\,(\mathrm{ID}) + 6.7\sqrt{A}; \quad r^2 = 0.94 \tag{6}$$

$$\mathrm{MT}_{2.4} = 122 + 127\,(\mathrm{ID}) + 4.1\sqrt{A}; \quad r^2 = 0.95 \tag{7}$$

$$\mathrm{MT}_{14.5} = 63 + 170\,(\mathrm{ID}) - 0.13\,(A * \mathrm{ID}); \quad r^2 = 0.95 \tag{8}$$

$$\mathrm{MT}_{38.6} = 29 + 206\,(\mathrm{ID}) - 0.19\,(A * \mathrm{ID}); \quad r^2 = 0.94. \tag{9}$$

For gains of 1.2 and 2.4, there was a small, but significant, effect of the amplitude of movement apart from that of ID on MT, accounting for 3% and 1% of variance, respectively. The addition of the $\sqrt{A}$ in the regressions for the low gain cases is a modification of Fitts' law that has been found necessary in a number of cases, where there is an effect of amplitude apart from that in ID. It arises from differences in the information processing rates of the distance-covering and homing-in phases of the movement [19] and is best described in this form rather than in logarithmic terms for the amplitude and target width [20].

At the higher gains of 14.5 and 38.6, the interaction terms became significant accounting for 9% and 12% of the regression variance as shown by the added step of the addition of the interaction in the stepwise regression. The regression at the higher gains show an effect not previously reported in that the MT is affected by an interaction between ID and movement amplitude, A. Thus, the common use of Fitts' law to describe data for mouse movement times needs modification to take into account this interaction. This effect is one further modification of Fitts' law found necessary, especially at higher gains. The smaller amplitude movements had a higher movement time than the larger amplitudes at the same ID value (see Fig. 1).

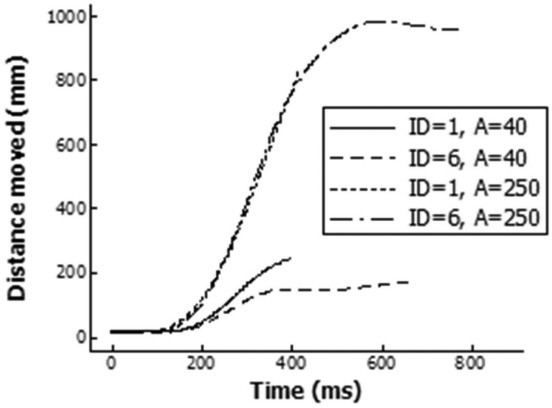
B. Two-Way Interactions

All three two-way interactions of (Amplitude $\times$ Gain), (Amplitude $\times$ ID), and (Gain $\times$ ID) were significant at $p < .001$. These interactions are discussed below.

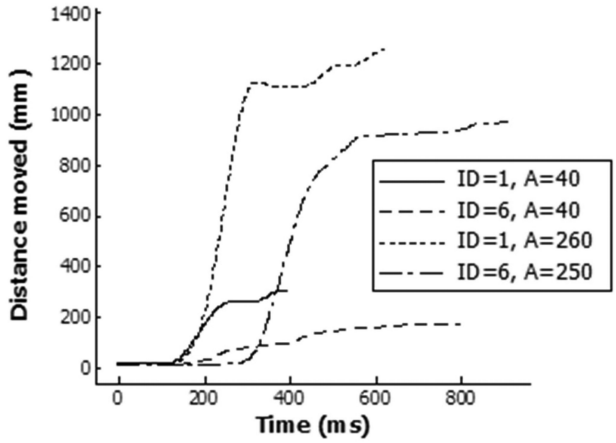
1) Amplitude $\times$ Gain Interaction: There is a clear pattern of there being an optimum value of control/display gain for each amplitude of movement (see Fig. 3). At the smallest amplitude (40 mm), the optimum gain was 2.4; at all other amplitudes, the optimum was 14.5. These were confirmed with a simple-effects analysis. Tukey HSD post-hoc tests showed that at an amplitude of 40 mm, the optimum gain of 2.4 had a significantly lower movement time than gains of 38.6 ($p < .01$) and 1.2 ($p < .05$) and gain of 38.6 had higher movement times than 1.2 and 2.4 ($p < .01$). At an amplitude of 90 mm, all differences between gains of 14.5, 1.2, and 38.6 were significantly different ($p < .01$), while at amplitude = 160 mm, only gain 14.5 was significantly faster than all other gains ($p < .01$). The pattern at a gain of 38.6 was different in that all comparisons, apart from that between gains of 2.4 and 38.6, were significantly different at $p < .01$.

These optimum values can be interpreted in terms of the displacement versus time plots for typical mouse movements. At the lowest gain, the time to reach the target area was high, while the settling time as the target was approached was low. At high gains (38.6), the time to reach the target area was small, but the settling time, due to the high sensitivity of the control, was considerably longer. Typical displacement versus time graphs is shown in Fig. 4 for Participant 8 and gains of 1.2 and 38.6.

2) Amplitude $\times$ ID Interaction: This interaction is illustrated in Fig. 5. The interaction was analyzed with Tukey HSD tests:



(a) Gain = 1.2 Subject 8



(b) Gain =38.6. Subject 8

Fig. 4. Sample displacement versus time profiles for Participant 8 with (a) gain of 1.2 and (b) gain of 38.6.

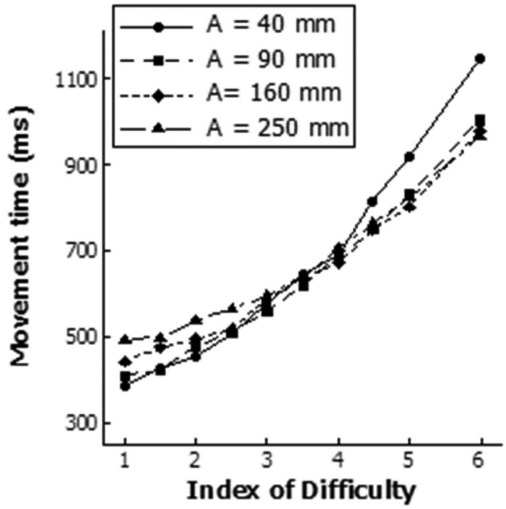


Fig. 5. Amplitude $*$ index of difficulty interaction for mouse operation (averaged over mouse gains).

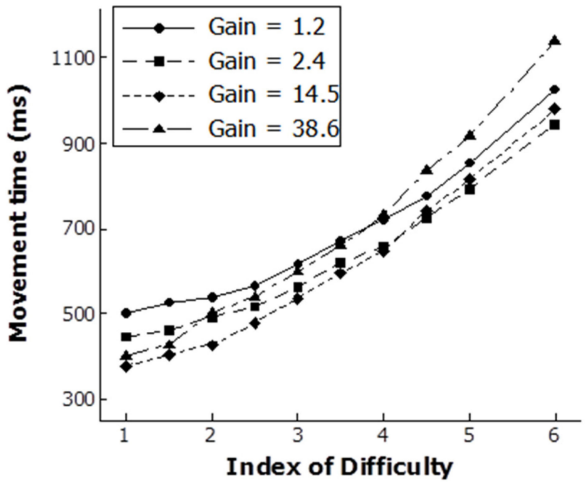


Fig. 6. Gain $*$ index of difficulty interaction (averaged over movement amplitudes).

At high ID values (4.5 to 6.0), the 40-mm amplitude movement takes longer than other amplitudes ($p < .01$); there is no significant effect of amplitude for IDs of 3 to 4.0. At low ID values, the MT at ID $= 1.0$ is shorter for the 40-mm amplitude than other amplitudes ($p < .01$); at ID $= 1.5$, only two comparisons are nonsignificant (40, 90 mm) and (90, 160 mm); other comparisons are significant at $p < .01$. At an ID $= 1.5$, the (160, 250 mm) and (40, 90 mm) are not significantly different, whereas all others are significant at $p < .01$. The crossover of the effects of ID with changing amplitude is seen in Fig. 5.

3) Gain $\times$ ID Interaction: The interaction is illustrated in Fig. 6. Tukey post-hoc tests showed a pattern, at IDs of 1.0 and 1.5, of gain rank order 14.5, 38.6, 2.4, and 1.2 for lowest to highest MT, with significant differences occurring between the 1.2 gain and all others (at least $p < .05$) and between 1.2 and 2.4 ($p < .01$). At the higher IDs (5 and 6), the rank order changed to 2.4, 14.5, 1.2, 38.6, with 38.6 having a higher movement time than other gains ($p < .01$) and 2.4 having a lower MT than 1.2 ($p < .01$).

C. Effect of Defined Target Area on Stopping Errors at Low ID

For ballistic movements, which are expected at the low IDs, Hoffmann [21] showed by means of a sensitivity analysis of the ballistic movement time equation, that end point variability is linearly related to the amplitude of the movement. The analysis showed that both the variability in force and the force application time (or impulse) affected the variability of the movement endpoint. The model was strongly supported by the research of Lin and Tsai [22]. Note, however, that this analysis was for movements where there was no defined target area.

We measured the end-point variability in this experiment and hence are able to determine the factors that affect variability when there is a defined target area, even though with the low ID values (less than 3) we might expect that there is a negligible effect due to the relatively large targets. Stepwise regression of the effects of amplitude, target width, and ID for the various gains yielded quite different models for the end-point variability at low ID, indicating a large effect of the presence of a target area that does not occur without a defined target, even though there is little final constraint to the movement. The best regressions were

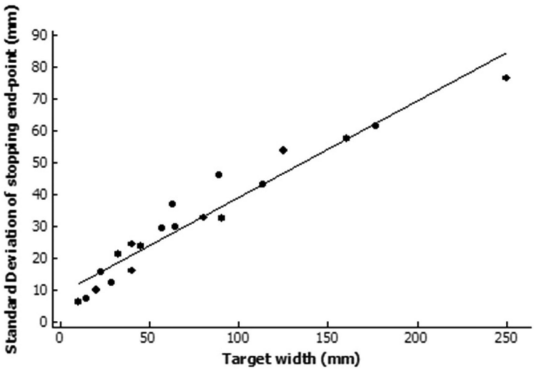


Fig. 7. Gain = 1.2. Standard deviation of the stopping end-point at the target.

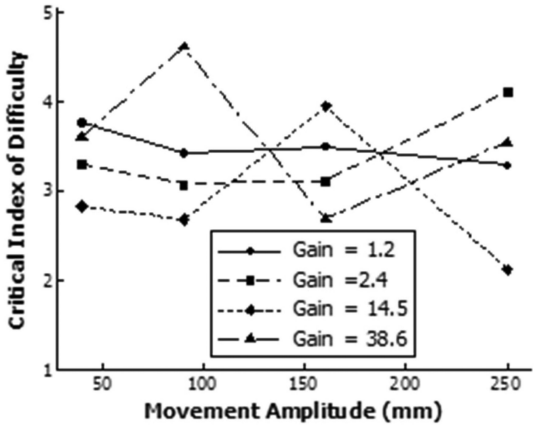


Fig. 8. Critical index of difficulty as a function of amplitude of movement and control/display gain.

Gain = 1.2:

$$\mathrm{SD}\,(1.2) = 9.74 + 0.185\mathrm{W} + 0.103\mathrm{A};$$
$$r^2 = 0.984, p < 0.001. \quad (10)$$

Target width accounted for 93.2% variance and amplitude 5.2%.

Gain = 2.4:

$$\mathrm{SD}\,(2.4) = -16.5 + 20.3\mathrm{ID} + 0.075\mathrm{A};$$
$$r^2 = 0.51, p < 0.001. \quad (11)$$

ID accounted for 43.6% of variance and amplitude 7.5%.

Gain = 14.5:

$$\mathrm{SD}\,(14.5) = -5.13 + 22.4\mathrm{ID}; \quad r^2 = 0.47, p < 0.001. \quad (12)$$

Gain = 38.6:

$$\mathrm{SD}\,(38.6) = -8.81 + 28.7\mathrm{ID}; \quad r^2 = 0.35, p < 0.01. \quad (13)$$

Thus, the end-point stopping errors changes from being largely determined by the target width at low gains (see Fig. 7) to being only mildly dependent on the ID of the movement at high gains. At low gains, participants are using the available target area and not having a stopping error governed by the natural variability in movement impulse, as would be expected for purely ballistic movements [21].

D. Critical ID

The ID below which movements could be made in a ballistic manner is called the critical ID [17]. This critical value is most clearly seen in the data of Gan and Hoffmann [15] for arm/hand movements, where there was a leveling out of the movement time at low ID and the times became dependent only on the amplitude of movement. They determined the critical ID to be given by (3). For $A = 100$, $\mathrm{ID}_{\mathrm{crit}} = 3.59$. As noted by Hoffmann [17], there was inadequate data on which to determine a critical value when participants are using a mouse for input to a computer. Due to the form of the experimental data in the present case, it is difficult to determine a value of the $\mathrm{ID}_{\mathrm{crit}}$. This is because the movements at low ID combined with high gain do not level out as in data for other limb movements. If we take an intersection point of the low ID movement times with those for high IDs, an approximate value may be obtained. Using this method, the $\mathrm{ID}_{\mathrm{crit}}$ values for each of the gains are as shown in Fig. 8. There is little pattern in these data; the best that can be said is that at a low gain of 1.2, the critical ID is about 3.5, which matches with the model of Gan and Hoffman [15]. However, at the high gains there is no clear pattern, and it appears that all movements, down to the lowest ID values, were made with ongoing visual control. These results are in contrast with those of Gan and Hoffmann, where it was possible to define a clear boundary between the ID values to make ballistic movements and those where ongoing visual control was required to capture the target.

III. Discussion

A. Gain Effects

The present experiment has covered a wide range of ID and mouse/screen gain in a design that has not been restricted by the use of clutching in mouse use or by quantization due to lack of screen resolution. Thus, we have been able to show clear values of optimum mouse gain as a function of the amplitude of movement, this being around 14.5 for movement amplitudes of 90, 160, and 250 mm and 2.4 for the smaller amplitude of 40 mm (see Fig. 3). A higher gain is primarily advantageous to move larger distances. With shorter distances, a higher gain will affect the targeting as more submovements will be needed to click on the target. This pattern of the effects of CD gain are well established for many different forms of arrangement [5], yielding an optimum gain for each of the amplitudes of movement.

It is suggested that the interaction term ($A * \mathrm{ID}$) arises from the structure of the movements at high gain. This interaction is essentially one between the distance-covering and homing-in phases of the movement as the amplitude term relates to the rapid phase of the movement, whereas the ID term relates to the slower end-phase of the movement as the target is approached and captured. This interaction thus incorporates the effects of high gains as distinct from the model at low gains where the effects, apart from that of the ID, are likely to be due to arm mass-moment-of-inertia. The effect of the term is to reduce the movement time as the product ($A * \mathrm{ID}$) increases. Thus, at a constant task difficulty (ID), increasing amplitude decreases MT as does increasing ID at a constant amplitude. It is difficult to

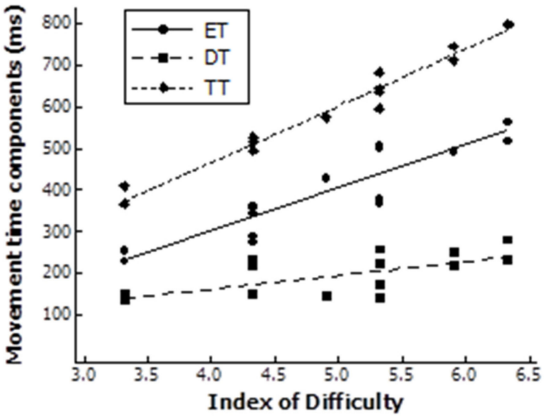


Fig. 9. Data of Walker *et al*. [24] for their Experiment 3, showing the times for moving to the target (ET), the mouse button time (DT) and the total time (TT).

explain these effects in terms of the structure of the movement without detailed measures of the submovements used at varying amplitude and ID.

At low gains, full arm movements were required to move the cursor to the required screen locations. Thus, arm inertia forces were involved as in arm moves without a mouse [15]. Note that the times given by (5) are much larger than those for the arm movements of Gan and Hoffmann [15]

$$\mathrm{MT}_{(G=1,\ \mathrm{ID}=1)} = 72 + 7.59\sqrt{A} \tag{14}$$

where A is in mm.

The reason for this difference lies, at least partly, in that at the end of a movement in the present experiment, the participant was required to make a button click when the cursor was within the target boundary. Note also that the times for these "ballistic" movements well exceed those beyond which the visual-control mechanisms could be used to improve final accuracy. For example, at $A = 100$ mm, the difference in MT is 313 ms after adjusting for the gain of 1.2 in the current experiment. The difference can be explained in two possible ways. With the keystroke model, a "preparation" or planning time of around 200 ms and a mouse click time of 113 ms can account for this difference when compared to the physical task where the touch is anticipated [23]. Thus, if these mouse movements at low ID are made ballistically, it is necessary to account for the time taken for the mouse button input. The second way is using Welford's method [19], where he discusses issues of this type and considers the time taken to be related to the difficulty of the task just completed. It is described as "decision clearance time." There is some research available to quantify the time taken [24]. Re-analysis of the Walker *et al*. data for their Experiments 1 and 3 shows that the time taken for this component of the task is reasonably linear in target width of the movement just completed, accounting for 88% of the variance in dwell time (DT). Data from their Experiment 3 are shown in Fig. 9. This result is in agreement with those of Chen *et al*. [25].

In its simplest form, the DT (ms) was given by

$$\mathrm{DT} = 330 - 28.9\,\mathrm{W};\ r^2 = 0.88, p < 0.001. \tag{15}$$

Subtracting these values from those of the actual total times for the mouse movements at low IDs puts the movement times to values that are in agreement with Fig. 1 of Gan and Hoffmann [15, p. 833].

B. Why Is There a Difference Between Experiments?

Previous research, such as that of Lin *et al*. [8], have found consistent results for the optimum gain for mouse input at about G = 2. Also, Casiez *et al*. [5] have found optimum values at much higher gains—approximately 4 to 8 for a standard screen and 15 for a high-resolution screen. Casiez *et al*. have pointed out that some of the differences have arisen because of the differences in apparatus. This is apparent from the differences found by those authors in their two experiments, where all that was changed was the resolution and size of the monitor.

Lin *et al*. [8] used a monitor with a resolution of .41 mm/pixel with an active area of 235 mm horizontally. The cursor was sampled at 32 Hz. Movements were discrete. This is compared with the Experiment 1 conditions of Casiez *et al.* [5] of 100 dpi monitor (.25 mm/pixel) and a mouse sampled at 60 Hz. In their Experiment 2, the monitor size was 4.7 m horizontally with a resolution of 1 mm/pixel. In both experiments, reciprocal movements were made between targets. Although not stated, the participants obviously had to be seated a large distance from the screen when using this monitor, hence the visual angle subtended by each pixel cannot be compared for the two experiments. The larger screen allowed the mouse to be used to high gains without quantization of the pixels (all could be selected). In the current experiment, the monitor had a resolution of .265 mm/pixel and a mouse reporting rate of 500 Hz.

With the number of differences in experimental equipment, it is difficult to determine the reasons for the great differences in optimum gains. In the Casiez *et al*. experiments, there was clutching at the low gains, whereas this was not occurring in either the Lin *et al*. or current experiments, due to the gains used and the size of mouse pads. With the higher mouse reporting rates of the current experiment it is more likely that effects at the higher gains would be detected than in the Casiez experiments. However, it is still difficult to account for the consistent and low optimum gains of the Lin *et al*. experiments.

C. Critical ID

At low gains, the pattern of MT versus ID was similar to that of Gan and Hoffmann [15], showing a leveling out of the movement time at low IDs. It would appear that at these low gains, the effect of the low ID along with the arm mass moments of inertia [as seen in the ballistic (2)] dominate the movement time at low ID. This is as expected and, although the MTs in this experiment are higher than of those in [15] for arm movements, the difference may be accounted for by the time required to click the mouse to signify the completion of the movement.

It was only at these low gains that there was a clear method by which to determine the critical ID by finding the intersection of the regression of the low ID data points (the ballistic movements) with the regression for the high ID data points (movements made under visual control). With this method, the critical ID came to be about 3.5 for mouse control, and relatively independent

of the amplitude of the movement. In this respect, the data are different to [15] for arm movement, where there was a linear relationship between the critical ID and the square-root of movement amplitude.

At higher gains, the similarity of pattern with arm movements disappeared, with no clear leveling of the movement times at low ID values; in fact, Fitts' law appeared to be valid to low values of ID and were independent of the amplitude of the movement. An attempt to find a critical ID and amplitude (see Fig. 8) showed no clear pattern. Thus, movements at high gain appear to have all required the use of ongoing visual control. This may be accounted for by the extreme sensitivity of the control—the first phase of the movement of distance to the target is very rapid and, due to control sensitivity, much more time is required in the homing-in phase of the movement to avoid excessive under- and overshooting as the target is captured. Thus, the major portion of the movement time is spent using closed-loop control and hence the MT versus ID plot appears linear even at low ID values.

IV. Conclusions and Limitations

The MT analysis revealed that there is a clear pattern showing an optimum value of control/display gain for each amplitude of movement. At the smallest amplitude (40 mm), the optimum gain was 2.4; at all other amplitudes, the optimum was 14.5.

For gains of 1.2 and 2.4, there was a small, but significant, effect of the amplitude of movement apart from that of ID on MT. At the higher gains of 14.5 and 38.6, the interaction of amplitude and ID became significant, accounting for 9% and 12% of the regression variance. Thus, these are important terms that should be accounted for in the Fitts' model as follows:

$$\mathrm{MT} = a + b\mathrm{ID} - c\,(A * \mathrm{ID})\,. \tag{16}$$

There are some limitations in this experiment. A high-precision mouse was used as opposed to a mouse that is normally used with a computer. It is assumed that the behavior of this type of mouse is similar to a mouse that is used for general applications. The reason for using such a high-precision mouse was to be able to generate a higher CD gain from the higher mouse gain. Also, clutching was not allowed by having a large mouse pad. This is normally not the case when using a mouse, even though under controlled conditions and for experimental purposes it may be a norm. Any form of clutching will no doubt increase movement time, but in terms of the aspects investigated, it would have no bearing.

References

[1] S. J. V. Nichols, "New interfaces at the touch of a fingertip," *Computer*, vol. 40, no. 8, pp. 12–15, Aug. 2007.

[2] A. Jain, D. Bhargava, and A. Rajput, "Touch-screen technology," *Int. J. Adv. Res. Comput. Sci. Electron. Eng.*, vol. 2, no. 1, pp. 74–78, 2013.

[3] Y. Tan, G. Tien, A. E. Kirkpatrick, B. B. Forster, and M. S. Atkins, "Evaluating eyegaze targeting to improve mouse pointing for radiology tasks," *J. Digit. Imag.*, vol. 24, no. 1, pp. 96–106, Feb. 2011.

[4] C. B. Gibbs, "Controller design: Interactions of controlling limbs, time-lags and gains in positional and velocity systems," *Ergonomics*, vol. 5, no. 2, pp. 385–402, 1962.

[5] G. Casiez, D. Vogel, R. Balakrishnan, and A. Cockburn, "The impact of control-display gain on user performance in pointing tasks," *Hum.-Comput. Interact.*, vol. 23, no. 3, pp. 215–250, Sep. 2008.

[6] R. Senanayake and R. S. Goonetilleke, "Pointing device performance in steering tasks," *Perceptual Motor Skills*, vol. 122, no. 3, pp. 886–910, Jun. 2016.

[7] H. D. Jellinek and S. K. Card, "Powermice and user performance," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1990, pp. 213–220. [Online]. Available: https://dl.acm.org/citation.cfm?id=97276

[8] M. L. Lin, R. G. Radwin, and G. C. Vanderheiden, "Gain effects on performance using a head-controlled computer input device," *Ergonomics*, vol. 35, no. 2, pp. 159–175, Feb. 1992.

[9] T. Johnsgard, "Fitts' law with a virtual reality glove and a mouse: Effects of gain," in *Proc. Grap. Interface*, 1994, pp. 8–15.

[10] M. Bohan, S. Thompson, D. Scarlett, and A. Chaparro, "Gain and target size effects on cursor-positioning time with a mouse," *Proc. Hum. Factors Ergonom. Soc.*, vol. 47, no. 4, pp. 737–740, Oct. 2003. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/154193120304700416

[11] J. Sandfeld and B. R. Jensen, "Effect of computer mouse gain and visual demand on mouse clicking performance and muscle activation in a young and elderly group of experienced computer users," *Appl. Ergonomics*, vol. 36, no. 5, pp. 547–555, Sep. 2005.

[12] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *J. Exp. Psychol.*, vol. 47, no. 6, pp. 381–91, Jun. 1954.

[13] P. M. Fitts and J. R. Peterson, "Information capacity of discrete motor responses," *J. Exp. Psychol.*, vol. 67, pp. 103–12, Feb. 1964.

[14] T. O. Kvalseth, "An alternative to Fitts' law," *Bull. Psychonomic Soc.*, vol. 16, no. 5, pp. 371–373, 1980.

[15] K.-C. Gan and E. R. Hoffmann, "Geometrical conditions for ballistic and visually controlled movements," *Ergonomics*, vol. 31, no. 5, pp. 829–839, May 1988.

[16] C. I. Howarth and W. D. A. Beggs, "The control of simple movements by multisensory information," in *Motor Behavior*. Berlin, Germany: Springer, 1985, pp. 125–151.

[17] E. R. Hoffmann, "Critical index of difficulty for different body motions: A review," *J. Motor Behav.*, vol. 48, no. 3, pp. 277–288, 2016.

[18] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY, USA: Routledge, 1988.

[19] A. T. Welford, *Fundamentals of Skill*, 2nd ed. London, U.K.: Methuen, 1968.

[20] A. H. S. Chan and E. R. Hoffmann, "Effect of movement direction and sitting/standing on leg movement time," *Int. J. Ind. Ergonom.*, vol. 47, pp. 30–36, May 2015.

[21] E. R. Hoffmann, "Fitts' law with an average of two or less submoves?" *J. Motor Behav.*, vol. 48, no. 4, pp. 319–331, 2016.

[22] R. F. Lin and Y. C. Tsai, "The use of ballistic movement as an additional method to assess performance of computer mice," *Int. J. Ind. Ergonom.*, vol. 45, pp. 71–81, Feb. 2015.

[23] S. K. Card, T. P. Moran, and A. Newell, *The Psychology of Human-Computer Interaction*, 1st ed. Hillsdale, NJ, USA: Erlbaum, 1983.

[24] N. Walker, D. A. Philbin, and A. D. Fisk, "Age-related differences in movement control: Adjusting submovement structure to optimize performance," *J. Gerontol. Ser. B, Psychological Sci. Social Sci.*, vol. 52, no. 1, pp. 40–53, 1997.

[25] Y. Chen, E. R. Hoffmann, and R. S. Goonetilleke, "Structure of hand/mouse movements," *IEEE Trans. Hum.-Mach. Syst.*, vol. 45, no. 6, pp. 790–798, Jun. 2015.

ICHMS 2020 Call for Papers

1st IEEE International Conference on Human-Machine Systems

April 6-8, 2020, Rome, Italy

Financial and Technical Sponsorship: IEEE Systems, Man, and Cybernetics Society

Conference Theme: Human-Centered Cyberphysical Systems

http://ichms.dimes.unical.it/

The 2020 IEEE International Conference on Human-Machine Systems (ICHMS) will be held in Rome, Italy. ICHMS will provide a remarkable opportunity for the academic and industrial communities to address new challenges and share solutions, and discuss future research directions in the important area of Human-Machines Systems. ICHMS specifically covers contributions on research related to integrated human/machine systems at multiple scales, and includes areas such as human/machine interaction; cognitive ergonomics and engineering; assistive/companion technologies; human/machine system modeling, testing and evaluation; and fundamental issues of measurement and modeling of human-centered phenomena in engineered systems. The organization of ICHMS is based on the following Track types: Conference Theme Track, Regular Track and Special Track (selected through a call). ICHMS will also feature plenary speeches, panel sessions, tutorials, workshops, PhD track and poster sessions. Contributions are expected from academia, industry, and management agencies. All accepted papers are expected to be included in IEEE Xplore.

Conference Theme Track

Human-centered Cyber-Physical Systems

Description: Cyberphysical Systems (CPS), like transport networks or large manufacturing facilities, are naturally human-centric; in fact, they are interacting with a considerable number of distributed computing elements for monitoring, control and management which can exchange information between them and notably with humans. Such complex and physically-entangled systems of systems are of crucial importance for the quality of life of the citizens and for a country economy. Human-machine systems-oriented issues arise at different levels: System, Computing and Interaction.

Topics:

- Cyber Transportation and Human Interaction in Connected and Autonomous Vehicles
- Human-centric Internet of Things: Interaction between Humans and Smart Objects
- Smart Communities living in Smart Cyber-physical Worlds: Towards Dynamic Provision of Cyberphysycal Services to Human Ensembles
- Human Manipulation of Motorized Cyber-Physical Systems
- Co-existence and cooperation of Humans and Robots in Smart Environments

Regular Tracks

Companion Technology (Leading Chair: Andreas Wendemuth): covers and combines trans-disciplinary research in fields such as computer science and artificial intelligence, cognitive science, engineering, psychology, and neurobiology.

Brain-Machine Interfaces and Systems (Leading Chair: Tiago H. Falk): includes assistive technology and research allowing locked-in individuals to communicate and control exoskeletons/devices to improve locomotion.

Human-AI Interaction and Cognitive Computing & Engineering (Leading Chairs: Changxu Wu, Yicong Zhou): encompasses human cognition in intelligent/AI system designs or/and development of AI algorithms, technology, methods for human beings.

Interactive and Wearable Computing and Systems (Leading Chairs: Min Chen, Angelika Peer): focuses on advances and developments in all aspects of interactive and/or wearable computing and systems.

Collaborative Intelligent Systems and Applications (Leading Chairs: Weiming Shen, Makoto Itoh): addresses advances on the development of scientific and engineering foundations, innovative technologies and solutions for IoT-enabled data-driven collaborative intelligent systems.

Human Factors Engineering (Leading Chairs: Nathan J McNeese, Jessie Y.C. Chen): focuses on the advancements in theory and practice related to human interaction with intelligent agents in a wide variety of environments.

PAPER SUBMISSION

Complete manuscripts must be electronically submitted through the conference website:

http://ichms.dimes.unical.it/

Submitted manuscripts could be *Regular* papers within six (6) pages or *Poster* papers within (3) pages in IEEE two-column format, including figures, tables, and references. Please use the templates at Manuscript Templates for IEEE Conference Proceedings from the conference website to prepare your paper. All submissions MUST be in PDF format.

IMPORTANT DATES

Special Track proposals deadline:	November 15, 2019
Paper submission deadline:	December 1, 2019
Notification of acceptance:	February 15, 2020
Camera-ready copy due:	March 15, 2020

Digital Identifier 10.1109/THMS.2019.2948234

IEEE SYSTEMS, MAN, AND CYBERNETICS SOCIETY

Areas of Interest: Large-scale systems, theory and applications; optimization, decision analysis, problem definition, modeling, simulation, test, and evaluation. Foundations of cybernetics, pattern recognition, adaptive, and learning systems, biocybernetics; man–machine systems. Representative applications include complex hardware, behavioral, biological, ecological, educational, environmental, health care, management, socioeconomic, transportation, and urban systems; national priorities.

Activities: Publication of this Transactions, Newsletter, and bibliographies. Sponsorship of annual conferences and workshops. Support of SMC Society Chapter and IEEE Section meetings and conferences.

Publications:

Transactions Editor-in-Chief, TSMC: C. L. P. Chen
Transactions Editor-in-Chief, TCYB: J. Wang
Transactions Editor-in-Chief, THMS: D. B. Kaber
Transactions Editor-in-Chief, TCSS: F.-Y. Wang
Magazine Editor-in-Chief: M. El-Hawary

eNewsletter: Editor, M. Dotoli
Webmaster: C.-H. Wang
Web Site: http://www.ieeesmc.org

Technical Committees

Cybernetics

Awareness Computing, T. Murata, G. Chakraborty, R. Kozma, and Q. Zhou
Big Data Computing, V. Snasel, I. Zelinka, and M. Wozniak
Computational Collective Intelligence, N. T. Nguyen
Computational Cybernetics, I. Rudas, P. Chen, and W. Pedrycz
Computational Intelligence, X. Wang and W. Y. Ng
Computational Life Science, M. R. Berthold, H. Yan, and D. Yeung
Cybernetics and Cyber-Physical Systems, S. Hu and A. Zomaya
Diagnostics & Prognostics, I. Makki, M. Franchek, and K. Grigoriadis
Evolving Intelligent Systems, P. Angelov
Granular Computing, S. Tsumoto, T.-P. Hong, and L. Wang
Information Assurance & Intelligent Multimedia-Mobile Communications, S. Agaian, P. Chen, and A. Arakelyan
Intelligent Industrial Systems, P. Vrba, T. Strasser, and A. M. Farid
Intelligent Internet Systems, J.W. T. Lee, S.-M. Chen, T.-H. Tan, and Y.-F. Huang
Intelligent Vehicular Control Systems and Control, J. Lu and T. Gordon
Knowledge Acquisition in Intelligent Systems, S. Rubin and S.-C. Chen
Machine Learning, W. Y. Ng, D. S. Yeung, and W. Pedrycz
Medical Informatics, Y. Hata, C. M. Helgason, and Y. Hata
Pattern Recognition, Y. Y. Tang and X. You
Soft Computing, A. Abraham, M. Koeppen, and H. Takagi

Human–Machine Systems

Biometrics and Applications, D. Zhang and Y. Xu
Brain-Machine Interface Systems*, M. Smith, S.-W. Lee, V. Prasad, and R. Chavarriaga
Cognitive Computing, Y. Zhou, Y. Yuan, W. Liu, and B. Hu
Computer-Supported Cooperative Work in Design, J.-P. A. Barthès, J. Luo, and W. Shen
Environmental Sensing, Networking and Decision Making, N.-B. Chang, M. Zhou, K. W. Hipel, and S. Mammar
Human-Centered Transportation Systems, T. Okazaki, P. Bajaj, and T. Imamura
Human–Computer Interaction, M. Bolton and Y. Xiao
Human Perception in Multimedia Computing, G. Lavoue and H. Wang
Information Systems for Design and Marketing, K. Yada and Y. Zhou
Interactive and Wearable Computing and Devices, P. X. Liu, G. Fortino, M. R. Yuce, and D. Chen
Shared Control, M. Itoh, T. Gibo, and E. Boer
Visual Analytics and Communication, W. Huang, Y. Luo, and H. Duh
*Shared with Cybernetics TA

Systems Science and Engineering

Conflict Resolution, L. P. Fang and K. W. Hipel
Cyber-Physical Cloud Systems, H. Tianfield
Discrete Event Systems, M. Pia Fanti and M. D. Jeng
Distributed Intelligent Systems, W. A. Gruver, V. Marik, and H. Zhu
Enterprise Architecture and Engineering, A. van der Merwe and A. Gerber
Enterprise Information Systems, L. Xu and M. Zhou
Grey Systems, S. Liu, R. Qiu, K.-L. Wen, J. Forrest, R. Guo, Y. Yang, and N.-B. Chang

Systems Science and Engineering (continued)

Homeland Security, H. Chen, D. Brown, D. Zeng, and D. Mendonça
Infrastructure Systems and Services, M. Weijnen
Intelligent Green Production Systems, H. A. Gabbar
Intelligent Learning in Control Systems, C.-C. Tsai, K.-S. Hwang, and H.-X. Li
Intelligent Power and Energy Systems, L. L. Lai and K. P. Wong
Intelligent Transportation Systems, B.-F. Wu and J.-W. Perng
Logistics Informatics and Industrial Security Systems, R. Zhang, M. Li, M. Dresner, and Z. Zhang
Medical Mechatronics, M.-Y. Lee C.-H. Kuo, and Y.-H. Liu
Model-Based Systems Engineering, D. Dori and A. M. Madni
Robotics and Intelligent Sensing, H. Zhang and S. Nahavandi
Service Systems and Organization, J. Chen
System of Systems, M. Johnson, M. Henshaw, and F. Sahin
Systems Biology, L. Chen

Standing Committee Chairpersons

Awards, M. Berthold
Chapter Coordinators, L. Fung
Conferences and Meetings, S. Kwong
Cybernetics, V. Marik
Distinguished Lecturer, I. Rudas
Electronic Communications, W. A. Gruver
Fellows, M. Berthold
Financial Transparency, M. Smith
History, M. Smith
Human-Machine Systems, C. Nemeth
Industrial Liaison Committee, H. Nakajima
Membership and Student Activities, I. Rudas
Nominations, C. L. P. Chen
Organizing and Planning, I. Engelson
Publications, V. Kreinovich
Publications Ethics, V. Kreinovich
Search Committee for the Nominations, C. L. P. Chen
Standards, W. Lumpkins
Student Activities, G. Eigner
Systems Science and Engineering, R. Roberts
Young Professionals, S.-F. Su

IEEE Councils and Committees

Big Data Initiative, M. Berthold
Biometric Council Society, Y. Xu
IEEE/CAA Journal of Automatica Sinica Steering Committee, L. Trajkovic
IEEE Cloud Computing Steering Committee, L. Fang
IEEE Press, M. Zhou
Nanotechnology Council AdCom, K. Hwang and C-C. Tsai
Region 8 Chapter Coordination Committee, E. Herrera-Viedma and I. Rudas
RFID Council, W. A. Gruver and K. Tam
Smart Grid, L. L. Lai
Society on Social Implications of Technology Board of Govenors, M. Smith
Standards Association, W. Lumpkins
Systems Council AdCom, M. Zhou and E. Tunstel
Technical Committee on RFID, W. A. Gruver
Transactions on Affective Computing Steering Committee, G. Chakraborty and W. Shen
Transactions on Big Data Steering Committee, C. L. P. Chen
Women in Engineering, M. P. Fanti

Annual Conferences

2015 Hong Kong, China, S. Kwong
2016 Budapest, Hungary, I. J. Rudas
2017 Banff, Canada, A. Basu
2018 Miyazaki, Japan, Y. Hata and T. Murata

MEMBERSHIP GRADE REQUIREMENTS

Member: A person who has demonstrated professional competence in the fields of activities listed above. For admission, a candidate shall either 1) belong to the IEEE or to an approved Society with the grade of member or higher, or 2) have graduated from a course of study of at least four academic years' duration or its equivalent, or 3) have demonstrated competence in work of a professional character for a period of at least five years and be recommended by three SMC Society Members. Persons desiring to join the IEEE as well as the Society should apply on IEEE application forms. Former IEEE members may join the Society without rejoining the IEEE only if they have not been an IEEE member for at least five years or if their professional field of interest differs from the general interest of the IEEE. Presently approved societies are

Acoustical Society of America
Aerospace Medical Association
American Institute of Industrial Engineers
American Institute of Physics
American Psychological Association
American Society of Mechanical Engineers
Armed Forces Communications and Electronics Association
Biophysical Society
Ergonomics Society
Human Factors & Ergonomics Society of America
Institute of Naval Architects
INFORMS
Optical Society of America
Society of Automotive Engineers
Society of Engineering Psychologists
Society for Information Displays

Student: Registered students carrying at least 30 percent of a normal full-time program who are interested in the Society's field of interest. Students should apply on IEEE application forms.

Digital Object Identifier 10.1109/THMS.2019.2948240

INFORMATION FOR AUTHORS

IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS is published bimonthly in February, April, June, August, October, and December. Each paper submitted is subjected to a thorough review procedure, and the publication decision by the editor-in-chief is based on reviewer and associate editor recommendations. Review management is generally under the direction of an associate editor. Our goal is to provide review results in approximately ten weeks.

A. Process for Submission of a Technical Paper to IEEE Transactions on Human-Machine Systems

1) All papers should be submitted electronically in Portable Document Format (PDF) or PostScript. The paper should print correctly on 8.5 by 11 inch paper. A text version of your abstract is required.
2) Please utilize our manuscript submission system at:
   http://mc.manuscriptcentral.com/thms
3) While each paper should be self-contained (i.e., fully readable and understandable independently from the multimedia material), we encourage the submission of attachments with the paper. An attachment (video, animations, applets, code, data, etc.) should be referenced in the text of the paper and enhance the presentation. We currently accept avi files (PowerPoint animations), MPEG video files, Java applets, code with specific compilation instructions and execution instructions and machine specific information, and data sets with a description of the data format. More information on multimedia materials can be found at:
   http://www.ieee.org/documents/MMdocumentation.pdf
4) Our primary objective is to publish technical material not otherwise available. If any portion of the manuscript has been presented, published, or submitted for publication elsewhere, a description of the prior publication must be included in the cover letter. In addition, the original publication must be cited in the manuscript.
5) If the manuscript or a considerable part of it has been presented, published, or submitted for publication elsewhere, you must inform the Editor. In the cover letter, please explain the contribution of the present work and describe and cite the prior work so that the contribution of the new work is clear. Failure to do so may result in the immediate rejection of your manuscript. Our primary objective is to publish technical material not otherwise available elsewhere.
6) The authors are required to suggest 6 reviewers. In your suggestion of reviewers, please do not suggest current associate editors for the journal. The list of associate editors appears in each issue.
7) Papers not meeting these criteria will be withdrawn.
8) **ORCID Required:** All IEEE journals require an Open Researcher and Contributor ID (ORCID) for all authors. To create an ORCID, please visit: https://orcid.org/register. The author will need a registered ORCID in order to submit a manuscript or review a proof in this journal.

B. Style for Manuscript

1) Use IEEE TRANSACTIONS (two-column) format for all submissions. Such formatting is required for the editor to determine the length of the manuscript. Please ensure that your text fits 8.5×11in. (21.6×27.9 cm) sheets.
2) Regular papers are 10 TRANSACTIONS pages in length, including any author photographs and biographies, when requested and provided. Technical correspondences are 5 TRANSACTIONS pages in. These page counts represent maximum submission lengths and are effective as of August 1, 2017 for all new submissions. Any additional page allocations must be approved by the editor-in-chief in advance of manuscript submission or through the review process, based on the recommendation of an associate editor or technical reviewer, and excess page charges will apply. Please see the below discussion on excess page charges when manuscripts exceed the maximum number of pages.
3) The TRANSACTIONS publishes only in English. Only manuscripts written in grammatically correct English will be accepted for publication. Papers not meeting this criteria will be withdrawn. English Language Editing Services. English language editing services can help refine the language of your article and reduce the risk of rejection without review. IEEE authors are eligible for discounts at several language editing services; visit the IEEE Author Center to learn more. Please note these services are fee-based and do not guarantee acceptance.
4) Provide an informative 100 to 250 word abstract at the head of the manuscript.
5) Provide a separate double-spaced sheet listing all footnotes, beginning with "Manuscript received… " (date to be filled in by the editor-in-chief) and affiliation of the author(s), and continuing with numeric citation. Acknowledgement of financial support is placed at the end of the first footnote.
6) References should appear in a separate section at the end of the paper, with items referred to by numerals in square brackets. References must be complete and in IEEE style: Style for papers: Author, first initials followed by last name, title in quotations, periodical, volume, page numbers, month, year. Style for books: Author, title. Location: publisher, year, chapter, and page numbers (if desired). See http://www.ieee.org/documents/style_manual.pdf for more details.
7) Provide a separate sheet listing all figure captions.
8) The publication will no longer be publishing authors' photos and biographies in all papers. Only editorials and guest editorials will include author photos and biographies.
9) Accepted papers will appear on the Web in the IEEE Xplore™ system approximately three weeks after corrected page proofs are received from an author. This will typically be before the paper appears in the print version of the journal. Hence, papers will be disseminated to the research community faster utilizing electronic submission.
10) "Information for IEEE Authors" and other information may be obtained at: http://www.ieee.org/publications_standards/publications/authors/authors_journals.html.
11) IEEE supports the publication of Chinese, Japanese, and Korean (CJK) author names in the native language alongside the English versions of the names in the author list of an article. For more information, please visit the IEEE Author Digital Tool Box at the following URL:
    http://www.ieee.org/publications_standards/publications/authors/auth_names_ native_lang.pdf.
12) This journal accepts graphical abstracts and they must be peer reviewed. For more information about graphical abstracts and their specifications, please visit the following link:
    http://www.ieee.org/publications_standards/publications/graphical_abstract.pdf.

C. *Style for Illustrations*

1) Originals for illustrations (including tables) should be sharp, noise-free, and of good contrast.
2) On graphs, show only the coordinate axes, or at most the major grid lines to avoid a dense hard-to-read result. Use patterns instead of colors to distinguish bar types of trend lines (unless paying for color printing).
3) All lettering should be large enough to permit legible reduction of the figure to column width, sometimes as much as 4 to 1. There should only be one illustration per page and only illustrations should be included in the set of illustrations. Please be certain that your lettering is sufficiently large such that it will be readable after photo-reduction.
4) If an illustration is to be reproduced in color the author is responsible for the incremental cost of printing in color.

D. Open Access

This publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted $2,045 OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. Any other application charges (such as over- length page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

E. Voluntary Page and Excess Page Charges

After a manuscript has been accepted for publication, the author's company or institution will be requested to pay a voluntary charge of $110 per printed page to cover part of the cost of publication. These page charges, like those for journals of other professional societies, are not obligatory nor are their payment a prerequisite for publication. A mandatory overlength page charge of $175 is required for each page in excess of the maximum 10 pages for a regular paper and the maximum 5 pages for technical correspondences. These charges are effective as of August 1, 2017 for all new submissions. Authors who are concerned with these latter charges are encouraged to estimate the length of their manuscripts prior to submission. Detailed instructions on payment of these charges will accompany the page proof.

F. Electronic Publishing

The final version of the manuscript must comply with IEEE Periodicals requirements for text and graphics processing. The author must submit final files electronically through the manuscript submission system at http://mc.manuscriptcentral.com/thms. The required files include source file(s) (e.g., Latex or Microsoft Word), figures, and a pdf file of the entire manuscript. Figures can also be embedded in the source file of the manuscript. Acceptable formats for figures include Word, eps, ps, tiff, ppt, and Excel. Further details are contained in the information for authors accompanying the final acceptance letters.

For important information on preparing final manuscript materials visit the Tools for Authors web site at:
http://www.ieee.org/publications_standards/publications/authors/authors_journals.html

G. Copyright

It is the policy of the IEEE to own the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers, and to facilitate the appropriate reuse of this material by others. To comply with the United States Copyright Law, authors are required to sign a digital version of the IEEE Copyright Form with their original submission. Information about the electronic IEEE Copyright Form is available online at:
https://www.ieee.org/publications/rights/copyright-main.html

H. Video Abstracts

Authors are encouraged to publish an on-line "video abstract" to complement a regular paper or technical correspondence publication in the Journal. The SMC Society will host any video abstract, as a supplement to a published paper, through the SMC "YouTube" channel. Your video will also be added to the SMC Society YouTube video playlist published at the SMC Society website. The Journal recommends author consideration of American Journal Experts (https://www.aje.com) for video abstract production. Authors can send video files to the current SMC Society webmaster, Dr. Saeid Nahavandi, using one of several file transfer services. Please contact Dr. Nahavandi for further information on file transfers and making available a video abstract as part of your publication through THMS."

IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS
DAVID B. KABER, *Editor-in-Chief*, University of Florida, Herbert Wertheim College of Engineering, Department of Industrial and Systems Engineering, 303 Weil Hall / P.O. Box 116595 Gainesville, FL 32611 USA, THMS-EIC@ieee.org

Digital Object Identifier 10.1109/THMS.2019.2942470